

ԵՐԵՎԱՆԻ ՊԵՏԱԿԱՆ ՀԱՄԱԼՍԱՐԱՆ

Ռ. Ա. ԽԱՉԱՏՐՅԱՆ

ՕՊՏԻՄԻԶԱՑԻԱՅԻ ՄԵԹՈԴՆԵՐԸ
ՄԵՔԵՆԱՅԱԿԱՆ
ՈՒՍՈՒՑՄԱՆ ԽՆԴԻՐՆԵՐՈՒՄ

ՈՒՍՈՒՄՆԱԿԱՆ ՁԵՌՆԱՐԿ

ՀՏԴ 004.8(07)

ԳՄԴ 16.63g7

Խ 282

Հրատարակության է երաշխավորել
ԵՊՀ գիտական խորհուրդը:

Գրախոսներ՝ ֆիզ. մաթ. գիտ. դոկտոր, պրոֆեսոր Խ. Խաչատրյան,
ֆիզ. մաթ. գիտ. թեկնածու Վ. Միքայելյան

Խմբագիր՝ տեխ.գիտ. դոկտոր Ս. Ստեփանյան

Խաչատրյան Ռ. Ա.
Խ 282 Օպտիմիզացիայի մեթոդները մեքենայական ուսուցման խնդիրներում:
Ուսումնական ձեռնարկ / Ռ. Ա. Խաչատրյան.- Եր.: ԵՊՀ հրատ., 2025.- 386 էջ:

Ձեռնարկը նախատեսված է ԵՊՀ ինֆորմատիկայի և կիրառական մաթեմատիկայի ֆակուլտետի ուսանողների համար:

Այն կարող է օգտակար լինել բնագիտական ֆակուլտետի ուսանողներին:

ՀՏԴ 004.8(07)

ԳՄԴ 16.63g7

ISBN 978-5-8084-2708-2

<https://doi.org/10.46991/YSUPH/9785808427082>

© ԵՊՀ հրատարակչություն, 2025

© Խաչատրյան Ռ. Ա., 2025

Բովանդակություն

Նախաբան	7
1 Ներածություն	8
2 Օպտիմալ կանխատեսող ալգորիթմ (պրեդիկտոր)	17
2.1 Մնհրաժեշտ տեղեկություններ և թերեմներ հավանականությունների տեսությունից	17
2.2 Պատահական մեծությունների կովարիացիա	24
2.3 Օպտիմալ և կոռեկտ պրեդիկտորներ	26
2.4 Հիպոթեզների բազմության VC չափողականությունը	34
2.5 Մեքենայական ուսուցման ֆունդամենտալ թերեմը	36
3 Ուռուցիկ անալիզի տարրեր	42
3.1 Ուռուցիկ բազմությունների անջատման թերեմները	42
3.2 Ուռուցիկ և β -ռդորկ ֆունկցիաներ	45
3.3 Ուռուցիկ ֆունկցիայի սուբդիֆերենցիալ	48
3.4 Կուն-Թակերի թերեմը	50
4 Ռեգրեսիա և ռեգուլյարիզացիա	54
4.1 Գծային ռեգրեսիայի խնդիրը	54
4.2 Տուգանային ֆունկցիաների մեթոդը	56
4.3 Կանոնավորման(ռեգուլյարիզացիայի) մեթոդը	58
4.4 Ժամանակային շարքեր և կանխատեսում	61
5 Ֆունկցիայի մինիմիզացիայի գրադիենտային և սուբգրադիենտային մեթոդներ	66
5.1 Ֆունկցիայի մինիմիզացիայի գրադիենտային եղանակների ընդհանուր նկարագրությունը	66
5.2 Գրադիենտային իջեցում հաստատուն քայլի դեպքում	69
5.3 Կոորդինատային իջեցման մեթոդը	71
5.4 Քայլի կիսման եղանակի գուգամիտության թերեմը	72
5.5 Մարիորի մեթոդի գուգամիտությունը	75
5.6 Գրադիենտի պրոյեկտման մեթոդը	78
5.7 Գծայնացման կամ պայմանական գրադիենտի մեթոդը	81
5.8 Ստոխաստիկ գրադիենտի մեթոդը հաստատուն քայլի դեպքում	86
5.9 AdaGrad (adaptive gradient) ալգորիթմը	88

5.10	Գրադիենտային արագացված մեթոդներ (Gradient decent with momentum and Adam method)	90
5.11	Ստոխաստիկ գրադիենտի մեթոդի կիրառությունը ռեալ ռիսկի մինիմիզացիայի խնդրում	96
5.12	Ստոխաստիկ գրադիենտի մեթոդի կիրառությունը ուռուցիկ և β -ողորկ ֆունկցիաների մինիմիզացիայի խնդիրներում	97
5.13	Ստոխաստիկ գրադիենտի պրոյեկտման մեթոդներ և կիրառություններ	98
5.14	Ստոխաստիկ սուբգրադիենտային մեթոդներ	109
5.15	L_1 և L_2 կարգավորիչների մասին	111
5.16	Օպտիմալ պրեդիկտորների գոյության պրոբլեմը ուռուցիկ խնդիրներում	116
5.17	Տիխոնովի ռեգուլյարիզացիա: Ալգորիթմների կայունություն և պրեդիկտորի օպտիմալություն	118
5.18	Ստոխաստիկ գրադիենտային իջեցում (SGD with Momentum)	122
5.19	Ֆունկցիայի մինիմիզացիայի Նյուտոն-Ռաֆսոնի մեթոդը	125
5.20	Մինիմիզացիայի երկրորդ կարգի մեթոդները էքսպոնենցիալ (exp-concave) ֆունկցիաների դասի համար	131
5.21	Նյուտոն-Գաուսի մեթոդը ոչ գծային ռեգրեսիայում	135
5.22	Գրադիենտային իջեցման մեթոդում քայլի ընտրության Պոլյակի եղանակը	136
6	Հենման վեկտորների մեթոդը (support vector machine- SVM)	139
6.1	Օպտիմալ անջատող հիպերհարթության կառուցման ալգորիթմը	141
6.2	Ռոգենբլատի ալգորիթմը	148
6.3	Հենման վեկտորների մեթոդը և ստոխաստիկ գրադիենտներ	150
6.4	Օպտիմալ բաժանող հիպերհարթության կառուցումը «ադմկոտ» նմուշներից բաղկացած հավաքածուի համար	151
6.5	Կորստի ֆունկցիաները բինար դասակարգման խնդիրներում	155
6.6	Միջուկների մեթոդը մեքենայական ուսուցման մեջ	157
6.7	Բազմությունների ռադեմախերյան բարդություն	163
6.8	Օպտիմալ պրեդիկտորներ գծային դասերում և հենման վեկտորների մեթոդում	168
7	Նեյրոնային ցանցերի ուսուցում	177
7.1	Միալանքի ետ տարածման մեթոդը (Error back propagation method)	179
7.2	SVD ալգորիթմը և երկշերտ նեյրոնային ցանցեր	186
8	Մեքենայական ուսուցման հավանականային մոդելներ	187
8.1	Պայմանական հավանականություններ, Բայեսի բանաձևը	187
8.2	Օպտիմալ բայեսյան կանոններ	190
8.3	Բայեսի պարզ դասակարգիչ	193

8.4	Առավելագույն հավանականության մեթոդը (Maximum likelihood function)	195
8.5	EM (expectation maximization) ալգորիթմը	200
8.6	Լոգիստիկ ռեգրեսիայի մեթոդը	202
8.7	Գծային ռեգրեսիայի մոդելի կայունության պայմանները	208
8.8	Հավանականային խտության և բաշխման ֆունկցիաների գնահատման մեթոդները	211
8.9	Գծային տարբերակիչ (դիսկրիմինանտ) անալիզ	217
8.10	Բազմաստիճան դասակարգում (multi-class classification)	220
8.11	Նյուտոն-Ռաֆսոնի մեթոդը լոգիստիկ ռեգրեսիայում	222
9	Ամրապնդմամբ ուսուցում (reinforcement learning)	225
9.1	Մարկովյան պրոցեսներ	225
9.2	Ամրապնդմամբ ուսուցում	230
10	Օնլայն օպտիմիզացիա: Օնլայն ուսուցման մոդելներ	241
10.1	Օնլայն (առցանց) օպտիմիզացիա	241
10.2	Perceptron օնլայն ալգորիթմը	242
10.3	Բայես-Լապլասի կանխատեսման մոդելը	245
10.4	Weighted-Majority ալգորիթմը	248
10.5	Առցանց կորուստների բաշխման օպտիմալ ալգորիթմ	252
10.6	Օնլայն կանխատեսման ընդհանուր ալգորիթմ	257
10.7	Ֆոն Նեյմանի թեորեմը մինիմաքսի մասին	259
10.8	Մեքենայական ուսուցման ֆունդամենտալ թեորեմը օնլայն ուսուցման համակարգում	263
11	Չափողականության կրճատում	267
11.1	PCA (Principal component analysis) ալգորիթմը	267
11.2	Ֆիշերի տարբերակիչը (դիսկրիմինանտը)	273
11.3	Չափողականության կրճատում պատահական գծային ձևափոխությամբ	278
11.4	SNE և UMAP ալգորիթմների մասին	281
12	Կլաստերավորում (կլաստերիզացիա)	292
12.1	k-միջինների ալգորիթմը	292
12.2	Կլաստերավորման մետրիկական մեթոդներ: DbSCAN ալգորիթմը	296
12.3	k-մոտիկ հարևանների մեթոդը	302
12.4	Կլաստերավորման տրամաբանական մեթոդներ: Որոշիչ ծառեր	307
12.5	Որոշիչ (վճիռների) ծառեր գծային ռեգրեսիայի և դասակարգման խնդիրներում	314
12.6	Կլաստերավորման GMM մոդելը (gaussian mixture model)	318
12.7	Apriori ալգորիթմը և ասոցիատիվ կանոններ	321
13	Մոդելների գնահատման չափանիշներ	327

13.1	Խաչաձև ստուգում (k-fold cross validation)	327
13.2	Դասակարգման մոդելների որակի գնահատման չափանիշներ	328
13.3	Վարկանիշային (ռանժիրովկայի) մեթոդներ և մոդելների գնահատման որոշ չափանիշներ	334
13.4	Կլաստերավորման մոդելների գնահատման մետրիկական չափանիշներ	341
14	Ուժեղացում (Boosting algorithms)	342
14.1	Թույլ հիպոթեզների դասեր	342
14.2	Բուստինգ դասակարգման խնդիրներում, Adaboost ալգորիթմը	344
14.3	Boosting ռեգրեսիայի խնդիրներում	350
14.4	AnyBoost ալգորիթմը	352
14.5	Գրադիենտային բուստինգ	356
15	Առաջարկությունների համակարգեր	363
15.1	Սատրիցների սինգուլյար վերլուծություն (SVD-singular value decomposition)	363
15.2	SVD մոդելի կիրառությունը առաջարկությունների համակարգերում	368
15.3	Memory-based մոդելներ	370
16	Հավելվածներ	372
16.1	Ավտոմատ դիֆերենցում և դիֆերենցիալ հավասարումների լուծումը նեյրոնային ցանցերով	372
16.2	Ստոխաստիկ գրադիենտի փուլային ալգորիթմը (դիսպերսիայի նվազեցում)	375
16.3	Ստոխաստիկ գրադիենտի մեթոդը ուժեղ աճի պայմանով (strong growth condition)	376
16.4	Օպտիմալ գրադիենտային մեթոդների մասին	378
	Резюме	379
	Abstract	380
	Գրականություն	381

ՆԱԽԱԲԱՆ

Ձեռնարկում շարադրվում են օպտիմիզացիայի մեթոդների կիրառություններ մեքենայական ուսուցման խնդիրներում: Նախապես ներկայացվում են մեքենայական ուսուցման հիմնական հասկացությունները՝ օբյեկտների հավաքածու, մոդելներ, կորստի ֆունկցիաներ և այլն:

Ներմուծվում է օպտիմալ կանխատեսող ալգորիթմների (պրեդիկտորների) գաղափար: Քննարկվում են դրանց գոյության և կառուցման ալգորիթմներ: Ուսումնասիրվում են այդ ալգորիթմների գուգամիտության, կայունության և օպտիմալության պայմանները: Տրվում են էմպիրիկ և ռեալ ռիսկերի տարբերության գնահատականներ, որոնք արտահայտվում են մեքենայական ուսուցման ֆունդամենտալ թեորեմում:

Նկարագրվում են գրադիենտային իջեցման մեթոդները և նրա մոդիֆիկացիաները (տարբերակները): Հատկապես առանձնացվում են ստոխաստիկ և արագացված գրադիենտային եղանակները, որոնք լայն կիրառություններ ունեն մեքենայական ուսուցման ալգորիթմներում:

Ձեռնարկը նախատեսված է կիրառական մաթեմատիկա ինչպես նաև համակարգչային գիտություն մասնագիտությամբ սովորող ուսանողների և մագիստրանտների համար:

Այն կարող է օգտակար լինել այն հետազոտողների համար, ովքեր ցանկանում են խորացնել իրենց տեսական և գործնական գիտելիքները մեքենայական ուսուցման բնագավառում:

Ենթադրվում է, որ ընթերցողը ծանոթ է հավանականությունների տեսության, գծային հանրահաշվի, մաթեմատիկական անալիզի, օպտիմիզացիայի մեթոդների և ալգորիթմների տեսության հիմունքներին:

1

Ներածություն

Մեքենայական ուսուցումը փորձի վրա հիմնված ուսումն է: Օրինակ՝ դա նման է այն մարդուն, ով սովորում է շախմատ խաղալ՝ դիտելով ուրիշների խաղը: Այս կերպ համակարգիչները կարող են ծրագրավորվել՝ իրենց տրամադրված տեղեկատվության հիման վրա: Համակարգիչները ունակություն են ձեռք բերում բացահայտելու նոր տվյալների բնութագրիչներ:

Մեքենայական ուսուցումը բաղկացած է հետևյալ փուլերից.

- տեղեկատվության հավաքագրում,
- տվյալների հավաքագրում,
- տվյալների վերլուծություն,
- ալգորիթմի մշակում,
- կատարել կանխատեսումներ մշակված ալգորիթմների օգնությամբ:

Ալգորիթմները բաժանվում են մի քանի խմբերի, որոնցից հիմնականներն են՝

- առանց վերահսկման ուսուցում (ուսուցում առանց ուսուցչի, unsupervised learning),
- վերահսկվող ուսուցում (ուսուցում ուսուցչով, supervised learning),
- կիսավերահսկվող ուսուցում (Semi-Supervised Learning),
- ամրապնդմամբ ուսուցում (reinforcement learning):

Ձեռնարկում քննարկվում են այս ալգորիթմներից առաջին երկուսը և չորրորդը (գլուխ 9):

Չվերահսկվող ուսուցման (unsupervised learning) ժամանակ համակարգիչը ստանում է միայն մուտքային տվյալների մի հավաքածու: Այնուհետև այն կապ է հաստատում այդ մուտքագրված տվյալների և ցանկացած այլ հիպոթետիկ տվյալների միջև: Մեքենան կատարում է տվյալների խմբավորում (կլաստերավորում):

Ընդհանուր առմամբ կան չափողականության կրճատման մի քանի հայտնի ալգորիթմներ: Դրանցից են՝ գլխավոր կոմպոնենտների ալգորիթմը PCA (principal component analysis), պատահական պրոյեկտավորման մեթոդը, t-SNE (t-distributed stochastic neighbor embedding) և UMAP (uniform manifold approximation and projection) ալգորիթմները: Ձեռնարկում քննարկում են այս բոլոր մեթոդները:

t-SNE (t-բաշխմամբ ստոխաստիկ հարևանի ներդրում) և UMAP ալգորիթմների օպտիմիզացիոն խնդիրները և դրանց լուծման ստոխաստիկ գրադիենտային մեթոդները ավելի մանրամասն շարադրված են [13, 14] հոդվածներում:

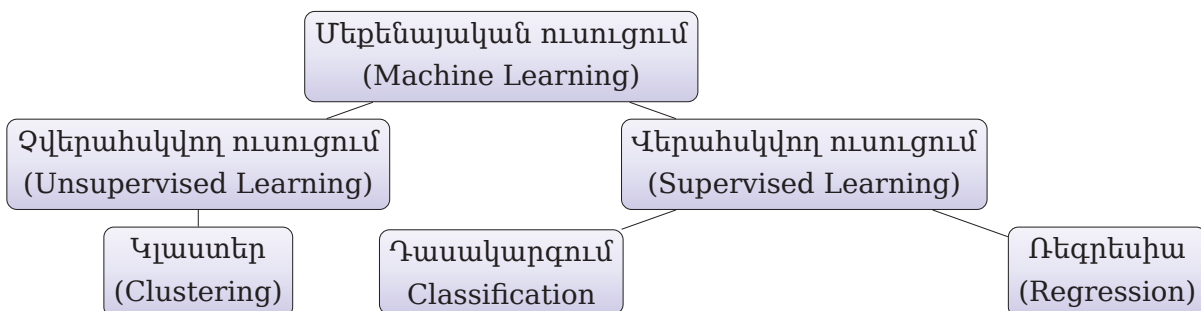
Այս ձեռնարկում ներկայացվում են այդ օպտիմիզացիոն խնդրի լուծման գրադիենտային եղանակը և նպատակային ֆունկցիայի հատկությունները (տե՛ս [27, 29]): UMAP ալգորիթմը SNE-ի մոդիֆիկացված տարբերակն է:

Այս մեթոդներով մեծ չափողականությամբ տվյալները ներդրվում են ավելի փոքր չափողականությամբ տարածություններում (հիմնականում երկու կամ երեք):

Վերահսկվող ուսուցումը (supervised learning) ներառում է համակարգչի կարողությունը՝ ճանաչելու էլեմենտները տրամադրված նմուշների հիման վրա: Համակարգիչը ուսումնասիրում է այն և զարգացնում է այս տվյալների հիման վրա նոր տվյալներ ճանաչելու ունակությունը:

Վերահսկվող ուսուցման հիմնական ալգորիթմներն են՝

- հենման վեկտորների մեթոդը,
- բայեսյան դասակարգիչները (կլասիֆիկատորները),
- որոշիչ ծառերը,
- k-մոտիկ հարևանների մեթոդը,
- գծային ռեգրեսիան:



Ձեռնարկում ներկայացվում են մեքենայական ուսուցման հիմնական հասկացությունները, մեթոդները և ալգորիթմները, որոնք խոշոր տվյալների մշակման և ինտելեկտուալ համակարգերի նախագծման տեսական հիմք են:

Այժմ ավելի մանրամասն նկարագրենք մեքենայական ուսուցման հիմնական խնդիրները և մեթոդները:

Մերենայական ուսուցումը արհեստական բանականության տեսության բաժին է, որի մեթոդներով կատարվում են օբյեկտների դասակարգում և կանխատեսում նախապես տրված օբյեկտների հավաքածուի միջոցով: Այժմ ձևակերպենք դասակարգման (կլասիֆիկացիայի) հիմնական խնդիրը: Դիցուք տրված են

- X և Y բազմությունները: X -ը անվանենք օբյեկտների բազմություն, իսկ Y -ը՝ պատասխանների կամ նշիչների բազմություն:
- Դիցուք Z -ը $X \times Y$ դեկարտյան արտադրյալի վերջավոր կամ հաշվելի ենթաբազմություն է, որի վրա տրված է հավանականային բաշխման ֆունկցիա՝

$$P : Z \rightarrow [0, 1], \sum_{z \in Z} p(z) = 1:$$

Դիտարկենք Z -ի վերջավոր մի ենթաբազմություն՝

$$S = \{z_i \equiv (x_i, y_i) \in X \times Y | i = 1, 2, \dots, l\}:$$

Անվանենք այն ուսուցանող հավաքածու (training sample): Վերահսկվող ուսուցման հիմնական խնդիրն է. S հավաքածուի տվյալների հիմնան վրա կառուցել $\hat{h} : X \rightarrow Y$ ֆունկցիա, այնպիսին, որ նա օժտված լինի այսպես ասած «ընդհանրական» հատկությամբ՝

$$\hat{h}(x) \approx y, (x, y) \in Z:$$

Այս ֆունկցիան կոչվում է պրեդիկտոր, հիպոթեզ կամ դասակարգիչ: Ճշտենք մոտավոր նշանի իմաստը:

Այդպիսի պրեդիկտորը կառուցվում է հետևյալ կերպ.

- սահմանվում է $h : X \rightarrow Y$ պրեդիկտորի սխալանք՝ $\text{error}(h) = E_{z \sim P}[h(x) \neq y]$,
- h պրեդիկտորի էմպիրիկ ռիսկ կոչվում է հետևյալ թիվը՝

$$L_S(h) = \frac{1}{l} \sum_{i=1}^l [h(x_i) \neq y_i]:$$

- Ընտրվում է h պրեդիկտորների ինչ-որ մի H դաս, և այդ դասի վրա մինիմիզացվում է էմպիրիկ ռիսկի $L_S(h)$ ֆունկցիան ըստ h -ի՝ կառուցելով $h_S(x)$ պրեդիկտորը՝

$$h_S = \arg \min_{h \in H} L_S(h):$$

- Այնուհետև գնահատվում է ռեալ ռիսկը, այսինքն՝ հետևյալ տարբերությունը՝

$$\text{error}(h_S) - \inf_{h \in H} \text{error}(h):$$

Դասակարգման հիմնական խնդիրը հետևյալն է՝ ուսումնասիրել, թե պրեդիկտորների ինչպիսի H դասերի և S հավաքածուների որ հզորությունների դեպքում

$\text{error}(h_S) - \inf_{h \in H} \text{error}(h)$ տարբերությունը կարելի է դարձնել փոքր նախապես տրված ϵ դրական թվից:

Նման ձևով կառուցված h_S պրեդիկտորը կոչվում է որոշիչ (decision): Փաստորեն h_S որոշիչ պրեդիկտորի օգնությամբ կարելի է կանխատեսել այնպիսի $x \in X$ օբյեկտների նշիչներ, որոնք հավաքածուից դուրս են:

Օբյեկտները կարող են լինել տարբեր բնույթի՝ մարդիկ, երկրներ, կազմակերպություններ, կայքեր, պատկերներ և այլն: Օբյեկտները նկարագրվում են հատկություններով (features) հետևյալ կերպ՝

- տրվում է օբյեկտների հատկությունների Γ բազմություն,
- յուրաքանչյուր $i \in \Gamma$ հատկությանը համապատասխանեցվում է նրա արժեքների D_i բազմություն,
- ամեն մի x օբյեկտի հետ կապվում է $i \in \Gamma$ հատկության $x^i \in D_i$ արժեք:

Այսպիսով՝ x օբյեկտը $(x^1, \dots, x^n) \in \mathbb{R}^n$ կոորդինատներով վեկտոր է:

Այժմ բերենք մեքենայական ուսուցման մի խնդիր, որը վերաբերվում է վարկավորմանը: Օբյեկտ-վեկտորները այստեղ հարցումներ են վարկ (կրեդիտ) ստանալու վերաբերյալ: Իսկ պատասխանների բազմությունն է՝ $Y = \{1, -1\}$: Ամեն մի օբյեկտի համար $h(x)$ -ը մեկնաբանվում է հետևյալ կերպ՝ բանկը վարկ տալու^o է, թե^o ոչ:

- «1» պատասխանը նշանակում է՝ վարկ տրամադրել x հարցման (դիմումի) հիման վրա,
- «-1» պատասխանը նշանակում է՝ մերժել դիմումը:

Բանկի այս որոշումները համարվում են ճիշտ, եթե

- հաճախորդը, որը ստացել է վարկը, վերադարձնում է այն ժամանակին:
- Հաճախորդը, որին մերժել են, մեծ հավանականությամբ չի կարող վերադարձնել վարկը ժամանակին:

Հայտնի է $S = \{(x_i, y_{x_i}), i = 1, 2, \dots, l\}$ բազմությունը, որտեղ

- x_i -ն վարկավորման դիմումն է,
- y_{x_i} -ն բանկի կայացրած որոշումն է:

Այս խնդրում օբյեկտները հաճախորդներ են, որոնք կարող են բնութագրվել հետևյալ հատկանիշներով՝ սեռը, կայուն աշխատանք ունենալը, մասնագիտությունը, աշխատունակությունը, աշխատավարձը, տարիքը, ընտանիքի եկամուտը և այլն:

Պահանջվում է կառուցել h_S որոշիչ պրեդիկտորը այնպես, որ ամեն մի նոր x հարցման դեպքում բանկը տա ճիշտ պատասխան:

Սովորաբար որպես պրեդիկտորների H դաս օգտագործում են ինչ-որ պարամետրերից կախված ֆունկցիաների ընտանիք: Օրինակ՝ դիտարկվող

խնդրում որպես պրեդիկտորների H դաս սովորաբար վերցնում են $h_w(x) = \text{sign}(w, x)$ տեսքի ֆունկցիաները, որտեղ w պարամետրը փոխվում է ինչ-որ M բազմության վրա: Կատարվում է Էմպիրիկ ռիսկի ֆունկցիայի մինիմիզացիա ըստ w պարամետրի:

Սակայն այն ունի իր առանձնահատկությունը՝ $l(h, (x, y)) = [h_w(x) \neq y]$ ֆունկցիան, որը կոչվում է կորստի ֆունկցիա (loss function), ողորկ չէ ըստ նշված պարամետրի: Կորստի այս խզվող ֆունկցիան վերևից մոտարկվում է անընդհատ կամ ողորկ ուռուցիկ ֆունկցիայով, և այնուհետև կատարվում է Էմպիրիկ ռիսկի մինիմիզացիա՝ օգտագործելով գրադիենտային կամ սուբգրադիենտային իջեցման տարբեր մեթոդներ, մասնավորապես՝ ստոխաստիկ գրադիենտային եղանակները [48, 6, 46, 39]:

Դիտարկենք դասակարգման խնդրի ևս մի օրինակ: Աղյուսակ 1.1-ում տրված է 100 բուժառուների (patients) տվյալներ:

Աղյուսակ 1.1			
	Չարկերակ (puls)	Հեմոգլոբին	Ախտորոշում
x_1	70	140	առողջ ($y=1$)
x_2	60	160	առողջ ($y=1$)
x_3	94	120	միոկարդիա ($y=-1$)
...	...	...	...
x_{100}	86	98	միոկարդիա ($y=-1$)

Այստեղ ուսուցանող հավաքածուն է

$$((70, 140), y = 1), ((60, 160), y = 1), ((94, 120), y = -1), \dots :$$

Պահանջվում է նոր բուժառուի համար կանխատեսել պատասխանը, օրինակ՝ $x = (75, 128)$, $y = ?$: Սա նույնպես դասակարգման խնդիր է և կոչվում է «ուսուցում ուսուցչի հետ»: Նման խնդիրներում հավաքածուի օբյեկտները ունեն իրենց պատասխանները, այսինքն՝ Y բազմությունը առկա է:

Ռեգրեսիայի խնդիրներում $X = \mathbb{R}^n$, $Y = \mathbb{R}^k$: Էմպիրիկ ռիսկն ունի հետևյալ տեսքը՝

$$L_S(h) = \frac{1}{l} \sum_{i=1}^l l(h, z_i),$$

որտեղ

$$l(h, z_i) = \|h(x_i) - y_i\|^2 \text{ կամ } l(h, z_i) = \|h(x_i) - y_i\|:$$

Իսկ $h : \mathbb{R}^n \rightarrow \mathbb{R}^k$ պրեդիկտորը (կանխատեսողը) ընտրվում է պարամետրից կախված ֆունկցիաների ինչ-որ ընտանիքից՝

$$h(x) = a(x, w), \quad w \in \mathbb{R}^n:$$

Մասնավորապես $a(x, w)$ ֆունկցիաները կարող են լինել գծային կամ ինչ-որ աստի-

ճանի բազմանդամներ:

Երբ հավաքածուում առկա են միայն օբյեկտներ՝ առանց պատասխանների, ապա կատարվում է կլաստերավորում, որի խնդիրն է օբյեկտները բաժանել խմբերի (կլաստերի) այնպես, որ մի կլաստերում լինեն իրար մոտ գտնվող առարկաներ, իսկ տարբեր կլաստերներում առարկաները զգալիորեն տարբերվեն: Կլաստերը կարող է բնութագրվել որպես ընդհանուր հատկություններ ունեցող օբյեկտների խումբ: Նման ալգորիթմները կոչվում են «ուսուցում առանց ուսուցչի»: Դիտարկենք նման մի օրինակ:

Աղյուսակ 1.2-ում ներկայացված են 100 մարդու տվյալներ: Պահանջվում է որոշել նրանց սեռը:

Աղյուսակ 1.2				
	Քաշ	Հասակ	Տարիք	Մազերի երկար.
x_1	96	170	42	կարճ
x_2	60	180	25	կարճ
x_3	54	165	30	երկար
...	...	...	...	...
x_{100}	108	193	32	երկար

Վերջին տարիներին մեքենայական ուսուցման նկատմամբ հետաքրքրությունը կտրուկ աճել է, քանի որ մեծ թվով տվյալների աճով պայմանավորված՝ մեծացել է նաև այդ տվյալների վերլուծության և մեկնաբանման արդյունավետ ալգորիթմների մշակումը: Ուռուցիկ օպտիմիզացիան այդ կարևոր մեթոդներից մեկն է: Դա պայմանական օպտիմիզացիայի խնդիր է, որում փնտրվում է ուռուցիկ ֆունկցիայի մինիմումի կետը գծային սահմանափակումների դեպքում: Մեքենայական ուսուցման շատ ալգորիթմներ մոդելավորվում են որպես ուռուցիկ օպտիմիզացիայի խնդիրներ: Դրա փայլուն օրինակ է բինար կլասիֆիկացիայի խնդրում օպտիմալ հիպերհարթության կառուցումը, որը տարբեր դասի օբյեկտները բաժանում է իրարից: Դա ձևակերպվում է որպես քառակուսային ծրագրավորման խնդիր՝ գծային սահմանափակումներով: Այդ իսկ պատճառով աշխատանքի առաջին մասում շարադրվում են ուռուցիկ անալիզի հիմնարար երկու՝ ուռուցիկ բազմությունների անջատման և Կուն-Թակերի թեորեմները՝ որպես ուռուցիկ ծրագրավորման խնդիրների լուծման հիմնական, ընդհանուր եղանակներ: Օգտագործելով Կուն-Թակերի թեորեմը բինար կլասիֆիկացիայի խնդրում՝ օպտիմալ հիպերհարթության կառուցման ալգորիթմը ձևակերպվում է որպես ուռուցիկ օպտիմիզացիայի խնդիր, որում սահմանափակումները տրվում են գծային հավասարությունների և անհավասարությունների տեսքով [27, 20, 38, 48]: Այս ալգորիթմը մեքենայական ուսուցումում կոչվում է հենման վեկտորների մեթոդ (SVM-support vector machines), որով կառուցվում է հիպերհարթություն, որը առավել ճշգրիտ է առանձնացնում տարբեր տիպի օբյեկտները իրարից [48, 27, 44, 20, 8, 29, 38]:

Աշխատանքում շարադրվում են գծային ռեգրեսիայի մոդելներ, որոնցով կանխատեսվում է գծային կախվածություն օբյեկտների և պատասխանների միջև

[48, 26, 27, 44, 8]: Կատարվում են նաև կանխատեսումներ ժամանակային շարքերի մոդելներով [25]:

Այնուհետև դիտարկվում է տրամաբանական (logistic) ռեգրեսիայի մոդելը, որով օբյեկտները խմբավորվում են երկու դասերում: Այդ մոդելը ելքում ունենում է $[0, 1]$ միջակայքի թիվ: Եթե այն փոքր է $1/2$ -ից, ապա օբյեկտը դասվում է առաջին խմբին, հակառակի դեպքում՝ երկրորդ դասին [48, 26, 27, 38]:

Աշխատանքում ուսումնասիրվում է օբյեկտների կլաստերավորման k միջինների և *DbSCAN* ալգորիթները, որոնք կարող են խմբավորել օբյեկտները՝ հիմնվելով նմանության վրա: Ըստ էության դա նշանակում է օբյեկտները բաժանել մի քանի խմբերի՝ մոտավորապես նմանատիպ հատկություններով [48, 20, 36]:

Դիտարկվում են նաև *Apriori* ալգորիթի միջոցով օբյեկտների խմբավորման ասոցիատիվ կանոններ [18]:

Ձեռնարկում առանձնակի տեղ են գրավում մեքենայական ուսուցման հավանականային բայեսյան դասակարգիչները (կլասիֆիկատորները): Ձևակերպվում է միջին ռիսկի ֆունկցիա, որի մինիմիզացիայի արդյունքը Բայեսի դասակարգիչն է [48, 27, 24, 8, 45]:

Քննարկվում է պայմանական հավանականության գնահատման իտերացիոն *EM* օպտիմիզացիոն ալգորիթը [23]:

Դիտարկվում են նաև օնլայն կանխատեսման ալգորիթներ՝ օգտագործելով փորձագետների ստրատեգիաներ [48, 27]:

Աշխատանքի իններորդ գլխում դիտարկվում են ամրապնդմամբ ուսուցման (reinforcement learning) օպտիմիզացիոն խնդիրները [12, 17]: Քննարկվում են *Q*-learning ալգորիթը Բելմանի հավասարումը լուծելու համար և ալգորիթի գույքամիտության պայմաններ: Ալգորիթը մեկնաբանվում է մի քանի օրինակներով:

Աշխատանքում նմուշների հավաքածուի օգնությամբ կատարվում է բաշխման և խտության ֆունկցիաների մոտարկումներ: Կառուցվում է, այսպես ասած, Էմպիրիկ բաշխման ֆունկցիաներ, և ցույց է տրվում, որ դրանք հավանականային իմաստով գույքամիտում են իրական բաշխման ֆունկցիային [36, 12]:

Ձեռնարկում քննարկվում են նաև գրադիենտային «բուստինգների» տարբեր ալգորիթներ և դրանց գույքամիտության հարցեր [27]:

Դիտարկվում է նաև սխալանքի ետ տարածման մեթոդը նեյրոնային ցանցերում, որը իրականում գրադիենտային իջեցման եղանակի իրականացումն է այդպիսի ցանցերի օգնությամբ: Այսինքն՝ կատարվում է նեյրոնային ցանցերի ուսուցում:

Աշխատանքում կարևոր տեղ է հատկացված նաև գլխավոր բաղադրիչների (կոմպոնենտների) մեթոդին, որով կատարվում է օբյեկտների չափողականության կրճատում՝ առանց ինֆորմացիայի մեծ կորստի [1]: Այն կատարվում է մատրիցների սինգուլյար վերլուծության հիման վրա (SVD-singular value decomposition):

Աշխատանքի վերջում դիտարկվում են մեքենայական ուսուցման երրորդ դասի խնդիրներ, երբ Y պատասխանների բազմությունը կարգավորված է:

Այստեղ օբյեկտները բաղկացած են երկու կոմպոնենտից: Առաջինը կոչվում է հարցում, երկրորդը՝ ցուցակ: Յուրաքանչյուր հարցման դեպքում ներկայացվում է փաստաթղթերի ցուցակ, որոնք կարգավորված են վարկանիշներով: Ներմուծվում են հարցման որակի գնահատման որոշակի չափանիշներ այդ վարկանիշների օգնությամբ: Այստեղ պահանջվում է կառուցել այնպիսի մոդել, ելնելով տրված X հավաքածուից և Y բազմությունից, որ այն նոր օբյեկտների համար լավագույնս կանխատեսի նրա վարկանիշը [20]:

Իսկ ինչպե՞ս ընտրել, թե որ ալգորիթմն է տվյալ խնդրի համար լավ: Դա արվում է այսպես ասած *cross – validation* մեթոդով: Աշխատանքում տվում է այդ մեթոդի շարադրանքը [48, 26, 8]:

Մեքենայական ուսուցման ալգորիթմները ունեն բազմաթիվ պրակտիկ կիրառություններ. մասնավորապես առաջարկությունների համակարգերում, բնական լեզվի մշակման, պատկերների ճանաչման և համակարգչային տեսողության (*computer vision*) բնագավառների խնդիրներում: Օրինակ՝ *Computer Vision* մոդելների օգնությամբ կարելի է կանխատեսել ճանապարհային խցիկների արձանագրած ավտոմեքենաների համարանիշները, որոնք այս կամ այն պատճառով աղավաղված են: Կամ՝ սմարթֆոնի միջոցով իմանալ՝ ով է տերը:

Աշխատանքում քննարկվում է առաջարկությունների համակարգերի կառուցման *SVD* մոդելը մատրիցների սինգուլյար վերլուծության հիման վրա, որը այժմ լայն տարածված է ապրանքների օնլայն վաճառքի ցանցերում [20]:

Ձեռնարկի հիմնական առանձնահատկությունն այն է, որ մաթեմատիկական խստությամբ ներկայացվում են ալգորիթմների գույճամիտության և կայունության թեորեմները [48, 27]:

Ձեռնարկում կան նաև բազմաթիվ օրինակներ և խնդիրներ, որոնք օգնում են յուրացնելու տեսական նյութը, իսկ տրված առաջադրանքները, որոնք կարող են դառնալ կուրսային աշխատանքների թեմաներ:

Աշխատանքում ընդգրկված մեթոդների հիմքը մաթեմատիկական վիճակագրությունն է և իհարկե ուռուցիկ անալիզը (մասնավորապես՝ ուռուցիկ բազմությունների և ֆունկցիաների հատկությունները և Կուն-Թակերի թեորեմը) [46, 42, 35, 43, 32, 39, 50]:

Պետք է նաև նշել այս ձեռնարկի ևս մի կարևոր առանձնահատկություն: Այստեղ որոշ պարագրաֆներ կարող են դառնալ ավարտական և մագիստրական թեզերի նյութեր (օրինակ՝ տասներորդ գլխում դիտարկված օնլայն ուսուցման կանխատեսման մոդելները և մեքենայական ուսուցման ֆունդամենտալ թեորեմը):

Ծատ հեղինակների կարծիքով՝ այս պահին մեքենայական ուսուցման ալգորիթմների վերաբերյալ լավագույն դասագրքեր են համարվում [48] և [22] -ը:

Ձեռնարկում ընդունված են հետևյալ նշանակումները. $S = z_1, z_2, \dots, z_l$ նշանակում է l օբյեկտների հաջորդականություն, որում կարող են լինել նաև կրկնություններ: Կամ՝ երբեմն S -ը կանվանենք նաև ուսուցանող հավաքածու:

D -ով կնշանակենք հավանականային բաշխում ինչ-որ Z բազմության վրա:

$z \sim D$ նշանակում է z օբյեկտը ընտրված է D բաշխումից:

$E_{z \sim D}[f(z)]$ -ը $f(z) : Z \rightarrow R$ պատահական մեծության մաթեմատիկական սպասումն է:

Ձեռնարկում որոշ տերմինների հայերեն համարժեքը տրված չէ (պահպանվել է անգլերեն տարբերակը) նրանց իմաստը չաղավաղելու նկատառումներով:

Ենթադրվում է, որ ընթերցողը տիրապետում է մաթեմատիկական անալիզի, գծային հանրահաշվի, օպտիմիզացիայի մեթոդների, հավանականությունների տեսության և կիրառական վիճակագրություն տեսությունների հիմնարար գաղափարներին և հասակացություններին բուհերի ֆիզիկամաթեմատիկական ֆակուլտետների ծրագրերի շրջանակներում: Ձեռնարկը նախատեսված է բուհերի ֆիզիկամաթեմատիկական ֆակուլտետների ուսանողների և ասպիրանտների համար: Այն կարող է օգտակար լինել նաև «մեքենայական ուսուցում» և «կանխատեսման տեսություն» առարկաներ դասավանդողներին:

Խնդրում ենք տեղեկացնել հեղինակին նկատված թերությունների և այս տեքստը բարելավելու վերաբերյալ առաջարկությունների մասին khachatryan.rafik@gmail.com հասցեով: Հեղինակը երախտապարտ կլինի ցանկացած դիտողության և առաջարկության համար:

2

Օպտիմալ կանխատեսող ալգորիթմ (պրեդիկտոր)

2.1 Անհրաժեշտ տեղեկություններ և թեորեմներ հավանականությունների տեսությունից

1. Հավանականության էլեմենտար սահմանում:

Դիցուք Ω -ն վերջավոր կամ հաշվելի բազմություն է: Ենթադրենք՝ նրա ամեն մի $\omega \in \Omega$ տարրին համապատասխանության մեջ է դրվում $p(\omega)$ թիվ, այնպիսին, որ

$$0 \leq p(\omega) \leq 1, \sum_{\omega \in \Omega} p(\omega) = 1:$$

ω էլեմենտները կոչվում են տարրական (էլեմենտար) պատահարներ, իսկ $p(\omega)$ -ն՝ ω -ի հավանականություն: p -ն կոչվում է հավանականային բաշխման ֆունկցիա Ω բազմության վրա: Եթե $A \subset \Omega$ ենթաբազմություն է, ապա այն կոչվում է պատահար, իսկ նրա հավանականություն է կոչվում հետևյալ մեծությունը՝

$$\sum_{\omega \in A} p(\omega):$$

2. Հավանականային տարածություն: Պատահարների $\{A_i \subset \Omega, i \in I\}$ ընտանիքը կոչվում է անհամատեղելի, եթե

$$A_i \cap A_j = \emptyset, i \neq j:$$

Սահմանում: Ω բազմության ենթաբազմությունների F ընտանիքը կոչվում է σ -հանրահաշիվ, եթե

- $\Omega \in F$,
- $A \in F \rightarrow A^c \in F$,
- $A_n \in F, n = 1, 2, \dots \rightarrow \bigcup_{i=1}^{\infty} A_i \in F$:

Սահմանում: $P : A \rightarrow [0, 1]$, $A \in F$ ֆունկցիան կոչվում է հավանականություն, եթե

- $P(\Omega) = 1$:
- Եթե $A_i \in F$, $i = 1, 2, \dots$ պատահարները անհամատեղելի են, ապա

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i):$$

(Ω, F, P) եռյակը կոչվում է հավանականային տարածություն:

3. Պայմանական հավանականություններ:

Սահմանում: A պատահարի պայմանական հավանականությունը B պայմանով կոչվում է

$$P(A|B) \equiv \frac{P(A \cap B)}{P(B)}$$

մեծությունը:

Սահմանում: A և B պատահարները կոչվում են անկախ, եթե

$$P(A \cap B) = P(A)P(B):$$

4. Պատահական մեծություններ: Դիցուք (Ω, F, P) հավանականային տարածության վրա տրված է $X : \Omega \rightarrow R$ ֆունկցիա, այնպիսին, որ կամայական $B \subset R$ չափելի բազմության համար

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in F:$$

Այս դեպքում $P(X^{-1}(B))$ -ն կոչվում է X պատահական մեծության B -ում ընկնելու հավանականություն: $F(x) = P(X < x)$ ֆունկցիան կոչվում է $X(\omega)$ պատահական մեծության բաշխման ֆունկցիա: Վերջինիս միջոցով կարելի է հաշվել կամայական $A \in F$ պատահարի հավանականությունը Սթիվենսի ինտեգրալի օգնությամբ՝

$$P(A) = \int_A dF(x):$$

Եթե բաշխման $F(x)$ ֆունկցիան բացարձակ անընդհատ է, ապա վերևի բանաձևը կարելի է ներկայացնել հետևյալ կերպ՝

$$P(A) = \int_A f(x)dx,$$

որտեղ f -ը կոչվում է X պատահական մեծության խտության ֆունկցիա:

5. Պատահական մեծությունների թվային բնութագրիչները: Դիցուք $f(x)$ -ը X պատահական մեծության խտության ֆունկցիան է: Այդ դեպքում

$$E[X] = \int_{-\infty}^{+\infty} xf(x)dx$$

թիվը կոչվում է X պատահական մեծության միջին արժեք կամ մաթեմատիկական սպասում:

$$D(X) = E[(X - E[X])^2] = \int_{-\infty}^{+\infty} (x - E[X])^2 f(x)dx$$

մեծությունը կոչվում է դիսպերսիա: Այս թվային բնութագրիչները օժտված են հետևյալ հատկություններով.

- $E[aX + b] = aE[X] + b, a, b \in R$:
- $D[aX + b] = E[aX + b - aE[X] - b]^2 = a^2 E[(X - E[X])^2] = a^2 D[X]$:
- $D[X] = E[(X - E[X])^2] = E[X^2 - 2XE[X] + (E[X])^2] =$
 $= E[X^2] - 2E[X]E[X] + (E[X])^2 = E[X^2] - (E[X])^2$:

Դիցուք X -ը պատահական մեծություն է, իսկ g -ն չափելի ֆունկցիա: $g(X)$ պատահական մեծության միջին արժեք կոչվում է հետևյալ մեծությունը՝

$$E[g(X)] = \int_R g(x)f(x)dx:$$

Այժմ ցույց տանք հետևյալ անհավասարությունը, որը կոչվում է Յենսենի անհավասարություն:

Թեորեմ 2.1. Դիցուք X -ը պատահական մեծություն է, իսկ $f(x)$ -ը ուռուցիկ ֆունկցիա է: Այդ դեպքում տեղի ունի

$$E[f(X)] \geq f(E[X])$$

անհավասարությունը:

Ապացույց: Ապացույցը կատարենք դիֆերենցելի f ուռուցիկ ֆունկցիայի համար: Հայտնի է, որ այս դեպքում ճշմարիտ է

$$f(x) - f(y) \geq (f'(y), x - y) \quad \forall x, y$$

անհավասարությունը: Այստեղ տեղադրելով

$$x = X, y = E[X]$$

կստանանք

$$f(X) \geq f(E[X]) + (f'(E[X]), X - E[X]):$$

Վերցնելով այս անհավասարության երկու մասերի մաթեմատիկական սպասումը՝ կստանանք

$$\begin{aligned} E[f(X)] &\geq E[f(E[X])] + E[f'(E[X]), X - E[X]] = \\ &= f(E[X]) + (f'(E[X]), E[X] - E[X]) = f(E[X]): \end{aligned}$$

Նշենք, որ եթե f -ը գոգավոր ֆունկցիա է, ապա առկա է հակառակ անհավասարությունը՝

$$E[f(X)] \leq f(E[X]):$$

Թեորեմ 2.2. (Չերիշևի անհավասարությունը): Կամայական ոչ բացասական g ֆունկցիայի և $\epsilon > 0$ թվի համար տեղի ունի հետևյալ անհավասարությունը՝

$$P(g(X) > \epsilon) \leq \frac{E[g(X)]}{\epsilon}:$$

Ապացույց: Եթե $E[g(X)] = +\infty$, ապա անհավասարությունը ակնհայտ է:

Դիտարկենք այն դեպքը, երբ մաթեմատիկական սպասումը վերջավոր թիվ է: Այդ դեպքում կունենանք

$$\begin{aligned} E[g(X)] &= \int_R g(x)f(x)dx = \int_{g(x) < \epsilon} g(x)f(x)dx + \int_{g(x) \geq \epsilon} g(x)f(x)dx \geq \\ &\geq \epsilon \int_{g(x) > \epsilon} f(x)dx = \epsilon P(g(X) > \epsilon): \end{aligned}$$

Այստեղից անմիջապես բխում է Չերիշևի անհավասարությունը:

Հետևանք 1: Ճշմարիտ է հետևյալ անհավասարությունը՝

$$E[|X - EX| > \epsilon] \leq \frac{D[X]}{\epsilon^2}:$$

Ապացույց: Ունենք

$$P(|X - E[X]| > \epsilon) = P(|X - E[X]|^2 > \epsilon^2) \leq \frac{D[X]}{\epsilon^2}:$$

Հետևանք 2: Դիցուք Z -ը պատահական մեծություն է, որը արժեքներ է ընդունում $[0, 1]$ հատվածից: Ենթադրենք նաև, որ $E[Z] = \theta$: Այդ դեպքում կամայական $a \in (0, 1)$ թվի համար տեղի ունի

$$P[Z > a] \geq \frac{\theta - a}{1 - a}$$

անհավասարությունը:

Ապացույց: Դիտարկենք $Y = 1 - Z$ պատահական մեծությունը: Քանի որ $Y \geq 0$ և $E[Y] = 1 - E[Z] = 1 - \theta$; ապա կիրառելով Չերիշևի անհավասարությունը՝ կստանանք

$$P[Z \leq 1 - a] = P[1 - Z \geq a] = P[Y \geq a] \leq \frac{E[Y]}{a} = \frac{1 - \theta}{a}:$$

Այստեղից

$$P[Z > 1 - a] \geq 1 - \frac{1 - \theta}{a} = \frac{a + \theta - 1}{a}:$$

Դրանից անմիջապես հետևում է, որ

$$P[Z > a] \geq \frac{\theta - a}{1 - a} :$$

Թեորեմ 2.3. (Չերիշևի մեծ թվերի օրենքը): Դիցուք ունենք իրարից անկախ X_i , $i = 1, 2, \dots$ պատահական մեծությունների հաջորդականություն, երբ

$$E[X_i] = a, \quad D[X_i] = \sigma, \quad i = 1, 2, \dots :$$

Այդ դեպքում

$$\lim_{n \rightarrow \infty} P(|\bar{X}_n - a| > \epsilon) = 0,$$

որտեղ

$$\bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n} :$$

Ապացույց: Կիրառենք Չերիշևի անհավասարությունը $g(X) = (\bar{X}_n - a)^2$ պատահական մեծության նկատմամբ, կստանանք

$$P(|\bar{X}_n - a| > \epsilon) = P(|\bar{X}_n - a|^2 > \epsilon^2) \leq \frac{E[(\bar{X}_n - a)^2]}{\epsilon^2} :$$

Ունենք նաև

$$E[(\bar{X}_n - a)^2] = E\left[\frac{1}{n} \sum_{i=1}^n (X_i - a)^2\right] = \frac{1}{n^2} \sum_{i=1}^n D[X_i] = \frac{\sigma^2}{n} :$$

Հետևաբար

$$P(|\bar{X}_n - a| > \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \rightarrow 0, \text{ երբ } n \rightarrow \infty :$$

6. Հոֆդինգի անհավասարությունը:

Թեորեմ 2.4. Դիցուք ունենք միևնույն ձևով բաշխված պատահական մեծությունների $Z_1, \dots, Z_l$ հաջորդականություն, այնպիսին, որ

$$E[\bar{Z}] = \beta, \quad P(Z_i \in [a, b]) = 1, \quad i = 1, 2, \dots, l, \quad \bar{Z} = \frac{1}{l} \sum_{i=1}^l Z_i :$$

Այդ դեպքում կամայական դրական $\epsilon > 0$ թվի համար տեղի ունի հետևյալ անհավասարությունը՝

$$P\left(\left|\frac{1}{l} \sum_{i=1}^l Z_i - \beta\right| > \epsilon\right) \leq \exp(-2l\epsilon^2/(b-a)^2) :$$

Թեորեմի ապացույցը հենվում է հետևյալ պնդման վրա:

Լեմմա 2.1. Դիցուք X -ը պատահական մեծություն է, որը ընդունում է արժեքներ $[a, b]$ հատվածից և $E[X] = 0$: Այդ դեպքում ցանկացած $\lambda > 0$ դրական թվի համար տեղի ունի հետևյալ անհավասարությունը՝

$$E[e^{\lambda X}] \leq e^{\lambda^2(b-a)^2/8} :$$

Ապացույց: Քանի որ $f(x) = e^{\lambda x}$ ֆունկցիան ուռուցիկ է, ապա կամայական $\alpha \in (0, 1)$ և $x \in [a, b]$ թվերի համար տեղի ունի $f(x) \leq \alpha f(a) + (1 - \alpha)f(b)$ անհավասարությունը: Այս անհավասարությունում տեղադրելով

$$\alpha = \frac{b - x}{b - a} \in (0, 1)$$

կստանանք

$$e^{\lambda x} \leq \frac{b - x}{b - a} e^{\lambda a} + \frac{x - a}{b - a} e^{\lambda b}:$$

Այնուհետև վերցնելով այս անհավասարության երկու մասի մաթեմատիկական սպասումները՝ կունենանք

$$\begin{aligned} E[e^{\lambda X}] &\leq \frac{b - E[X]}{b - a} e^{\lambda a} + \frac{E[X] - a}{b - a} e^{\lambda b} = \\ &= \frac{b}{b - a} e^{\lambda a} - \frac{a}{b - a} e^{\lambda b}: \end{aligned}$$

Նշանանակենք $h = \lambda(b - a)$, $p = -a(b - a)$, $L(h) = -hp + \ln(1 - p + pe^h)$: Այս դեպքում անհավասարության աջ մասը կունենա $e^{L(h)}$ տեսքը: Ապացույցի ավարտի համար բավական է ցույց տալ, որ $L(h) \leq h^2/8$:

Քանի որ

$$L(0) = L'(0) = 0, \quad L''(h) \leq \frac{1}{4},$$

ապա վերլուծելով $L(h)$ ֆունկցիան ըստ Թեյլորի բանաձևի՝ կստանանք $L(h) \leq h^2/8$ անհավասարությունը:

Այժմ անցնենք Հոֆդինգի անհավասարության ապացույցին: Նշանակենք

$$X_i = Z_i - E[Z_i], \quad \bar{X} = \frac{1}{l} \sum_{i=1}^l X_i:$$

Համաձայն Չերիշևի անհավասարության՝ ունենք

$$P(\bar{X} \geq \epsilon) = P(e^{\lambda \bar{X}} \geq e^{\lambda \epsilon}) \leq e^{-\lambda \epsilon} E[e^{\lambda \bar{X}}]:$$

Քանի որ X_i , $i = 1, 2, \dots, l$ պատահական մեծությունները անկախ են, ապա

$$E[e^{\lambda \bar{X}}] = E\left[\prod_i e^{\lambda X_i/l}\right] = \prod_i E[e^{\lambda X_i/l}]:$$

Կիրառելով այստեղ լեմմա 2.1-ը՝ կստանանք

$$P(\bar{X} \geq \epsilon) \leq e^{-\lambda \epsilon} \prod_i e^{\lambda^2(b-a)^2/(8l^2)} \leq e^{-\lambda \epsilon + \lambda^2(b-a)^2/8l}:$$

Այնուհետև տեղադրելով այստեղ $\lambda = 4l\epsilon/(b-a)^2$, կստանանք

$$P(\bar{X} \geq \epsilon) \leq e^{-\frac{2l\epsilon^2}{(b-a)^2}}:$$

Անալոգ դատողություններ անելով $-\overline{X}$ պատահական մեծության նկատմամբ՝ կստանանք

$$P(\overline{X} \leq -\epsilon) \leq e^{-\frac{2\epsilon^2}{(b-a)^2}}:$$

Հետևաբար՝ ստացված այս երկու անհավասարություններից ամփոփապես բխում է Հոֆդինգի անհավասարությունը:

7. Սահմանային թեորեմներ, նորմալ բաշխում: Դիցուք փորձի արդյունքում հնարավոր է երկու ելք՝ հանդես է գալիս A պատահարը կամ նրա հակադիր պատահարը: Դիցուք կատարվում են n անկախ փորձեր: Ենթադրենք յուրաքանչյուր փորձում A պատահարի հանդես գալու հավանականությունը հավասար է p -ի՝ $P(A) = p$: Այդ դեպքում հակադիր պատահարի հանդես գալու հավանականությունը հավասար կլինի $(1-p)$: Հետևաբար n փորձերի ընթացքում A պատահարի k անգամ հանդես գալու հավանականությունը հավասար կլինի

$$P_n(k) = C_n^k p^k (1-p)^{n-k}:$$

Այս բանաձևը կրում է Է Բեռնուլի անունը կամ կոչվում է բինոմական բաշխման բանաձև:

Թեորեմ 2.5. Դիցուք X պատահական մեծությունը ունի բինոմական բաշխում: Այդ դեպքում

$$\lim_{n \rightarrow \infty} P\left(\left|\frac{X}{n} - p\right| > \epsilon\right) = 0:$$

Ապացույց: Հայտնի է, որ բինոմական բաշխման ժամանակ $E[X] = np$, $D[X] = np(1-p)$: Կիրառելով Չեբիշևի անհավասարությունը՝ կստանանք

$$P\left(\left|\frac{X}{n} - p\right| \geq \epsilon\right) \leq \frac{D[X/n]}{\epsilon^2} = \frac{np(1-p)/n^2}{\epsilon^2} = \frac{p(1-p)}{n\epsilon^2} \rightarrow 0, \text{ երբ } n \rightarrow \infty:$$

Տեղի ունի նաև սահմանային հեռույալ թեորեմը:

Թեորեմ 2.6. Դիցուք X պատահական մեծությունը ունի բինոմական բաշխում: Այդ դեպքում

$$\lim_{n \rightarrow \infty} P\left(a \leq \frac{X - np}{\sqrt{np(1-p)}} \leq b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2} dx:$$

8. Պայմանական մաթեմատիկական սպասումներ: Դիցուք ունենք պատահական մեծությունների զույգ (X, Y) : Նշանակենք

$$f^{Y|X}(y|x) = P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)} = \frac{f^{X,Y}(x, y)}{f^X(x)}:$$

Սա պայմանական հավանականությունն է այն բանի, որ Y պատահական մեծությունը ընդունում է y արժեք այն պայմանով, որ X -ը ընդունում է x արժեքը:

Y պատահական մեծության պայմանական մաթեմատիկական սպասում այն

պայմանով, որ X -ը ընդունում է x արժեք, կոչվում է հետևյալ մեծությունը՝

$$E[Y|X = x] = \int_R f^{Y|X}(y|x)y dy:$$

Եթե X և Y դիսկրետ պատահական մեծություններ են, ապա

$$E[Y|X = x] = \sum_y f^{Y|X}(y|x)y:$$

Տեղի ունի հետևյալ պնդումը:

Թեորեմ 2.7. $E[Y|X]$ պայմանական մաթեմատիկական սպասումը բավարարում է հետևյալ պայմանները:

1. $E^X E[Y|X] = E[Y]$:
2. Եթե X և Y պատահական մեծությունները անկախ են, ապա

$$E[Y|X] = E[Y]:$$

Ապացույց: Թեորեմի առաջին կետի ապացույցը կատարենք դիսկրետ դեպքի համար: Ունենք

$$\begin{aligned} E^X E[Y|X] &= \sum_x E[Y|X = x]P(X = x) = \\ &= \sum_x \sum_y y f^{Y|X}(y|x)P(X = x) = \sum_y y \sum_x f^{Y|X}(y|x)P(X = x) = \sum_y y P(Y = y) = E[Y]: \end{aligned}$$

Թեորեմի առաջին կետի ապացույցը անմիջապես բխում է այն փաստից, որ եթե X և Y պատահական մեծությունները անկախ են, ապա

$$P(Y = y|X = x) = P(Y = y):$$

2.2 Պատահական մեծությունների կովարիացիա

Դիցուք X -ը և Y -ը պատահական մեծություններ են: Այս պատահական մեծությունների կովարիացիա է կոչվում հետևյալ թիվը՝

$$Cov(X, Y) = E[(X - E[X])(Y - E[Y])]:$$

Ցույց տանք, որ այս մեծությունը կարելի է հաշվել հետևյալ բանաձևով՝

$$Cov(X, Y) = E[XY] - E[X]E[Y]:$$

Իրոք, ունենք

$$\begin{aligned} Cov(X, Y) &= E[(X - EX)(Y - EY)] = E[XY - XE[Y] + E[X]E[Y] - YE[X]] = \\ &= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y] = \end{aligned}$$

$$= E[XY] - E[X]E[Y]:$$

Նշենք կովարիացիայի մի քանի հատկություններ:

- Դիցուք $X_1, X_2, \dots, X_n$ պատահական մեծություններ են, իսկ

$$A = \sum_{i=1}^n \alpha_i X_i, \quad B = \sum_{i=1}^m \beta_i X_i$$

նրանց գծային կոմբինացիաներն են: Այդ դեպքում

$$Cov(A, B) = \sum_{i=1}^n \sum_{j=1}^m \alpha_i \beta_j Cov(X_i, X_j):$$

- $Cov(X, X) = E[X^2] - (E[X])^2$, որը X պատահական մեծության դիսպերսիան է:

Պնդում: Եթե X -ը և Y -ը անկախ պատահական մեծություններ են, ապա

$$Cov(X, Y) = 0:$$

Ապացույց: Ունենք

$$Cov(X, Y) = E[XY] - E[X]E[Y] = E[X]E[Y] - E[X]E[Y] = 0:$$

Այստեղ օգտվեցինք այն փաստից, որ անկախ պատահական մեծությունների դեպքում

$$E[XY] = E[X]E[Y]:$$

Դիցուք ունենք $X = (x_1, x_2, \dots, x_n)$, $Y = (y_1, y_2, \dots, y_n)$ պատահական վեկտորներ: Այս պատահական վեկտորների կովարիացիա է կոչվում հետևյալ մատրիցը՝

$$Var(X, Y) \equiv Cov(X, Y) = E[(X - E[X])(Y - E[Y])^T]:$$

Բացված տեսքով այն կունենա հետևյալ կառուցվածքը՝

$$Var(X, Y) = \begin{pmatrix} E[(x_1 - E[x_1])(y_1 - E[y_1])] & \dots & E[(x_1 - E[x_1])(y_n - E[y_n])] \\ E[(x_2 - E[x_2])(y_1 - E[y_1])] & \dots & E[(x_2 - E[x_2])(y_n - E[y_n])] \\ \dots & \dots & \dots \\ E[(x_n - E[x_n])(y_1 - E[y_1])] & \dots & E[(x_n - E[x_n])(y_n - E[y_n])] \end{pmatrix}:$$

Պատահական վեկտորների կովարիացիոն մատրիցը օժտված է հետևյալ հատկություններով՝

- $Var(X) \geq 0$ (դրական կիսատրոշյալ մատրից):
- $Var(X, Y_1 + Y_2) = Var(X, Y_1) + Var(X, Y_2)$:
- $Var(X, Y) = 0 \rightarrow Var(X + Y) = Var(X) + Var(Y)$:

2.3 Օպտիմալ և կոռեկտ պրեդիկտորներ

Դիցուք X -ը օբյեկտների բազմությունն է, իսկ Y -ը՝ նշիչների (Label, Markers): Այս երկուսի դեկարտյան արտադրյալը՝ $Z=X \times Y$, կանվանենք նմուշների բազմություն:

$h: X \rightarrow Y$ արտապատկերումը կոչվում է պրեդիկտոր կամ հիպոթեզ: Դիցուք H -ը պրեդիկտորների (հիպոթեզների) բազմությունն է: Կամայական ոչ բացասական

$$l: H \times Z \rightarrow R_+$$

ֆունկցիա անվանենք կորստի ֆունկցիա: Ենթադրենք՝ օբյեկտների Z բազմության վրա տրված է հավանականային D բաշխում:

$h \in H$ հիպոթեզի ռիսկ $l(h, z)$ կորստի ֆունկցիայի դեպքում կոչվում է հետևյալ մաթեմատիկական սպասումը՝

$$L_D(h) \equiv E_{z \sim D}[l(h, z)]:$$

Նմանապես, եթե ունենք օբյեկտների $S = \{z_1, z_2, \dots, z_m\} \in Z$ հավաքածու, ապա $h \in H$ հիպոթեզի էմպիրիկ (փորձնական) ռիսկ սահմանվում է հետևյալ բանաձևով՝

$$L_S(h) = \frac{1}{m} \sum_{i=1}^m l(h, z_i):$$

Դիցուք ինչ-որ կանոնով նմուշների ցանկացած $S = \{z_1, z_2, \dots, z_m\}$ հավաքածուին համապատասխանության մեջ է դրվում որևէ պրեդիկտոր $h \in H$: Այդ համապատասխանությունը կոչվում է ալգորիթմ և նշանակվում է հետևյալ կերպ՝

$$A: S \rightarrow H:$$

Այսինքն՝ ալգորիթմը արտապատկերում է, որը S բազմության յուրաքանչյուր էլեմենտին համապատասխանեցնում է H բազմության ինչ-որ էլեմենտ: S -ը կոչվում է ալգորիթմի մուտք, իսկ $A(S) \in H$ -ը՝ ելք:

Որպես ալգորիթմի նմուշ կարող ենք դիտարկել հետևյալը՝

$$A(S) = \arg \min_{h \in H} L_S(h):$$

Այն սովորաբար նշանակվում է ERM_H սիմվոլով, և մենք հետագայում կընդունենք $h_S \equiv ERM_H(S)$ գրառումը:

Սահմանում: Դիցուք H -ը հիպոթեզների, իսկ Z -ը նմուշների բազմություններ են, $l(h, z)$ կորստի ֆունկցիա է: Ենթադրենք՝ գոյություն ունեն $m_H: (0, 1)^2 \rightarrow N$ ֆունկցիա և A ալգորիթմ՝ բավարարող հետևյալ պայմանը. ցանկացած $\epsilon, \delta \in (0, 1)$ թվերի, կամայական հավանականային D բաշխման ու ցանկացած պատահական $S = \{z_1, z_2, \dots, z_l\}$, $l \geq m_H(\epsilon, \delta)$ հավաքածուի համար, որը տրվում է A ալգորիթմի

մուտքին, այն վերադարձնում է $h \equiv A(S) \in H$ հիպոթեզ, այնպիսին, որ տեղի ունի

$$P(E_{z \sim D} l(A(S), z) \leq \min_{h \in H} E_{z \sim D} l(h, z) + \epsilon) > 1 - \delta$$

անհավասարությունը: Կամ՝ որ նույն է՝ առնվազն $1 - \delta$ հավանականությամբ տեղի ունի համարժեք հետևյալ անհավասարությունը՝

$$L_D(A(S)) \leq \min_{h \in H} L_D(h) + \epsilon:$$

Այդ դեպքում պրեդիկտորների H դասը կոչվում է օպտիմալ, իսկ $A(S)$ հիպոթեզը՝ օպտիմալ պրեդիկտոր նշված դասում:

Հարց է առաջանում՝ պրեդիկտորների ինչպիսի դասերի համար գոյություն ունեն օպտիմալ պրեդիկտորներ և ինչպես դրանք կառուցել: Այս հարցը մեքենայական ուսուցման հիմնական խնդիրն է: Այս դասընթացում ճշտվում են այդ դասերը, և տրվում է օպտիմալ պրեդիկտորների կառուցման գրադիենտային և սուբգրադիենտային եղանակների լրիվ և ամբողջական շարադրանքը:

Այժմ քննարկենք պարզ դեպք, երբ պրեդիկտորների բազմությունը վերջավոր է:

Թեորեմ 2.8. Եթե հիպոթեզների H բազմությունը վերջավոր է և $l(h, z) \in [0, 1]$, ապա այդ դասում գոյություն ունի օպտիմալ պրեդիկտոր:

Նախ ցույց տանք հետևյալ պնդումը:

Լեմմա 2.2. Դիցուք $\epsilon, \delta \in (0, 1)$ ֆիքսած թվեր են: Այդ դեպքում գոյություն ունի այնպիսի m_H թիվ, որ $m \geq m_H$ երկարությամբ կամայական $S = \{z_1, z_2, \dots, z_m\}$ նմուշների հավաքածուի և D հավանականային բաշխման համար տեղի ունի

$$P\{S : \forall h \in H, |L_S(h) - L_D(h)| \leq \epsilon\} \geq 1 - \delta$$

անհավասարությունը:

Ապացույց: Ունենք

$$\{S : \exists h \in H, |L_S(h) - L_D(h)| > \epsilon\} = \bigcup_{h \in H} \{S : |L_S(h) - L_D(h)| > \epsilon\}:$$

Այստեղից

$$P(\{S : \exists h \in H, |L_S(h) - L_D(h)| > \epsilon\}) \leq \sum_{h \in H} P(\{S : |L_S(h) - L_D(h)| > \epsilon\}):$$

Ցույց տանք, որ այս անհավասարության աջ մասի յուրաքանչյուր գումարելի ֆիքսած h -ի դեպքում ձգտում է զրոյի, երբ $m \rightarrow \infty$: Հիշենք, որ

$$L_D(h) = E_{z \sim P} l(h, z), \quad L_S(h) = \frac{1}{m} \sum_{i=1}^m l(h, z_i):$$

Քանի որ z_i պատահարները իրարից անկախ են և ունեն միևնույն հավանա-

կանային բաշխում, ապա $l(h, z_i)$ պատահական մեծության մաթեմատիկական սպասումը հավասար է $L_D(h)$ -ի: Մաթեմատիկական սպասման գծային հատկությունից հետևում է նաև, որ $L_D(h)$ -ը $L_S(h)$ -ի մաթեմատիկական սպասումն է: Հետևաբար $|L_D(h) - L_S(h)|$ տարբերությունը $L_S(h)$ պատահական մեծության և նրա մաթեմատիկական սպասման շեղումն է: Գնահատենք այս շեղման մեծությունը Հոֆդինգի անհավասարության միջոցով:

Պնդում (Հոֆդինգի անհավասարությունը): Դիցուք $\theta_1, \dots, \theta_m$ իրարից անկախ միևնույն բաշխումը ունեցող պատահական մեծություններ են: Բացի դրանից ենթադրենք, որ

$$E[\theta_i] = \alpha, \quad P(a \leq \theta_i \leq b) = 1:$$

Այդ դեպքում կամայական $\epsilon > 0$ թվի համար տեղի ունի

$$P(|\frac{1}{m} \sum_{i=1}^m \theta_i - \alpha| > \epsilon) \leq 2 \exp(-2m\epsilon^2/(b-a)^2)$$

անհավասարությունը:

Դիցուք մեր օրինակում $\theta_i = l(h, z_i)$: Այդ դեպքում

$$L_S(h) = \frac{1}{m} \sum_{i=1}^m \theta_i, \quad L_D(h) = \alpha:$$

Ըստ ենթադրության նաև $l(h, z)$ ֆունկցիայի արժեքների տիրույթը $[0, 1]$ հատվածն է: Այդ դեպքում, ըստ Հոֆդինգի անհավասարության, կունենանք

$$P(S : |L_D(h) - L_S(h)| > \epsilon) = P(|\frac{1}{m} \sum_{i=1}^m \theta_i - \alpha| > \epsilon) \leq 2 \exp(-2m\epsilon^2):$$

Այստեղից կստանանք

$$P(\{S : \exists h \in H, |L_S(h) - L_D(h)| > \epsilon\}) \leq \sum_{h \in H} 2 \exp(-2m\epsilon^2) = 2|H| \exp(-2m\epsilon^2):$$

Եթե m -ը ընտրենք այնպես, որ

$$m \geq \frac{\ln(2|H|/\delta)}{2\epsilon^2},$$

ապա

$$P(S : \exists h \in H, |L_S(h) - L_D(h)| > \epsilon) \leq \delta:$$

Լեմման ապացուցվեց:

Թեորեմի պնդումը անմիջապես հետևում է այս լեմմայից: Իրոք, նշանակենք

$$h_S = \arg \min_{h \in H} L_S(h):$$

Այդ դեպքում ցանկացած $h \in H$ համար, ըստ լեմմայի պնդման, կունենանք

$$\begin{aligned} L_D(h_S) &\leq L_S(h_S) + \frac{\epsilon}{2} \leq L_S(h) + \frac{\epsilon}{2} \leq \\ &\leq L_D(h) + \frac{\epsilon}{2} + \frac{\epsilon}{2} = L_D(h) + \epsilon : \end{aligned}$$

Վերջում նշենք, որ թեորեմի պայմաններին բավարարում է կորստի $l(h, z) = l(h, (x, y)) \equiv [h(x) \neq y]$ ֆունկցիան, որը ընդունում է 1 արժեք, եթե միջակ փակագծի արտահայտությունը ճիշտ է, հակառակ դեպքում՝ 0: Այդ դեպքում՝

$$L_D(h) = P_{(x,y) \sim D}[h(x) \neq y], \quad L_S(h) = \frac{|\{i \in [m] : h(x) \neq y\}|}{m} :$$

Ավելացնենք նաև, որ այս դեպքում որպես A ալգորիթմ հանդես է գալիս այն կանոնը, որը մինիմիզացնում է էմպիրիկ $L_S(h)$ ռիսկը՝ ելքում տալով $h_S = \arg \min_{h \in H} L_S(h)$ հիպոթեզը: Թեորեմն ապացուցվեց:

Այժմ տանք հավանականային իմաստով կոռեկտ պրեդիկտորների բազմության սահմանումը:

Սահմանում: Հիպոթեզների H բազմությունը կոչվում է PAC -կոռեկտ (probably approximatly correct), եթե գոյություն ունեն $m_H : (0, 1)^2 \rightarrow N$ ֆունկցիա և A ալգորիթմ, որոնք բավարարում են հետևյալ պայմանները:

- Դիցուք $\epsilon, \delta \in (0, 1)$ կամայական թվեր են,
- D -ն հավանականային բաշխում է նմուշների Z բազմության վրա, իսկ $f : X \rightarrow \{0, 1\}$ օբյեկտների նշիչների այնպիսի ֆունկցիա է, որ $L_D(f) = 0$:

Այդ դեպքում ցանկացած $S = \{z_1, \dots, z_m\}$, $m \geq m_H$ հավաքածուի դեպքում ($z_i = (x_i, f(x_i))$, $i = 1, 2, \dots, m$), որը տրվում է A ալգորիթմի մուտքին, ապա ելքում ալգորիթմը վերադարձնում է այնպիսի $h = A(S) \in H$ հիպոթեզ, որը առնվազն $1 - \delta$ հավանականությամբ բավարարում է $L_D(A(S)) \leq \epsilon$ անհավասարությանը:

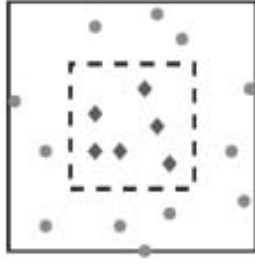
Այս անհավասարությանը բավարարող $A(S)$ պրեդիկտորը ինչպես նաև A ալգորիթմը կոչվում են PAC -կոռեկտ:

Հեշտ է նկատել, որ եթե հիպոթեզների H բազմությունը օպտիմալ է, ապա այն կոռեկտ է:

Այժմ բերենք մի օրինակ, որը ցույց է տալիս, որ ոչ բոլոր H դասերն են PAC -կոռեկտ: Դիտարկենք տվյալներ, որոնք պատկերված են 2.1 նկարում: Ենթադրենք կետերը բաշխված են հավասարաչափ քառակուսու ներսում, իսկ նշիչների f ֆունկցիան գործում է հետևյալ կերպ. « 1 » նշիչ ստացել են այն օբյեկտները, որոնք քառակուսու ներսում են, մնացածները ունեն « 0 » նշիչներ: Մեծ քառակուսու մակերեսը հավասար է 2-ի, իսկ փոքրիկը՝ 1-ի:

Դիտարկենք կամայական S հավաքածու: Որպես կորստի $l(h, z)$ ֆունկցիա վերցնենք հետևյալը՝

$$l(h, z) = [h(x) \neq y]:$$



Նկ. 2.1: Գոյություն չունի օպտիմալ պրեդիկտոր

Այդ դեպքում ակնհայտ է, որ

$$h_S(x) = \begin{cases} y_i, \exists i \in [1 : m] : x = x_i, \\ 0, \text{ հակառակ դեպքում} \end{cases}$$

պրեդիկտորը բավարարում է $h_S(x) = \arg \min_{h \in H} L_S(h)$, $L_S(h_S) = 0$ պայմանները: Այսինքն՝ Էմպիրիկ սխալանքը հավասար է զրոյի: Բայց իրական սխալանքը $E_{z \sim P}(l(h_S, z)) = 1/2$: Հետևաբար նշված դասը PAC-կոռեկտ չէ:

Նշենք, որ օպտիմալ պրեդիկտորներ գոյություն ունեն H հիպոթեզների որոշ անվերջ դասերի համար:

Այժմ դիտարկենք հիպոթեզների անվերջ մի դաս, որի համար օպտիմալ հիպոթեզ գոյություն ունի:

Դիտարկենք աստիճանային ֆունկցիաների դաս թվային առանցքի վրա՝

$$\{h_a : a \in R\},$$

որտեղ h_a -ն որոշվում է հետևյալ կանոնով՝ $h_a : R \rightarrow \{0, 1\}$, $h_a(x) = 1_{x < a}$: Հիշեցնենք, որ $h_a(x) = 1$, եթե $x < a$, հակառակ դեպքում՝ 0:

Թեորեմ 2.9. Դիցուք նմուշների S բազմության վրա տրված է հավանականային D բաշխում, և գոյություն ունի այնպիսի $h^* \in H$ հիպոթեզ, որ $L_D(h^*) = 0$: Այդ դեպքում ցանկացած $\epsilon, \delta \in (0, 1)$ թվերի համար տեղի ունի

$$P(L_S(h_S) \leq \epsilon) > 1 - \delta, \quad S = \{z_1, z_2, \dots, z_m\}, \quad m \geq \ln(1/\delta)/\epsilon$$

անհավասարությունը, որտեղ $h_S = \text{ERM}_H(S)$:

Ապացույց: Դիցուք h^* -ին համապատասխանող աստիճանային ֆունկցիան է $1_{x < a^*}$ -ը: Ենթադրենք նաև, որ (a_0, a_1) այն միջակայքն է, որ

$$a_0 < a^* < a_1, \quad P_{x \sim D_x}(x \in (a_0, a^*)) = P_{x \sim D_x}(x \in (a^*, a_1)) = \epsilon:$$

Տվյալների S հավաքածուի համար նշանակենք

$$b_0 = \max\{x : (x, 1) \in S\}, \quad b_1 = \min\{x : (x, 0) \in S\}:$$

Դիցուք b_S -ը h_S հիպոթեզին համապատասխանող աստիճանային ֆունկցիան է: Այստեղից $b_s \in (b_0, b_1)$: Հետևաբար որպեսզի կատարվի $L_D(h_S) \leq \epsilon$ անհավասարությունը, բավական է, որ

$$b_0 \geq a_0, \quad b_1 \leq a_1 :$$

Ուրիշ խոսքով՝

$$P_{S \sim D^m}(L_D(h_S)) > \epsilon \leq P_{S \sim D^m}(b_0 < a_0) + P_{S \sim D^m}(b_1 > a_1):$$

$b_0 < a_0$ պատահարը տեղի ունի միայն այն դեպքում, երբ հավաքածուի բոլոր օբյեկտները գտնվում են (a_0, a^*) միջակայքից դուրս: Այստեղից

$$P_{S \sim D^m}(b_0 < a_0) = P_{S \sim D^m}(\forall (x, y) \in S, \quad x \notin (a_0, a^*)) = (1 - \epsilon)^m \leq e^{-\epsilon m} :$$

Քանի որ ըստ ենթադրության $m > \ln(2/\delta)/\epsilon$, հետևաբար՝

$$e^{-\epsilon m} \leq \delta/2:$$

Նման կերպ ապացուցվում է, որ

$$P_{S \sim D^m}(b_1 > a_1) \leq \delta/2:$$

Վերջնականում կստանանք

$$P_{S \sim D^m}(L_D(h_S)) > \epsilon \leq \delta:$$

Թեորեմ 2.10. Դիցուք X -ը անվերջ բազմություն է, իսկ H -ը բոլոր ֆունկցիաներ են, որոնք գործում են X -ից $\{0, 1\}$ բազմության վրա: Այդ դեպքում H -ը PAC-կոռեկտ չէ:

Նախ ապացուցենք հետևյալ պնդումը:

Լեմմա 2.3. Դիցուք X -ը անվերջ բազմություն է, իսկ A -ն՝ քինար դասակարգման ալգորիթ: Դիցուք m -ը բնական թիվ է: Այդ դեպքում գոյություն ունի այնպիսի D բաշխում $X \times \{0, 1\}$ բազմության վրա, որի համար բավարարվում են հետևյալ պայմանները՝

- գոյություն ունի այնպիսի $f : X \rightarrow \{0, 1\}$ ֆունկցիա, որ $L_D(f) = 0$:
- $P_{S \sim D^m}[L_D(A(S)) \geq 1/8] > 1/7$:

Ապացույց: Դիցուք $C \subset X$, $|C| = 2m$: Այդ դեպքում գոյություն ունեն $T = 2^{2m}$ քանակով ֆունկցիաներ, որոնք գործում են C -ից $\{0, 1\}$ բազմության վրա: Նշանակենք այդ ֆունկցիաները $f_1, f_2, \dots, f_T$: Յուրաքանչյուր f_i ֆունկցիայի դեպքում $C \times \{0, 1\}$ բազմության վրա սահմանենք հավանականային D_i բաշխում

հետևյալ բանաձևով՝

$$D_i(x, y) = \begin{cases} \frac{1}{|C|}, & \text{երթե } y = f_i(x) \\ 0, & \text{հակառակ դեպքում:} \end{cases}$$

Սա նշանակում է, որ (x, y) գույգը ընտրելու հավանականությունը հավասար է $1/|C|$, երթե y -ը համընկնում է f_i -ի արժեքի հետ, հակառակ դեպքում՝ 0: Այստեղից հետևում է, որ $L_{D_i}(f_i) = 0$: Հետևաբար թեորեմի առաջին եզրակացությունը կատարվում է բոլոր D_i բաշխումների համար: Այժմ ցույց տանք, որ երկրորդ պայմանը կատարվում է այս բաշխումներից ինչ-որ մեկի համար: Դրա համար բավական է ցույց տալ, որ

$$\max_{i \in [T]} E_{S \sim D_i^m} [L_{D_i}(A(S))] \geq \frac{1}{4}, \quad (2.1)$$

որտեղ S -ը m երկարության օրինակների հաջորդականություն է, որը տրվում է A ալգորիթմի մուտքին, իսկ այն վերադարձնում է $A(S) : C \rightarrow \{0, 1\}$ ֆունկցիա: Այստեղ կիրառենք

$$P[Z > a] \geq \frac{E[Z]}{1 - a}$$

անհավասարությունը, որը ապացուցվել է Չեբիշևի անհավասարության երկրորդ հետևանքում: Հետևաբար՝

$$P_{S \sim D_i^m} [L_{D_i}(AS) > 1/8] \geq \frac{1/4 - 1/8}{1 - 1/8} = \frac{1}{7}:$$

Այժմ ապացուցենք (2.1) անհավասարությունը: m երկարության բոլոր S հաջորդականությունների քանակը հավասար է $k = (2m)^m$: Նշանակենք այդ հաջորդականությունները՝ $\{S_1, S_2, \dots, S_k\}$: Դիցուք

$$S_i^i = \{(x_1, f_i(x_1)), (x_2, f_i(x_2)), \dots, (x_m, f_i(x_m))\} :$$

Եթե ֆիքսենք D_i բաշխումը, ապա նրան համապատասխանող հավաքածուներ կլինեն՝

$$S_i^i, S_2^i, \dots, S_k^i,$$

և դրանցից յուրաքանչյուրի հանդես գալու հավանականությունը նույնն է: Հետևաբար՝

$$E_{S \sim D^m} [L_{D_i}(A(S))] = \frac{1}{k} \sum_{j=1}^k L_{D_i}(A(S_j^i)): \quad (2.2)$$

Քանի որ թվերի միջին արժեքը ընկած է դրանց մինիմալ և մաքսիմալ արժեքների միջև, ապա

$$\max_{i \in [T]} \frac{1}{k} \sum_{j=1}^k L_{D_i}(A(S_j^i)) \geq \frac{1}{T} \sum_{i=1}^T \frac{1}{k} \sum_{j=1}^k L_{D_i}(A(S_j^i)) =$$

$$= \frac{1}{k} \sum_{j=1}^k \frac{1}{T} \sum_{i=1}^T L_{D_i}(A(S_j^I)) \geq \min_{j \in [k]} \frac{1}{T} \sum_{i=1}^T L_{D_i}(A(S_j^I)) : \quad (2.3)$$

Դիցուք

$$S_j = \{x_1, x_2, \dots, x_m\}, \{v_1, v_2, \dots, v_p\} = C \setminus S_j :$$

Քանի որ S_j -ին օրինակների հաջորդականությունն է՝ ընտրված C բազմությունից, ապա ակնհայտ է, որ $p \geq m$: Հետևաբար կամայական $h : C \rightarrow \{0, 1\}$ ֆունկցիայի և յուրաքանչյուր i -ի դեպքում կունենանք

$$\begin{aligned} L_{D_i}(h) &= \frac{1}{2m} \sum_{x \in C} \mathbb{1}_{[h(x) \neq f_i(x)]} \geq \frac{1}{2m} \sum_{r=1}^p \mathbb{1}_{[h(v_r) \neq f_i(v_r)]} \\ &\geq \frac{1}{2p} \sum_{r=1}^p \mathbb{1}_{[h(v_r) \neq f_i(v_r)]} : \end{aligned} \quad (2.4)$$

Հետևաբար՝

$$\begin{aligned} \frac{1}{T} \sum_{i=1}^T L_{D_i}(A(S_j^I)) &\geq \frac{1}{T} \sum_{i=1}^T \frac{1}{2p} \sum_{r=1}^p \mathbb{1}_{[A(S_j^I)(v_r) \neq f_i(v_r)]} = \\ &= \frac{1}{2p} \sum_{r=1}^p \frac{1}{T} \sum_{i=1}^T \mathbb{1}_{[A(S_j^I)(v_r) \neq f_i(v_r)]} \geq \frac{1}{2} \min_{r \in [p]} \frac{1}{T} \sum_{i=1}^T \mathbb{1}_{[A(S_j^I)(v_r) \neq f_i(v_r)]} : \end{aligned} \quad (2.5)$$

Ֆիքսենք $r \in [p]$:

$f_1, f_2, \dots, f_T$ ֆունկցիաներից կազմենք $T/2$ քանակով զույգեր (f_i, f'_i) այն սկզբունքով, որ $f_i(c) \neq f'_i(c)$ միայն այն դեպքում, երբ $c = v_r$: Քանի որ այդ զույգերի համար ունենք $S_j^i = S_j^{i'}$, ապա

$$\mathbb{1}_{[A(S_j^i)(v_r) \neq f_i(v_r)]} + \mathbb{1}_{[A(S_j^{i'})(v_r) \neq f_i(v_r)]} = 1 :$$

Հետևաբար՝

$$\frac{1}{T} \sum_{i=1}^T \mathbb{1}_{[A(S_j^i)(v_r) \neq f_i(v_r)]} = \frac{1}{2} : \quad (2.6)$$

Ակնհայտորեն (2.4) – (2.6) անհավասարություններից բխում է (2.1) անհավասարությունը: Լեմման ապացուցվեց:

Այժմ անցնենք թեորեմ 2.10-ի ապացույցին: Ենթադրենք հակառակը, և դիցուք հիպոթեզների H բազմությունը PAC -կոռեկտ դաս է: Դա նշանակում է, որ $\epsilon < 1/8$, $\delta < 1/7$ թվերի համար գոյություն ունի A ալգորիթմ և $m = m(\epsilon, \delta)$ բնական թիվ, այնպիսին, որ կամայական D բաշխման դեպքում տրված $X \times \{0, 1\}$ բազմության վրա տեղի ունի հետևյալ պնդումը: Եթե որևէ $f : X \rightarrow \{0, 1\}$ ֆունկցիայի դեպքում $L_D(f) = 0$, ապա առնվազն $1 - \delta$ հավանականությամբ տեղի ունի

$$L_D(A(S)) \leq \epsilon, \quad S \sim D^m$$

անհավասարությունը: Բայց քանի որ $|X| > 2m$, ապա ըստ Լեմմա 2.1-ի, A ալգորիթմի համար գոյություն ունի D բաշխում, այնպիսին, որ $1/7 > \delta$

հավանականությամբ ճշմարիտ է

$$L_D(A(S)) > 1/8 > \epsilon$$

անհավասարությունը, ինչը հակասում է սկզբնական ենթադրությանը:

2.4 Հիպոթեզների բազմության VC չափողականությունը

Այս պարագրաֆում տրվում է հիպոթեզների H բազմության $VCdim(H)$ չափողականության գաղափարը: Այն ներմուծվել է [6] աշխատանքում (Vapnik):

Սահմանում: Դիցուք H -ը ֆունկցիաների դաս է՝ որոշված X բազմության վրա, և եթե $h \in H$, ապա $h : X \rightarrow \{0, 1\}$: Դիցուք

$$C = \{c_1, c_2, \dots, c_l\} \subset X, H_C = \{((h(c_1), h(c_2), \dots, h(c_l)), h \in H)\}:$$

Այն մաքսիմալ $|C|$ թիվը, որի համար տեղի ունի $|H_C| = 2^{|C|}$ պայմանը, կոչվում է H բազմության VC չափողականությունը և նշանակվում է հետևյալ կերպ՝ $VCdim(H)$:

Դիտարկենք որոշ օրինակներ:

Օրինակ 1: Որպես H բազմություն վերցնենք աստիճանաձև ֆունկցիաների բազմությունը, որոշված թվային առանցքի վրա՝ $h_a(x) \equiv \mathbb{1}_{x < a}$: Դիցուք $C = \{c_1\}$: Եթե վերցնենք $a = c_1 + 1$, ապա $h_a(c_1) = 1$, իսկ եթե $a = c_1 - 1$, ապա $h_a(c_1) = 0$: Դիտարկենք այժմ $C = \{c_1, c_2\}$, $c_1 < c_2$: Այս դեպքում ոչ մի աստիճանային $h_a(x)$ ֆունկցիայով $(0, 1)$ վեկտորը չի ստացվի, քանի որ եթե $h_a(c_1) = 0$, ապա $h_a(c_2) = 0$: Հետևաբար՝ $VCdim(H) = 1$:

Օրինակ 2: Դիցուք $H = \{h_{(a,b)}, a < b\}$, որտեղ

$$h_{(a,b)}(x) = \mathbb{1}_{x \in (a,b)}:$$

Դիցուք $C = \{1, 2\}$: Այդ դեպքում բոլոր

$$(0, 0), (0, 1), (1, 0), (1, 1)$$

վեկտորները «հասանելի» են: Օրինակ՝ $(1, 0)$ կստացվի, եթե վերցնենք $a = 0.5$, $b = 1.5$: Քանի որ $1 \in (0.5, 1.5)$, ապա $h_{a,b}(1) = 1$: Իսկ $h_{a,b}(2) = 0$, քանի որ $2 \notin (a, b)$: Դիցուք այժմ $C = \{c_1, c_2, c_3\}$, $c_1 < c_2 < c_3$: Այդ դեպքում հեշտ է տեսնել, որ $(1, 0, 1)$ վեկտորը հասանելի չէ ոչ մի (a, b) միջակայքով: Հետևաբար՝ $VCdim(H) = 2$:

Օրինակ 3: Դիցուք H -ը վերջավոր բազմություն է: Ակնհայտ է, որ կամայական C բազմության համար $|H_C| \leq |H|$: Այստեղից անմիջապես հետևում է, որ

$$VCdim(H) \leq \log_2 |H|:$$

Օրինակ 4: Դիտարկենք աֆինական ֆունկցիաների դասը՝

$$L_d = \{h_{w,b}, : w \in \mathbb{R}^d, b \in \mathbb{R}\},$$

որտեղ

$$h_{w,b}(x) = (w, x) + b = \sum_{i=1}^d w_i x_i + b:$$

Կատարենք հետևյալ նշանակումները՝

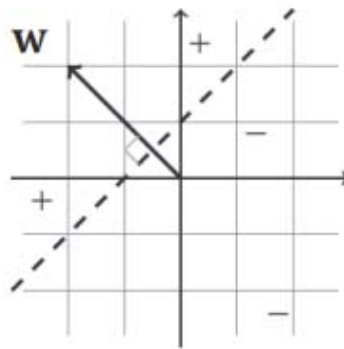
$$w' = (b, w_1, w_2, \dots, w_d), \quad x' = (1, x_1, x_2, \dots, x_d):$$

Այդ դեպքում աֆինական $h_{w,b}$ ֆունկցիան կարող ենք գրել $d + 1$ չափանի տարածության մեջ սկալյար արտադրյալի տեսքով՝ $h_{w,b} = (w', x')$: Բինար դասակարգման խնդրում ունենք $X = \mathbb{R}^d$, $Y = \{-1, 1\}$: Ենթատարածությունների դասը որոշվում է հետևյալ կերպ՝

$$HS_d = \text{sign} L_d = \{x \rightarrow \text{sign}(h_{w,b}(x)) : h_{w,b} \in L_d\}:$$

Այլ կերպ ասած՝ յուրաքանչյուր HS_d ենթատարածություն պարամետրիզացված է $w \in \mathbb{R}^d$ վեկտորով և b սկալյարով: Ստանալով x վեկտորը՝ հիպոթեզը վերադարձնում է $\text{sign}((w, x) + b)$ նշանը: $d=2$ -ի դեպքում ենթատարածությունների դասը մեկնաբանվում է հետևյալ կերպ՝

«+» նշիչ ունի այն ենթատարածությունը, որի վեկտորները սուր անկյուն են կազմում w վեկտորի հետ: Իսկ w -ի վեկտորի հետ բութ անկյուն կազմող վեկտորները դասվում են «-» նշիչ ունեցող ենթատարածությունում (տե՛ս նկ. 2.2):



Նկ. 2.2: Ենթատարածություն

Թեորեմ 2.11. Տեղի ունի հետևյալ հավասարությունը՝ $VCdim(HS_d) = d + 1$:

Ապացույց: Բավական է ցույց տալ, որ համասեռ ենթատարածությունների VC չափողականությունը հավասար է d -ի: Դիտարկենք միավոր վեկտորների բազիսը՝

$$C = \{e_1, e_2, \dots, e_d\}:$$

Այդ դեպքում կամայական $\{y_1, y_2, \dots, y_d\}$ նշիչների դեպքում վերցնելով $w = (y_1, y_2, \dots, y_d)$

վեկտորը՝ կատանանք

$$(w, e_i) = y_i, \quad i = 1, 2, \dots, d:$$

Դիցուք այժմ ունենք կամայական $x_1, x_2, \dots, x_{d+1}$ վեկտորների համախումբ: Այդ դեպքում գոյություն ունեն $\alpha_1, \alpha_2, \dots, \alpha_{d+1}$ թվեր, որոնցից գոնե մեկը գրոյից տարբեր է, այնպիսին, որ

$$\sum_{i=1}^{d+1} \alpha_i x_i = 0:$$

Դիցուք

$$I_+ = \{i : \alpha_i > 0\}, \quad I_- = \{i : \alpha_i < 0\}:$$

Ենթադրենք այս երկու բազմությունները դատարկ չեն: Այդ դեպքում

$$\sum_{i \in I_+} \alpha_i x_i = \sum_{i \in I_-} |\alpha_i| x_i:$$

Դիցուք գոյություն ունի այնպիսի w վեկտոր, որ

$$(w, x_i) > 0, \quad i \in I_+, \quad (w, x_i) < 0, \quad i \in I_-:$$

Այդ դեպքում

$$\begin{aligned} 0 &< \sum_{i \in I_+} \alpha_i (w, x_i) = \left(\sum_{i \in I_+} \alpha_i x_i, w \right) = \left(\sum_{i \in I_-} |\alpha_i| x_i, w \right) = \\ &= \sum_{i \in I_-} |\alpha_i| (x_i, w) < 0: \end{aligned}$$

Ստացանք հակասություն: Հակասություն կատանանք նաև այն դեպքերում, երբ I_+ , I_- բազմություններից որևէ մեկը դատարկ է: Այսպիսով ապացուցվեց թեորեմի պնդումը:

2.5 Մեքենայական ուսուցման ֆունդամենտալ թեորեմը

Այս պարագրաֆում H բազմության $V \dim(H)$ չափողականության միջոցով տրվում է անհրաժեշտ և բավարար պայման օպտիմալ պրեդիկտորների գոյության համար: Միաժամանակ դուրս է բերվում հիպոթեզների բազմության կոռեկտության բավարար պայման բինար դասակարգման խնդիրներում:

Այդ փաստերը ներկայացվում են մեքենայական ուսուցման ֆունդամենտալ թեորեմում:

Ճշմարիտ է հետևյալ պնդումը, որը համարվում է մեքենայական ուսուցման ֆունդամենտալ թեորեմ:

Թեորեմ 2.12. Դիցուք հիպոթեզների H բազմությունը բաղկացած է $h : X \rightarrow \{0, 1\}$ ֆունկցիաներից: Ենթադրենք նաև, որ որպես կորստի ֆունկցիաներ

հանդես են գալիս $l(z, h) \equiv [h(x) \neq y]$ ֆունկցիաները: Դիցուք նաև

$$VCdim(H) = d < \infty, \quad \epsilon, \delta \in (0, 1):$$

Այդ դեպքում կամայական հավանականային D բաշխման և պատահական

$$S \sim D^l, \quad l \geq 4 \frac{2d}{(\delta\epsilon)^2} \ln\left(\frac{2d}{(\delta\epsilon)^2}\right) + \frac{4d \ln(2e/d)}{(\delta\epsilon)^2}$$

հավաքածուի համար տեղի ունի հետևյալ անհավասարությունը՝

$$P(L_S(h_S) \leq \min_{h \in H} L_D(h) + \epsilon) > 1 - \delta:$$

Ապացույց: Դիտարկենք հետևյալ ոչ բացասական պատահական մեծությունը՝

$$\sup_{h \in H} |L_D(h) - L_S(h)|:$$

Նկատենք, որ կամայական $h \in H$ հիպոթեզի համար ունենք

$$L_D(h) = E_{S' \sim D^l}[L_{S'}(h)],$$

որտեղ $S' = \{z'_1, \dots, z'_l\}$ լրացուցիչ հավաքածու է՝ անկախ միանման բաշխված օբյեկտներով: Հետևաբար՝

$$E_{S \sim D^l} \left[\sup_{h \in H} |L_D(h) - L_S(h)| \right] = E_{S \sim D^l} \left[\sup_{h \in H} |E_{S' \sim D^l} L_{S'}(h) - L_S(h)| \right]:$$

Ըստ եռանկյան անհավասարության՝ ունենք

$$|E_{S' \sim D^l}[L_{S'}(h) - L_S(h)]| \leq E_{S' \sim D^l} |L_{S'}(h) - L_S(h)|:$$

Այսպիսով՝

$$\begin{aligned} E_{S' \sim D^l} \left[\sup_{h \in H} |L_D(h) - L_S(h)| \right] &\leq E_{S', S \in D^l} \left[\sup_{h \in H} |L_{S'}(h) - L_S(h)| \right] = \\ &= E_{S', S \sim D^l} \left[\sup_{h \in H} \frac{1}{l} \left| \sum_{i=1}^l l(h, z'_i) - l(h, z_i) \right| \right]: \end{aligned}$$

Քանի որ այս անհավասարության աջ մասում կարող ենք վեկտորների տեղերը փոխել, ապա կառաջանա նաև $-(l(h, z'_i) - l(h, z_i))$ անդամը: Հետևաբար անհավասարության աջ մասը հավասար կլինի

$$E_{S', S \sim D^l} \left[\sup_{h \in H} \frac{1}{l} \left| \sum_{i=1}^l \sigma_i (l(h, z'_i) - l(h, z_i)) \right| \right],$$

որտեղ $\sigma_i \in \{-1, 1\}$: Այսպիսով, քանի որ անհավասարությունը ճիշտ է ցանկացած

$\sigma \in \{+, -1\}^l$ նշանների դեպքում, ապա այն տեղի կունենա նաև այդ նշանների հավասարաչափ բաշխման դեպքում: Նշանակենք այդ բաշխումը $U_{+,-}$: Հետևաբար անհավասարության աջ մասը կարող ենք գրել հետևյալ կերպ՝

$$E_{\sigma \sim U_{+,-}^l} E_{S', S \sim D^l} \left[\sup_{h \in H} \frac{1}{l} \left| \sum_{i=1}^l \sigma_i (l(h, z'_i) - l(h, z_i)) \right| \right] :$$

Իսկ մաթեմատիկական սպասման գծային հատկության համաձայն վերևի արտահայտությունը կարող ենք գրել այսպես՝

$$E_{S, S' \sim D^l} E_{\sigma \sim U_{+,-}^l} \left[\sup_{h \in H} \frac{1}{l} \left| \sum_{i=1}^l \sigma_i (l(h, z'_i) - l(h, z_i)) \right| \right] :$$

Ֆիքսենք S, S' հավաքածուները և որպես C համարենք

$$z_1, z_2, \dots, z_l, z'_1, z'_2, \dots, z'_l$$

հաջորդականությունը: Այդ դեպքում սուլաբեմումը կարող ենք վերցնել ըստ $h \in H_C$ հիպոթեզների, այսինքն՝

$$\begin{aligned} E_{\sigma \sim U_{+,-}^l} \left[\sup_{h \in H} \frac{1}{l} \left| \sum_{i=1}^l \sigma_i (l(h, z'_i) - l(h, z_i)) \right| \right] &= \\ = E_{\sigma \sim U_{+,-}^l} \left[\max_{h \in H_C} \frac{1}{l} \left| \sum_{i=1}^l \sigma_i (l(h, z'_i) - l(h, z_i)) \right| \right] : \end{aligned}$$

Ֆիքսենք կամայական $h \in H_C$ և դիտարկենք պատահական

$$\theta_h = \frac{1}{l} \sum_{i=1}^l \sigma_i (l(h, z'_i) - l(h, z_i))$$

մեծությունը: Ակնհայտ է, որ այս պատահական մեծությունը արժեքներ է ընդունում $[-1, 1]$ հատվածից և $E[\theta_h] = 0$: Հետևաբար համաձայն Հոֆմանի անհավասարության՝ կունենանք

$$P[|\theta_h| > \rho] \leq 2 \exp(-2l\rho^2) :$$

Այստեղից հետևում է, որ

$$P[\max_{h \in H_C} |\theta_h| > \rho] \leq 2|H_C| \exp(-2l\rho^2) :$$

Հետևաբար՝

$$E[\max_{h \in H_C} |\theta_h|] \leq \frac{4 + \sqrt{\ln |H_C|}}{\sqrt{2l}} :$$

Այս հետևությունը բխում է հետևյալ լեմմայից:

Լեմմա 2.4. Դիցուք X -ը պատահական մեծություն է, իսկ x' -ը սկալյար: Ենթադրենք նաև, որ գոյություն ունեն $a > 0$, $b \geq e$ թվեր, այնպիսին, որ ցանկացած $t \geq 0$ համար տեղի ունի հետևյալ անհավասարությունը՝

$$P(|X - x'| > t) \leq 2be^{-t^2/a^2}:$$

Այդ դեպքում՝

$$E[|X - x'|] \leq a(2 + \sqrt{\ln b}):$$

Ապացույց: Դիտարկենք մոնոտոն աճող $t_i = a(i + \sqrt{\ln b})$, $i = 0, 1, 2, \dots$ հաջորդականությունը: Հետևաբար՝

$$E[|X - x'|] \leq a\sqrt{+} + \sum_{i=1}^{\infty} t_i P[|X - x'| > t_{i-1}]:$$

Ըստ լեմմայի ենթադրության՝ կունենանք

$$\begin{aligned} & \sum_{i=1}^{\infty} t_i P[|X - x'| > t_{i-1}] \leq \\ & \leq 2ab \sum_{i=1}^{\infty} (i + \sqrt{\ln b}) e^{-1(i-1+\sqrt{\ln b})^2} \leq \\ & \leq 2ab \int_{1+\sqrt{\ln b}}^{\infty} x e^{-(x-1)^2} dx = \\ & = 2ab \int_{\sqrt{\ln b}}^{\infty} (y+1) e^{-y^2} dy \leq \\ & \leq 4ab \int_{\sqrt{\ln b}}^{\infty} y e^{-y^2} dy = 2a: \end{aligned}$$

Լեմման ապացուցվեց:

Հայտնի է նաև, որ եթե $l > d + 1$, ապա

$$|H_C| \leq (el/d)^d:$$

Այս անհավասարությունը կոչվում է Չաուերայի լեմմա, որի ապացույցը կարող եք տեսնել [1]-ում (Լեմմա A 5 , էջ 407): Օգտագործելով դա վերևի անհավասարությունում՝ կստանանք

$$E_{S \sim D^l}[\sup_{h \in H} |L_D(h) - L_S(h)|] \leq \frac{4 + \sqrt{\ln b}}{\sqrt{2l}}:$$

Այսպիսով կիրառելով Չերիշևի անհավասարությունը, կարող ենք ասել, որ

առնվազն $1 - \delta$ հավանականությամբ տեղի ունի

$$|L_S(h) - L_D(h)| \leq \frac{4 + \sqrt{d \ln(2el/d)}}{\delta \sqrt{2l}}$$

անհավասարությունը: Կարող ենք ենթադրել, որ $\sqrt{d \ln(2el)/d} > 4$: Այդ դեպքում

$$|L_S(h) - L_D(h)| \leq \frac{1}{\delta} \sqrt{\frac{2d \ln(2el/d)}{l}} :$$

Որպեսզի այս անհավասարության աջ մասը լինի փոքր ϵ -ից, պետք է, որ

$$l \geq \frac{2d \ln l}{(\delta \epsilon)^2} + \frac{2d \ln(2e/d)}{(\delta \epsilon)^2} :$$

Այժմ քննարկենք, թե ինչ բավարար պայմանի դեպքում տեղի կունենա այս անհավասարությունը: Դրա համար նախ ապացուցենք հետևյալ պնդումները:

Լեմմա 2.5. Դիցուք $a > 0$: Այդ դեպքում

$$x \geq 2a \ln a \Rightarrow x \geq a \ln x :$$

Ապացույց: Դիցուք $a \in (0, \sqrt{e})$, ապա $x \geq a \ln x$ անհավասարությունը տրիվիալ է: Դիցուք այժմ $x > \sqrt{e}$: Այս բազմության վրա դիտարկենք $f(x) = x - a \ln x$ ֆունկցիան: Քանի որ նրա ածանցյալը դրական է $x > a$ բազմության վրա, ապա ֆունկցիան աճում է այդտեղ: Բացի դրանից ունենք

$$\begin{aligned} f(2a \ln a) &= 2a \ln a - a \ln(2a \ln a) = \\ &= 2a \ln a - a \ln a - a \ln(2 \ln a) = \\ &= a \ln a - a \ln(2 \ln a) : \end{aligned}$$

Այստեղից հետևում է լեմմայի պնդումը, քանի որ

$$a - 2 \ln a > 0 \quad \forall a :$$

Լեմմա 2.6. Դիցուք $a \geq 1$, $b > 0$: Այդ դեպքում

$$x \geq 4a \ln 2a + 2b \Rightarrow x \geq a \ln x + b :$$

Ապացույց: Բավական է ցույց տալ, որ $x \geq 4a \ln(2a) + 2b$ պայմանից հետևում է, որ $x \geq 2a \ln x$ և $x \geq 2b$: Քանի որ $a \geq 1$, ապա $x \geq 2b$: Մյուս կողմից, քանի որ $b > 0$, ապա $x \geq 4a \ln(2a)$: Այստեղից, համաձայն նախորդ լեմմայի, կստանանք $x \geq 2a \ln x$: Լեմման ապացուցվեց:

Այս լեմմայից հետևում է, որ եթե

$$l \geq 4 \frac{2d}{(\delta \epsilon)^2} \ln \frac{2d}{(\delta \epsilon)^2} + \frac{4d \ln(2e/d)}{(\delta \epsilon)^2},$$

ապա տեղի կունենա

$$l \geq \frac{2d \ln l}{(\delta\epsilon)^2} + \frac{2d \ln(2e/d)}{(\delta\epsilon)^2}$$

անհավասարությունը:

Շարունակենք հիմնական թեորեմի ապացույցը: Մենք ստացանք, որ առնվազն $1 - \delta$ հավանականությամբ տեղի ունի $|L_D(h) - L_S(h)| \leq \epsilon$, $\forall h \in H$ անհավասարությունը S հավաքածուի l երկարության համապատասխան ձևով ընտրության դեպքում: Այդ դեպքում կամայական $H_S \in \arg \min_{h \in H} L_S(h)$ հիպոթեզի համար կունենանք

$$\begin{aligned} L_D(h_S) &\leq L_S(h_S) + \epsilon/2 \leq L_S(h) + \epsilon/2 \leq \\ &\leq L_D(h) + \epsilon/2 + \epsilon/2 = L_D(h) + \epsilon: \end{aligned}$$

Թեորեմն ապացուցվեց:

3

Ուռուցիկ անալիզի տարրեր

3.1 Ուռուցիկ բազմությունների անջատման թեորեմները

Այստեղ նկարագրվում են ուռուցիկ բազմությունների անջատման թեորեմներ, որոնք հետագայում հիմք են հանդիսանալու ուռուցիկ խնդիրներում միմյանմի անհրաժեշտ ու բավարար պայմաններ ստանալու համար:

Սահմանում 3.1. Դիցուք $p \in \mathbb{R}^n$ -ն ոչ զրոյական վեկտոր է, իսկ α -ն իրական թիվ է:

$$H = \{x \in \mathbb{R}^n / (p, x) = \alpha\}$$

բազմությունը կոչվում է հիպերհարթություն, իսկ

$$H_+ = \{x \in \mathbb{R}^n / (p, x) \geq \alpha\}, \quad H_- = \{x \in \mathbb{R}^n / (p, x) \leq \alpha\}$$

բազմությունները այդ հիպերհարթությամբ ծնված կիսատարածություններ:

Սահմանում 3.2. $M_1, M_2 \subseteq \mathbb{R}^n$ բազմությունները կոչվում են **անջատվող հիպերհարթությամբ**, եթե գոյություն ունի այնպիսի $p \neq 0$ վեկտոր, որ

$$(p, x) \leq (p, y) \quad \forall x \in M_1, \quad \forall y \in M_2:$$

Սա նշանակում է, որ գոյություն ունի այնպիսի H հիպերհարթություն, որ

$$M_1 \subseteq H_-, \quad M_2 \subseteq H_+:$$

Թեորեմ 3.1. Դիցուք $M_1, M_2 \subseteq \mathbb{R}^n$ այնպիսի ուռուցիկ բազմություններ են, որ $M_1 \cap M_2 = \emptyset$: Այդ դեպքում դրանք անջատվում են հիպերհարթությամբ:

Ապացույց: Թեորեմի ապացույցը հենվում է հետևյալ երկու պնդումների վրա:

Լեմմա 3.1. Եթե M -ը փակ ուռուցիկ բազմություն է և $a \notin M$, ապա այդ կետը հիպերհարթությամբ անջատվում է M բազմությունից:

Ապացույց: Գտնենք a կետի պրոյեկցիան M բազմության վրա: Յույց տանք, որ

այն գոյություն ունի և միակն է: Վերցնենք a կենտրոնով և $r \equiv \inf_{x \in M} \|a - x\| + \varepsilon$ շառավղով $B_r(a)$ գունդը, որտեղ $\varepsilon > 0$ ֆիքսած թիվ է: Քանի որ գունդը փակ սահմանափակ բազմություն է, ապա $M \cap B_r(a)$ հատումը կոմպակտ է: Հետևաբար a կետի հեռավորությունը այդ հատումից հասանելի է: Դժվար չէ համոզվել, որ այդ կետը ամենամոտիկն է նաև M բազմությունից, այսինքն՝ դա $\Pi_M(a)$ կետն է (a կետի պրոյեկցիան M բազմության վրա): Ցույց տանք $\Pi_M(a)$ -ի միակությունը: Ենթադրենք՝ գոյություն ունի երկու ամենամոտիկ կետ: Դիցուք դրանք b, c կետերն են: Եթե a, b, c կետերը եռանկյուն են կազմում, ապա այն հավասարասրուն է, որի սրունքը հավասար է $d \equiv \inf_{x \in M} \|x - a\|$, իսկ հիմքը $[c, b]$ հատվածն է: Քանի որ M -ը ուռուցիկ է, ապա $[c, b] \subset M$: Հետևաբար՝ a գագաթից տարած բարձրության հիմքը $[c, b]$ հատվածի միջնակետն է: Քանի որ բարձրությունը փոքր է սրունքից, և սրունքը մինիմալ հեռավորությունն է, ապա ստացանք հակասություն: Նշենք նաև a, b, c կետերը միևնույն գծի վրա գտնվել չեն կարող, քանի որ $a \notin M$: Այժմ $[\Pi_M(a), a]$ հատվածի g միջնակետով տանենք հիպերհարթություն, որի նորմալը $a - \Pi_M(a)$ վեկտորն է: Ցույց տանք, որ այդ հիպերհարթությունը անջատում է a կետը M բազմությունից: Դիցուք H_+ -ը այն կիսատարածությունն է, որը պարունակում է a կետը, իսկ H_- կիսատարածությունը չի պարունակում a -ն: Ցույց տանք, որ $M \subset H_-$: Նախ պարզ է, որ g կետի պրոյեկցիան M բազմության վրա $\Pi_M(a)$ կետն է: Ենթադրենք, որ գոյություն ունի $e \in M \cap H_+$: Այդ դեպքում $g, \Pi_M(a), e$ գագաթներով եռանկյան մեջ g գագաթից տարված բարձրության հիմքը կպատկանա $[\Pi_M(a), e] \subset M$ հատվածին, և այդ բարձրությունը փոքր է $[g, \Pi_M(a)]$ հատվածի երկարությունից, որը հակասություն է, քանի որ այն g կետի մինիմալ հեռավորությունն է M բազմությունից:

Լեմմա 3.2. Դիցուք M -ը ուռուցիկ բազմություն է, իսկ $a \notin M$: Այդ դեպքում a կետը հիպերհարթությամբ անջատվում է M բազմությունից:

Ապացույց: Եթե $a \notin \overline{M}$, ապա հանգում ենք նախորդ դեպքին, քանի որ այս դեպքում a կետը կարելի է հիպերհարթությամբ անջատել $\overline{M}$ բազմությունից, հետևաբար M բազմությունից: Այժմ ենթադրենք, որ a -ն M -ի եզրային կետ է: Այդ դեպքում գոյություն ունի այնպիսի $\{x^k\} \notin \overline{M}$ հաջորդականություն, որ $x^k \rightarrow a$: Համաձայն լեմմա 3.1-ի՝ յուրաքանչյուր x^k կետ կարելի է անջատել M բազմությունից հիպերհարթությամբ: Դա նշանակում է, որ գոյություն ունեն այնպիսի p^k միավոր վեկտորներ, որ

$$(p^k, x) \leq (p^k, x^k) \quad \forall x \in M:$$

Զսահմանափակելով ընդհանրությունը, կարելի է ենթադրել, որ $p^k \rightarrow p^0 \neq 0$: Այստեղ անցնելով սահմանի՝ կստանանք

$$(p^0, x) \leq (p^0, a) \quad \forall x \in M,$$

ինչը նշանակում է a կետի և M բազմության անջատում հիպերհարթությամբ:

Այժմ անցնենք թեորեմի ապացույցին: Նշանակենք՝ $M \equiv M_1 - M_2$: M -ը ուռուցիկ է, և քանի որ $M_1 \cap M_2 = \emptyset$, ապա $0 \notin M$: Համաձայն լեմմա 3.2-ի՝ 0 կետը կարելի է

անջատել M բազմությունից: Դա նշանակում է, որ գոյություն ունի այնպիսի $p \neq 0$ վեկտոր, որ

$$(p, x - y) \leq 0 \quad \forall x \in M_1, \quad \forall y \in M_2:$$

Այստեղից

$$(p, x) \leq (p, y) \quad \forall x \in M_1, \quad \forall y \in M_2:$$

Հետևանք 3.1. Դիցուք $M \subseteq \mathbb{R}^n$ -ը ուռուցիկ բազմություն է, իսկ a -ն՝ նրա եզրային կետը: Այդ դեպքում a կետով կարելի է տանել հիպերհարթություն, այնպիսին, որ M բազմությունը ընկած լինի այդ հիպերհարթությանը ծնված կիսատարածություններից որևէ մեկում: Այդպիսի հիպերհարթությունը կոչվում է հենման հիպերհարթություն:

Սահմանում 3.3. Դիցուք M_1 և $M_2 \subseteq \mathbb{R}^n$: Կասենք, որ այդ բազմությունները «խիստ» են անջատվում հիպերհարթությանը, եթե գոյություն ունեն այնպիսի p վեկտոր և $\varepsilon > 0$ թիվ, որ

$$(p, x) \leq (p, y) - \varepsilon \quad \forall x \in M_1, \quad \forall y \in M_2:$$

Տեղի ունի նաև ուռուցիկ բազմությունների «խիստ» անջատման հետևյալ թեորեմը:

Թեորեմ 3.2. Դիցուք $M_1 \subset \mathbb{R}^n$ փակ ուռուցիկ բազմություն է, իսկ $M_2 \subset \mathbb{R}^n$ ուռուցիկ կոմպակտ է և

$$M_1 \cap M_2 = \emptyset:$$

Այդ դեպքում այս բազմությունները «խիստ» անջատվում են հիպերհարթությանը:

ԽՆԴԻՐՆԵՐ

1. Գրել այն հիպերհարթության հավասարումը, որը անջատում է $(-1, 2, 1, -3)$ կետը $M \subseteq \mathbb{R}^4$ բազմությունից, երբ այն տրվում է անհավասարությունների հետևյալ համակարգով.

$$\begin{cases} 5x_1 + x_2 - 5x_3 - 3x_4 \leq 1, \\ -3x_2 - 2x_3 + 5x_4 \leq 2, \\ 3x_2 + x_3 + 2x_4 \leq 0, \\ x_1 + x_2 + 3x_3 - x_4 \leq 9: \end{cases}$$

2. Ապացուցել հետևյալ պնդումը: Որպեսզի $M \subseteq \mathbb{R}^n$ փակ բազմությունը լինի ուռուցիկ, անհրաժեշտ է և բավարար, որ ցանկացած $b \notin M$ կետ խիստ անջատվի հիպերհարթությանը M բազմությունից:
3. Դիցուք ուռուցիկ բազմության տրված կետով կարելի է տանել երկու հենման հիպերհարթություն: Ապացուցել, որ այդ կետով անցնում են անթիվ բազմության հենման հիպերհարթություններ:

4. Դիցուք $M_1, M_2 \subset \mathbb{R}^n$ այնպիսի ուռուցիկ բազմություններ են, որ

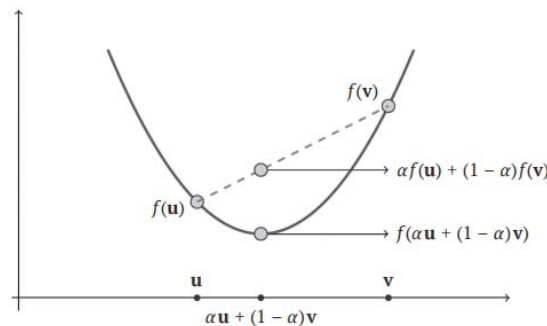
$$\text{int}M_2 \neq \emptyset \text{ և } M_1 \cap \text{int}M_2 = \emptyset:$$

Ապացուցել, որ M_1 և M_2 բազմությունները անջատվում են հիպերհարթությամբ:

3.2 Ուռուցիկ և β -ողորկ ֆունկցիաներ

$f: C \rightarrow \mathbb{R}$ ֆունկցիան կոչվում է ուռուցիկ C ուռուցիկ բազմության վրա, եթե

$$f(\alpha u + (1 - \alpha)v) \leq \alpha f(u) + (1 - \alpha)f(v), \quad u, v \in C, \quad \alpha \in [0, 1] \text{ (տե՛ս նկ. 3.1):}$$



Նկ. 3.1: Ուռուցիկ ֆունկցիա

Եթե $f: \mathbb{R}^n \rightarrow \mathbb{R}$ ֆունկցիան դիֆերենցելի է, ապա ուռուցիկությունը համարժեք է հետևյալ անհավասարությանը՝

$$f(v) - f(w) \geq (f'(w), v - w), \quad w, v \in C:$$

Պնդում 1: Դիցուք $f: \mathbb{R}^d \rightarrow \mathbb{R}$ ֆունկցիան ներկայացվում է հետևյալ կերպ՝

$$f(w) \equiv g((w, x) + y),$$

որտեղ $x \in \mathbb{R}^d, y \in \mathbb{R}, g: \mathbb{R} \rightarrow \mathbb{R}$ ուռուցիկ ֆունկցիա է:

Այդ դեպքում f -ը ուռուցիկ է:

Ապացույց: Ցանկացած $w_1, w_2 \in \mathbb{R}^n$ վեկտորների և $\alpha \in [0, 1]$ թվի համար ունենք

$$\begin{aligned} f(\alpha w_1 + (1 - \alpha)w_2) &\leq g((\alpha w_1 + (1 - \alpha)w_2, x) + y) = g(\alpha((w_1, x) + y) + (1 - \alpha)((w_2, x) + y)) \leq \\ &\leq \alpha g((w_1, x) + y) + (1 - \alpha)g((w_2, x) + y) \leq \\ &\leq \alpha f(w_1) + (1 - \alpha)f(w_2): \end{aligned}$$

Պնդումն ապացուցվեց:

Կասենք, որ $f(w)$ ֆունկցիան ρ -լիպշիցյան է, եթե

$$|f(w_1) - f(w_2)| \leq \rho \|w_1 - w_2\| \quad \forall w_1, w_2 \in \mathbb{R}^n:$$

Պնդում: Դիցուք $f(w) = g_1(g_2(w))$, որտեղ g_1 -ը և g_2 -ը համապատասխանաբար ρ_1 և ρ_2 - լիպշիցյան ֆունկցիաներ են: Այդ դեպքում $f(w)$ -ը $\rho_1\rho_2$ - լիպշիցյան է:

Ապացույց: Ունենք

$$|f(w_1) - f(w_2)| = |g_1(g_2(w_1)) - g_1(g_2(w_2))| \leq$$

$$\leq \rho_1 \|g_2(w_1) - g_2(w_2)\| \leq \rho_1\rho_2 \|w_1 - w_2\|:$$

Կասենք, որ $f(w)$ ֆունկցիան β -ոդորկ է, եթե նրա գրադիենտը β -լիպշիցյան է: Այսինքն՝

$$\|f'(w_1) - f'(w_2)\| \leq \beta \|w_1 - w_2\|:$$

Կարելի է ցույց տալ, որ եթե f -ը β ոդորկ է, ապա տեղի ունի հետևյալ անհավասարությունը՝

$$f(v) \leq f(w) + (f'(w), v - w) + \frac{\beta}{2} \|v - w\|^2:$$

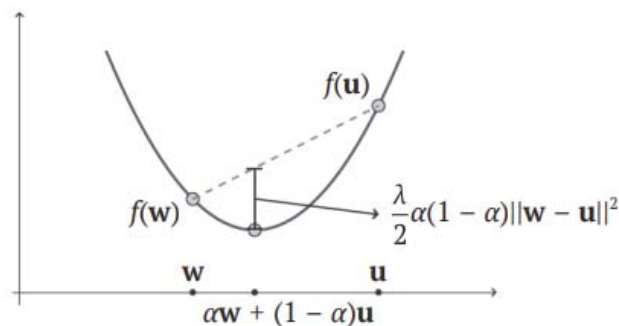
Դիցուք այժմ f -ը ոչ բացասական, β -ոդորկ և ուռուցիկ ֆունկցիա է: Այդ դեպքում վերևի անհավասարության մեջ տեղադրելով $v = w - (1/\beta)f'(w)$ և հաշվի առնելով ուռուցիկության $f(v) \geq f(w) + (f'(w), v - w)$ անհավասարությունը՝ կստանանք

$$\|f'(w)\|^2 \leq 2\beta f(w):$$

Այս անհավասարությունը մենք կօգտագործենք ռեգուլյարիզացիայի մոդելի կայունության ապացուցման ընթացքում:

Սահմանում: f ֆունկցիան կոչվում է λ -ուժեղ ուռուցիկ, եթե տեղի ունի հետևյալ անհավասարությունը՝

$$f(\alpha w + (1 - \alpha)v) \leq \alpha f(w) + (1 - \alpha)f(v) - \frac{\lambda}{2} \alpha(1 - \alpha) \|w - v\|^2, \quad u, v \in \mathbb{R}^n, \quad \alpha \in [0, 1]:$$



Նկ. 3.2: Ուժեղ ուռուցիկ ֆունկցիա

Ճշմարիտ է հետևյալ կարևոր պնդումը:

Թեորեմ 3.3. Դիցուք $l(w)$ ուռուցիկ ֆունկցիա է: Այդ դեպքում կամայական $\lambda > 0$ դրական թվի համար

- $f(w) = l(w) + \lambda \|w\|^2$ ֆունկցիան 2λ -ուժեղ ուռուցիկ է:
- f -ը ունի միակ մինիմումի w^* կետ $\mathbb{R}^n$ -ի վրա, և տեղի ունի

$$f(w) - f(w^*) \geq \frac{\lambda}{2} \|w - w^*\|^2$$

անհավասարությունը:

- Եթե նաև f -ը դիֆերենցելի է, ապա

$$f(w) - f(u) \geq (f'(u), w - u) + \frac{\lambda}{2} \|w - u\|^2:$$

ԽՆԴԻՐՆԵՐ

- Դիցուք $x \in \mathbb{R}^d$, $y \in \mathbb{R}$: Ցույց տալ, որ $f(w) = ((w, x) - y)^2$ ուռուցիկ ֆունկցիա է:
- Դիցուք $x \in \mathbb{R}^d$, $y \in \{-1, 1\}$: Ցույց տալ, որ $f(w) = \ln(1 + \exp(-y(w, x)))$ ֆունկցիան ուռուցիկ է:
- Դիցուք $f_i(w)$, $i = 1, 2, \dots, m$ ֆունկցիաները ուռուցիկ են: Ապացուցել, որ ուռուցիկ են նաև հետևյալ ֆունկցիաները՝

ա) $g_1(w) = \max_{i \in [1, m]} f_i(w)$,

բ) $g_2(w) = \sum_{i=1}^m \alpha_i f_i(w)$, $\alpha_i \geq 0$, $i = 1, 2, \dots, m$:

- Ցույց տալ, որ

$$f(x) = \ln(1 + \exp(x))$$

ֆունկցիան 1-լիպշիցյան է $\mathbb{R}$ -ի վրա:

- Ցույց տալ, որ

$$f(x) = \ln(1 + \exp(x))$$

ֆունկցիան $(1/4)$ -ոդորկ է:

- Դիցուք $f(w) = g((w, x) + y)$, որտեղ $g: \mathbb{R} \rightarrow \mathbb{R}$ -ն β -ոդորկ ֆունկցիա է: Ցույց տալ, որ f -ը $(\beta \|x\|^2)$ -ոդորկ է:
- Դիցուք $w \in \mathbb{R}^d$, $y \in \mathbb{R}$: Ցույց տալ, որ $f(w) = ((w, x) - y)^2$ ֆունկցիան $(2\|x\|^2)$ -ոդորկ է:
- Ցույց տալ, որ $f(w) = \ln(1 + \exp(-y(w, x)))$, $y \in \{-1, +1\}$ ֆունկցիան $(\|x\|^2/4)$ -ոդորկ է:

3.3 Ուռուցիկ ֆունկցիայի սուբդիֆերենցիալ

Դիցուք f -ը ուռուցիկ ֆունկցիա է՝ որոշված $\mathbb{R}^n$ -ի վրա: f ֆունկցիայի սուբդիֆերենցիալ x_0 կետում կոչվում է հետևյալ բազմությունը՝

$$\partial f(x_0) = \{v : f(x) - f(x_0) \geq (v, x - x_0) \forall x \in \mathbb{R}^n\}:$$

Տրված անհավասարությանը բավարարող v վեկտորը կոչվում է սուբգրադիենտ: Ճշմարիտ են հետևյալ պնդումները.

- Եթե f ուռուցիկ ֆունկցիան դիֆերենցելի է x_0 կետում, ապա $\partial f(x_0) = \{f'(x_0)\}$:
- Դիցուք $f_i, i = 1, 2, \dots, m$ ուռուցիկ ֆունկցիաները անընդհատ են x_0 կետում: Այդ դեպքում

$$\partial\left(\sum_{i=1}^m f_i\right)(x_0) = \sum_{i=1}^m \partial f_i(x_0):$$

- Դիցուք $f_i, i \in I$ ուռուցիկ ֆունկցիաները դիֆերենցելի են x_0 կետում, և

$$f(x) = \max_{i \in I} f_i(x):$$

Այդ դեպքում

$$\partial f(x_0) = \text{conv}\left\{\bigcup_{i \in I(x_0)} f'_i(x_0)\right\},$$

որտեղ $I(x_0) = \{i \in I : f(x_0) = f_i(x_0)\}$:

- Որպեսզի x_0 կետը լինի f ուռուցիկ ֆունկցիայի մինիմումի կետ անհրաժեշտ է և բավարար, որ $0 \in \partial f(x_0)$:

Բերենք սուբդիֆերենցիայի հաշվման օրինակներ:

Օրինակ 1: Դիցուք $f(x) = |x| = \max\{x, -x\}$: Այդ դեպքում

$$\partial f(x) = \begin{cases} 1, & x > 0 \\ [-1, 1], & x = 0, \\ -1, & x < 0: \end{cases}$$

Օրինակ 2: Դիցուք

$$f(x) = \|x\|_1 = \sum_{i=1}^m f_i(x), \quad f_i(x) = |x_i|, \quad i = 1, 2, \dots, m:$$

Այդ դեպքում

$$\partial f_i(x) = \begin{cases} \{\text{sign}(x_i)e_i\}, & x_i \neq 0, \\ [-e_i, e_i], & x_i = 0: \end{cases}$$

Իրոք, համաձայն գումարի սուբդիֆերենցիալի վերաբերյալ կանոնի, ունենք

$$\partial f(x) = \sum_{i=1}^m \partial f_i(x) = \sum_{i \in I_1(x)} \text{sign}(x_i) e_i + \sum_{i \in I_0(x)} [-e_i, e_i],$$

$$I_0(x) = \{i : x_i = 0\}, I_1(x) = \{i : x_i \neq 0\} : \text{Այսպիսով՝}$$

$$\partial f(x) = \{z \in \mathbb{R}^m : z_i = \text{sign} x_i, i \in I_1(x), |z_i| \leq 1, i \in I_0(x)\}:$$

Օրինակ 3: Դիտարկենք օպտիմիզացիայի հետևյալ խնդիրը՝

$$\frac{1}{2} \|x - y\|_2^2 + \lambda \|x\|_1 \rightarrow \min_{x \in \mathbb{R}^m}:$$

Եթե x^* -ը խնդրի լուծումն է, ապա համաձայն մինիմումի անհրաժեշտ ու բավարար պայմանի՝ կունենանք

$$0 \in \partial f(x^*) = x^* - y + \lambda \partial \|x^*\|_1 :$$

Այստեղից հետևում է, որ

$$y_i - x_i^* = \lambda \text{sign}(x_i^*), x_i^* \neq 0,$$

$$|y_i - x_i^*| = |y_i| \leq \lambda, x_i^* = 0;$$

Այստեղից հետևում է, որ

$$x_i^* = \begin{cases} y_i - \lambda, & y_i > \lambda, \\ 0, & |y_i| \leq \lambda, \\ y_i + \lambda, & y_i < -\lambda: \end{cases}$$

Օրինակ 4: Դիցուք $f : \mathbb{R}^d \rightarrow \mathbb{R}$ -ն ուռուցիկ ֆունկցիա է, $A : \mathbb{R}^k \rightarrow \mathbb{R}^d$ ՝ գծային օպերատոր: Նշանակենք՝ $h(x) \equiv f(Ax + b)$, որտեղ $b \in \mathbb{R}^d$: Այդ դեպքում

$$\partial h(x) = A^T \partial f(Ax + b):$$

Երբ f -ը դիֆերենցելի ֆունկցիա է, ապա այս բանաձևը ուղղակի հետևանք է Թեյլորի բանաձևի: Իրոք, ունենք

$$f(A(x + \epsilon\gamma) + b) - f(Ax + b) = f(Ax + b + \epsilon A\gamma) - f(Ax + b) =$$

$$= \langle f'(Ax + b), \epsilon A\gamma \rangle + o(\epsilon) = \epsilon (A^T f'(Ax + b), \gamma) + o(\epsilon), \gamma \in \mathbb{R}^k:$$

Հետևաբար՝

$$\partial h(x) = A^T \partial f(Ax + b):$$

Օրինակ 5: Դիցուք $h(x) = \|Ax - b\|_1$: Հաշվենք h -ի սուբդիֆերենցիալը: Դիցուք a_i վեկտորը A մատրիցի i -րդ սյունն է: Նշանակենք $g(y) = \|y\|_1$: Օգտվելով օրինակ

2-ի արդյունքից՝ կստանանք

$$\partial g(Ax + b) = \sum_{i \in I_1(x)} \text{sign}((a_i, x) - b_i) e_i + \sum_{i \in I_0(x)} [-e_i, e_i],$$

որտեղ

$$I_0(x) = \{i : (a_i, x) = b_i\}, \quad I_1(x) = \{i : (a_i, x) \neq b_i\} :$$

Այստեղից, համաձայն նախորդ օրինակի,

$$\begin{aligned} \partial h(x) &= A^T \partial g(Ax + b) = \sum_{i \in I_1(x)} \text{sign}((a_i, x) - b_i) A^T e_i + \sum_{i \in I_0(x)} [-A^T e_i, A^T e_i] = \\ &= \sum_{i \in I_1(x)} \text{sign}((a_i, x) - b_i) a_i^T + \sum_{i \in I_0(x)} [-a_i^T, a_i^T] : \end{aligned}$$

3.4 Կուն-Թակերի թեորեմը

Այս պարագրաֆում ապացուցվում է Կուն-Թակերի թեորեմը, որը մաթեմատիկական ծրագրավորման ուռուցիկ խնդիրներում մինիմումի կետերի բնութագրման հիմնական պայմանն է: Հետագայում այս թեորեմի պայմանները մասնավորեցվում են մեքենայական ուսուցման հենման վեկտորների մեթոդում օպտիմալ հիպերհարթության կառուցման խնդրում:

Դիցուք $f_i(x)$, $i \in [0 : m]$, ֆունկցիաները ուռուցիկ են $\mathbb{R}^n$ -ի վրա: Դիտարկենք $f_0(x)$ ֆունկցիայի մինիմիզացիայի խնդիրը $M \equiv \{x \in \mathbb{R}^n / f_i(x) \leq 0, i \in [1 : m]\}$ բազմության վրա.

$$f_0(x) \rightarrow \min, \quad x \in M : \quad (3.1)$$

Հեշտ է ցույց տալ, որ M -ը ուռուցիկ բազմություն է:

(3.1) խնդիրը կոչվում է **ուռուցիկ ծրագրավորման** խնդիր: $f_0(x)$ ֆունկցիայի մինիմումի կետերը M բազմության վրա կոչվում են (3.1) խնդրի լուծումներ: Կազմենք **Լագրանժի** ֆունկցիան՝

$$L(x, \lambda) = \lambda_0 f_0(x) + \lambda_1 f_1(x) + \dots + \lambda_m f_m(x),$$

որտեղ $\lambda = (\lambda_0, \lambda_1, \dots, \lambda_m) \in \mathbb{R}^{m+1}$:

Թեորեմ 3.4. (Կուն-Թակեր: Անհրաժեշտությունը): Դիցուք x^* կետը (3.1) խնդրի լուծումն է: Այդ դեպքում գոյություն ունեն $\lambda_0 \geq 0, \lambda_1 \geq 0, \dots, \lambda_m \geq 0$ թվեր, որոնցից գոնե մեկը զրո չէ, այնպիսին, որ

$$1) \quad L(x, \lambda) \geq L(x^*, \lambda) \quad \forall x \in \mathbb{R}^n,$$

$$2) \quad \lambda_i f_i(x^*) = 0, \quad i \in [1 : m]:$$

Ապացույց: Չսահմանափակելով ընդհանրությունը՝ կարելի է ենթադրել, որ $f_0(x^*) = 0$: Հակառակ դեպքում կդիտարկենք $\tilde{f}_0(x) = f_0(x) - f_0(x^*)$ ֆունկցիան, որը

նույնպես ուռուցիկ է, և $\tilde{f}_0(x^*) = 0$: Դիտարկենք հետևյալ բազմությունը՝

$$C = \{(\mu_0, \mu_1, \dots, \mu_m) \in \mathbb{R}^{m+1} / \exists x \in \mathbb{R}^n :$$

$$f_0(x) < \mu_0, f_i(x) \leq \mu_i, i \in [1 : m]\} :$$

Այս բազմությունը ուռուցիկ է: Այն անմիջապես հետևում է սահմանումից: C -ն դատարկ չէ: Իսկապես, քանի որ

$$f_0(x^*) < 1, f_i(x^*) \leq 1, i \in [1 : m],$$

ապա

$$(1, 1, \dots, 1) \in C :$$

Ակնհայտ է նաև, որ $0 \notin C$, քանի որ հակառակ դեպքում գոյություն կունենար այնպիսի x կետ, որ

$$f_0(x) < 0, f_i(x) \leq 0, i \in [1 : m],$$

ինչը հակասություն է: 0 կետը անջատենք հիպերհարթությամբ C ուռուցիկ բազմությունից: Դա նշանակում է, որ գոյություն ունեն $\lambda_i, i \in [0 : m]$ թվեր, որոնցից գոնե մեկը զրո չէ, այնպիսին, որ

$$\sum_{i=0}^m \lambda_i \mu_i \geq 0 \quad \forall (\mu_0, \mu_1, \dots, \mu_m) \in C : \quad (3.2)$$

Ցույց տանք, որ $\lambda_i \geq 0, i \in [0 : m]$: Ենթադրենք, որ ինչ-որ i_0 ինդեքսի համար տեղի ունի $\lambda_{i_0} < 0$ անհավասարությունը: Ակնհայտ է, որ ցանկացած $\delta > 0$ և $\mu_{i_0} > 0$ թվի համար $\mu \equiv (\delta, 0, \dots, \mu_{i_0}, 0, \dots, 0) \in C$: Տեղադրելով μ վեկտորի կոորդինատները (3.2) անհավասարությունում՝ կստանանք

$$\delta \lambda_0 + \mu_{i_0} \lambda_{i_0} \geq 0 : \quad (3.3)$$

Այստեղ անցնելով սահմանի, երբ $\delta \rightarrow 0$ և $\mu_{i_0} \rightarrow +\infty$, կստանանք, որ (3.3) անհավասարության ձախ մասը ձգտում է $-\infty$, իսկ աջ մասը զրո է, որը հակասություն է:

Այժմ ապացուցենք թեորեմի երկրորդ պնդումը: (3.2) անհավասարությունում տեղադրելով

$$\mu_0 = \delta > 0, \mu_1 = 0, \dots, \mu_i = f_i(x^*), \mu_{i+1} = 0, \dots, \mu_m = 0$$

թվերը՝ կստանանք

$$\delta \lambda_0 + \lambda_i f_i(x^*) \geq 0 :$$

Այստեղ անցնելով սահմանի, երբ δ ձգտի զրոյի, կստանանք

$$\lambda_i f_i(x^*) \geq 0 : \quad (3.4)$$

Բայց քանի որ $\lambda_i \geq 0$ և $f_i(x^*) \leq 0$, ապա (3.4)-ից հետևում է, որ

$$\lambda_i f_i(x^*) = 0 : \quad (3.5)$$

Այժմ ապացուցենք թեորեմի եզրակացության առաջին կետը: Իրոք, ցանկացած $x \in \mathbb{R}^n$ -ի համար (3.3) անհավասարությունում տեղադրելով

$$\mu_0 = \delta + f_0(x), \mu_1 = f_1(x), \dots, \mu_m = f_m(x)$$

կոորդինատներով μ վեկտորը՝ կստանանք

$$\lambda_0(\delta + f_0(x)) + \lambda_1 f_1(x) + \dots + \lambda_m f_m(x) \geq 0:$$

Անցնելով սահմանի, երբ $\delta \rightarrow 0$, կստանանք

$$\lambda_0 f_0(x) + \lambda_1 f_1(x) + \dots + \lambda_m f_m(x) \geq 0 \quad \forall x \in \mathbb{R}^n: \quad (3.6)$$

(3.5)-(3.6)-ից հետևում է, որ

$$L(x, \lambda) = \sum_{i=0}^m \lambda_i f_i(x) \geq 0 =$$

$$= \lambda_0 f_0(x^*) + \lambda_1 f_1(x^*) + \dots + \lambda_m f_m(x^*) = L(x^*, \lambda) \quad \forall x \in \mathbb{R}^n:$$

Թեորեմ 3.5. (Կուլն-Թակեր: Բավարարությունը): Դիցուք x^* կետը բավարարում է (3.1) խնդրի բոլոր սահմանափակումներին՝

$$f_i(x^*) \leq 0, \quad i \in [1 : m],$$

և գոյություն ունեն այնպիսի $\lambda_0 > 0, \lambda_1 \geq 0, \lambda_2 \geq 0, \dots, \lambda_m \geq 0$ թվեր, որ տեղի ունեն թեորեմ 3.1-ի 1)-2) պայմանները: Այդ դեպքում x^* վեկտորը (3.1) խնդրի լուծում է:

Ապացույց: Չսահմանափակելով ընդհանրությունը՝ կարող ենք ենթադրել, որ $\lambda_0 = 1$: Այդ դեպքում, հաշվի առնելով 1) և 2) պայմանները, կունենանք

$$f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) \geq f_0(x^*) + \sum_{i=1}^m \lambda_i f_i(x^*) = f_0(x^*) \quad \forall x \in \mathbb{R}^n: \quad (3.7)$$

Եթե x վեկտորը այնպիսին է, որ $f_i(x) \leq 0 \quad \forall i \in [1 : m]$, ապա (3.7) անհավասարությունից կստանանք $f_0(x) \geq f_0(x^*)$: **Ռեգուլյարության պայմանը:** Եթե գոյություն ունի այնպիսի $\bar{x}$ վեկտոր, որ

$$f_i(\bar{x}) < 0, \quad i \in [1 : m],$$

ապա λ_0 գործակիցը կարելի է վերցնել հավասար մեկի:

Իրոք, եթե $\lambda_0 = 0$, ապա թեորեմ 3.1-ի 1) և 2) եզրակացություններից հետևում է, որ

$$\sum_{i=1}^m \lambda_i f_i(\bar{x}) \geq 0,$$

որը հակասություն է, քանի որ

$$\sum_{i=1}^m \lambda_i f_i(\bar{x}) < 0:$$

Նշենք, որ ռեգուլյարության պայմանը ուռուցիկ ծրագրավորման խնդրում նշանակում է, որ M բազմությունը ունի ներքին կետ:

Մասնավոր դեպքեր: Դիցուք ուռուցիկ ծրագրավորման խնդիրը ունի հետևյալ տեսքը՝

$$f_0(x) \rightarrow \min, \quad f_i(x) = (a_j, x) - b_j = 0, \quad j = 1, 2, \dots, m, \quad x_i \geq 0, \quad i = 1, 2, \dots, n:$$

Այդ դեպքում Կուն - Թակերի պայմանները կընդունեն հետևյալ տեսքը՝

ա) $\frac{\partial L(x, \lambda)}{\partial x_i} \geq 0, \quad i = 1, 2, \dots, n,$

բ) $x_i \frac{\partial L(x, \lambda)}{\partial x_i} = 0, \quad i = 1, \dots, n,$

գ) $x_i \geq 0, \quad i = 1, 2, \dots, n:$

Այստեղ $L(x, \lambda) = f_0(x) + \sum_{j=1}^m \lambda_j f_j(x):$

4

Ռեգրեսիա և ռեգուլյարիզացիա

4.1 Գծային ռեգրեսիայի խնդիրը

Գծային ռեգրեսիայի խնդիրը մեկ կամ մի քանի փոփոխականի կախվածության մոդել է մի ուրիշ փոփոխականից: Այդ կախվածությունը արտահայտվում է գծային ֆունկցիայի միջոցով:

Դիտարկենք պրակտիկ մի օրինակ: Ասենք ունենք տվյալներ պաղպաղակի վաճառքի և վաճառքի օրերին օդի ջերմաստիճանի վերաբերյալ: Կառուցենք այդ կախվածության գրաֆիկը: t առանցքը թող լինի օդի ջերմաստիճանը, իսկ $a(t)$ առանցքը՝ վաճառված պաղպաղակի քանակը: Ելնելով այդ էմպիրիկ տվյալներից՝ դուրս բերենք վաճառվող պաղպաղակի քանակի կախվածությունը օդի ջերմաստիճանից: Դրանից հետո մենք այլևս վաճառքի տվյալների կարիք չենք ունենա: Այսինքն՝ իմանալով ջերմաստիճանը և օգտագործելով ուղղակի կախվածության բանաձևը՝ կարող ենք կանխատեսել՝ քանի պաղպաղակ կարելի է վաճառել տարբեր ջերմաստիճանների դեպքում: Ենթադրենք այդ կախվածությունը արտահայտվում է գծային ֆունկցիայի տեսքով՝

$$a(t) = w_0 + w_1 t,$$

որտեղ t -ին օդի ջերմաստիճանն է, w_0 -ն և w_1 -ը մոդելի պարամետրներն են, իսկ $a(t)$ -ն վաճառված պաղպաղակի քանակը:

Որպես կորստի ֆունկցիա այստեղ կարելի է վերցնել միջին քառակուսային շեղումը՝

$$Q(a, t) = \frac{1}{l} \sum_{i=1}^l ((a(t_i) - y_i)^2,$$

որտեղ y_i -ն վաճառված պաղպաղակի քանակն է t_i ջերմաստիճանի դեպքում: Խնդիրը այն է, որ պետք է մինիմիզացնել կորստի այս ֆունկցիան ըստ w_0 և w_1 պարամետրերի: Մինիմիզացիայի միջոցով գտնում ենք այդ պարամետրերը և $a(t)$ ֆունկցիայի օգնությամբ կատարում ենք կանխատեսում:

Այժմ ընդհանրացնենք այս մոդելը: Դիցուք տրված է նմուշների ինչ-որ հավաքածու՝

$$S = \{(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)\} \subseteq \mathbb{R}^n \times \mathbb{R} :$$

Պետք է կառուցել (կանխատեսել) $f(x)$ ֆունկցիա՝ իմանալով, որ $y_i = f(x_i)$, $i = 1, 2, \dots, l$: Գծային ռեգրեսիայի դեպքում փնտրվում է գծային $a_S(x) = (w, x) - w_0$ տեսքի ֆունկցիա, այնպիսին, որ

$$Q(a_S) = \frac{1}{l} \sum_{i=1}^l L(a_S, x_i) = \frac{1}{l} \sum_{i=1}^l (a_S(x_i) - y_i)^2 \rightarrow \min:$$

Նկատենք, որ $a_S(x_i) = (w, x_i) - w_0 = (\mathbf{w}, \mathbf{x}_i)$, որտեղ $\mathbf{w} = (w, w_0)$, $\mathbf{x}_i = (x_i, -1)$: Համաձայն այս նշանակումների՝ պետք է գտնել այնպիսի $\mathbf{w}$ վեկտոր, որը մինիմում հաղորդի

$$Q((\mathbf{w}) \equiv \frac{1}{l} \sum_{i=1}^l ((\mathbf{w}, \mathbf{x}_i) - y_i)^2$$

ֆունկցիային (միջին քառակուսային շեղմանը):

Վեկտորական տեսքով սա կարելի է գրել հետևյալ ձևով՝

$$Q(\mathbf{w}) = \frac{1}{l} \|X\mathbf{w} - Y\|^2 \rightarrow \min, \quad (4.1)$$

որտեղ

$$X = \begin{pmatrix} \mathbf{x}_1 \\ \dots \\ \mathbf{x}_l \end{pmatrix}, \quad Y = \begin{pmatrix} y_1 \\ \dots \\ y_l \end{pmatrix} :$$

Ստուգենք, որ $Q(\mathbf{w})$ ֆունկցիան ուռուցիկ է: Իրոք,

$$\begin{aligned} \|X(\alpha\mathbf{w} + (1-\alpha)\mathbf{w}' - Y\|^2 &= \|\alpha(X\mathbf{w} - Y) + (1-\alpha)(X\mathbf{w}' - Y)\|^2 = \\ &= \alpha^2 \|X\mathbf{w} - Y\|^2 + 2\alpha(1-\alpha)(X\mathbf{w} - Y, X\mathbf{w}' - Y) + (1-\alpha)^2 \|X\mathbf{w}' - Y\|^2 \leq \\ &\leq \alpha^2 \|X\mathbf{w} - Y\|^2 + \alpha(1-\alpha)(\|X\mathbf{w} - Y\|^2 + \|X\mathbf{w}' - Y\|^2) + (1-\alpha)^2 \|X\mathbf{w}' - Y\|^2 = \\ &= \alpha \|X\mathbf{w} - Y\|^2 + (1-\alpha) \|X\mathbf{w}' - Y\|^2: \end{aligned}$$

Քանի որ Q -ն ուռուցիկ ֆունկցիա է, ապա $\mathbf{w}$ վեկտորը մինիմումի կետ է միայն այն դեպքում, երբ

$$\frac{\partial Q}{\partial \mathbf{w}_j} = 0, \quad j = 0, 2, \dots, l:$$

Վեկտորական տեսքով սա կարելի է ներկայացնել

$$X\mathbf{w}X = YX$$

հավասարության տեսքով, կամ որ նույնն է՝

$$X^T X\mathbf{w} = X^T Y: \quad (4.2)$$

Այստեղից՝ եթե $X^T X$ մատրիցը վերասերվող չէ, ապա

$$\mathbf{w} = (X^T X)^{-1} X^T Y : \quad (4.3)$$

4.2 Տուգանային ֆունկցիաների մեթոդը

Տուգանային ֆունկցիաների մեթոդով պայմանական օպտիմիզացիայի խնդիրները վերածվում են ոչ պայմանական օպտիմիզացիայի խնդիրների։ Կազմվում են պարամետրից կախված ֆունկցիաների ընտանիք, որոնք մինիմիզացվում են ամբողջ տարածության վրա։ Ստացված մինիմումի կետերի հաջորդականությամբ մոտարկվում են սկզբնական խնդրի օպտիմալ լուծումները։

Սահմանում 4.1. $P(x)$ անընդհատ ֆունկցիան կոչվում է տուգանք M փակ բազմության համար, եթե

- 1) $P(x) = 0$ այն և միայն այն դեպքում, երբ $x \in M$,
- 2) $P(x) > 0$, երբ $x \notin M$:

Թեորեմ 4.1. դիցուք՝

- 1) $\{c_k\}$ դրական թվերի այնպիսի հաջորդականություն է, որ

$$c_{k+1} > c_k, \quad c_k \rightarrow +\infty:$$

- 2) x_k կետը $q(c_k, x) \equiv f(x) + c_k P(x)$ ֆունկցիայի գլոբալ մինիմումի կետն է $\mathbb{R}^n$ -ի վրա։

Այդ դեպքում x_k հաջորդականության սահմանային կետը f -ի մինիմումի կետ է M բազմության վրա։

Ապացույց. Դիցուք x^* -ը f -ի մինիմումի կետն է M բազմության վրա։ Այդ դեպքում

- 1) $q(c_k, x_k) \leq q(c_{k+1}, x_{k+1})$;
- 2) $P(x_k) \geq P(x_{k+1})$;
- 3) $f(x_k) \leq f(x_{k+1})$;
- 4) $f(x^*) \geq q(c_k, x_k) \geq f(x_k)$;

Իրոք, ունենք

- 1) $q(c_{k+1}, x_{k+1}) = f(x_{k+1}) + c_{k+1} P(x_{k+1}) \geq f(x_{k+1}) + c_k P(x_{k+1}) \geq f(x_k) + c_k P(x_k) = q(c_k, x_k)$;

$$2) f(x_k) + c_k P(x_k) \leq f(x_{k+1}) + c_k P(x_{k+1}) \text{ և } f(x_{k+1}) + c_{k+1} P(x_{k+1}) \leq f(x_k) + c_{k+1} P(x_k) :$$

Այս երկու անհավասարություններից կատանանք

$$(c_{k+1} - c_k)P(x_k) \geq (c_{k+1} - c_k)P(x_{k+1}) :$$

Հետևաբար՝ $P(x_k) \geq P(x_{k+1})$:

$$3) \text{ Ունենք } f(x_{k+1}) + c_k P(x_{k+1}) \geq f(x_k) + c_k P(x_k) \text{ և } P(x_k) \geq P(x_{k+1}) :$$

Հետևաբար՝ $f(x_{k+1}) \geq f(x_k)$:

$$4) f(x_k) \leq f(x_k) + c_k P(x_k) \leq f(x^*) + c_k P(x^*) = f(x^*) :$$

Դիցուք այժմ $\bar{x}$ -ը $\{x_k\}$ հաջորդականության սահմանային կետն է: Դա նշանակում է, որ $f(x_{k_m}) \rightarrow f(\bar{x})$: Հետևաբար, համաձայն (4)-րդ հատկության՝ ունենք

$$\lim_{m \rightarrow \infty} c_{k_m} P(x_{k_m}) = \lim_{m \rightarrow \infty} [q(c_{k_m}, x_{k_m}) - f(x_{k_m})] \equiv q^* - f(\bar{x}) :$$

Բայց քանի որ $c_{k_m} \rightarrow +\infty$, ապա

$$\lim_{m \rightarrow \infty} P(x_{k_m}) = 0 :$$

Այսինքն՝ $P(\bar{x}) = 0$: Մյուս կողմից $f(x_{k_m}) \leq f(x^*)$: Անցնելով սահմանի՝ կատանանք $f(\bar{x}) \leq f(x^*)$: Իսկ սա նշանակում է, որ $\bar{x}$ -ը f ֆունկցիայի մինիմումի կետն է M բազմության վրա:

Նորից քննարկենք գծային ռեգրեսիայի խնդիրը: Դիցուք ենթադրենք, որ X մատրիցի տողերը գծորեն կախված են: Դա նշանակում է, որ գոյություն ունի այնպիսի $\mathbf{v} \neq 0$ վեկտոր, որ $X\mathbf{v} = 0$: Իսկ, եթե $\mathbf{w}^*$ վեկտորը (4.2) հավասարման լուծումն է, ապա ցանկացած ϵ թվի համար կունենանք

$$X^T X(\mathbf{w}^* + \epsilon \mathbf{v}) = X^T Y,$$

կնշանակի՝ $\mathbf{w}^* + \epsilon \mathbf{v}$ վեկտորը նույնպես լուծում է: Իսկ դա նշանակում է, որ օպտիմալ լուծումների բազմությունը սահմանափակ չէ, և հետևաբար խնդիրը կայուն չէ:

Երբեմն գծային ռեգրեսիայի խնդրում $\mathbf{w}$ պարամետրի վրա դրվում է սահմանափակում, և (4.1) խնդրի փոխարեն դրվում է հետևյալ խնդիրը՝

$$Q(\mathbf{w}) \rightarrow \min, \|\mathbf{w}\|^2 \leq m : \quad (4.4)$$

Այս խնդիրը կարելի է լուծել տուգանային ֆունկցիաների մեթոդով՝ որպես տուգանք վերցնելով հետևյալ ֆունկցիան՝

$$P(\mathbf{w}) = [\max(0, (\|\mathbf{w}\|^2 - m))]^2 :$$

Կազմենք $c > 0$ պարամետրից կախված

$$G_c(\tilde{w}) \equiv Q(\mathbf{w}) + c[\max(0, \|\mathbf{w}\|^2 - m)]^2$$

ֆունկցիան: Համաձայն տուգանային ֆունկցիաների մեթոդի՝ c պարամետրի մեծ արժեքների դեպքում $\mathbf{w}(\alpha)$ վեկտորը մոտ է (4.4)-րդ խնդրի որևէ լուծման:

Այժմ նկարագրենք տուգանքային ֆունկցիաների մեթոդի ընդհանուր ընթացակարգը: Այդ ալգորիթմը կազմված է հետևյալ քայլերից.

- Վերցնում ենք մոնոտոն աճող դրական թվերի հաջորդականություն՝

$$c_{k+1} > r_k, \quad c_k \rightarrow +\infty:$$

Դիցուք $x(c_0) = x^0$, $\epsilon > 0$:

- k -րդ իտերացիայի ժամանակ լուծում ենք ոչ պայմանական օպտիմիզացիայի

$$x(c_k) = \arg \min_{x \in \mathbb{R}^n} (f(x) + c_k P(x))$$

խնդիրը՝ որպես սկզբնական մոտավորություն վերցնելով $x(c_{k-1})$ -ն:

- Եթե $P(x(c_k)) < \epsilon$, ապա ընդունում ենք $x(c_k)$ -ն որպես խնդրի լուծում և կանգ ենք առնում: Հակառակ դեպքում k -ն ավելացնում ենք մեկով և վերադառնում նախորդ քայլին:

Քանի որ յուրաքանչյուր իտերացիայում ոչ պայմանական օպտիմիզացիայի խնդրի լուծման համար պետք է կիրառել մոտավոր մի մեթոդ, օրինակ՝ գրադիենտային իջեցում, ապա որպես $x(c_k)$ պետք է վերցնել այդ մոտավոր արժեքները:

Հաջորդ երկու պարագրաֆները նվիրված են Տիխոնովի ռեգուլյարիզացիայի մեթոդի շարադրմանը և այդ ալգորիթմների գույքամիտության հարցերին:

4.3 Կանոնավորման (ռեգուլյարիզացիայի) մեթոդը

Ինչպես տեսանք, գծային ռեգրեսիայի խնդրի օպտիմալ լուծումների բազմությունը ընդհանուր դեպքում սահմանափակ չէ: Երբեմն պետք է լինում այդ լուծումների բազմությունից գտնել այն կետը, որը ավելի մոտիկ է նախօրոք ֆիքսված ինչ-որ կետի: Այդ խնդիրը լուծվում է կարգավորիչների (ռեգուլյարիզատորների) օգնությամբ: Ըստ էության ոչ կոռեկտ խնդրի օպտիմալ լուծումները մոտարկվում են կոռեկտ խնդիրներ լուծումների հաջորդականությամբ: Ռեգուլյարիզացիայի տեսությունը առաջին անգամ մշակվել է Ա. Ն. Տիխոնովի կողմից:

Դիտարկենք $f(x) \rightarrow \min$, $x \in \mathbb{R}^n$ խնդիրը: Ներմուծենք հետևյալ ֆունկցիան, որը կոչվում է ռեգուլյարիզացված ֆունկցիա՝

$$N_\alpha(x) = f(x) + \alpha \Omega(x),$$

որտեղ $\Omega(x) = \|x - x_0\|^2$ (x_0 -ն նախօրոք ֆիքսված վեկտոր է): Ω -ն կոչվում է կարգավորիչ (ռեգուլյարիզատոր): Նշանանակենք

$$y_\alpha = \arg \min \{N_\alpha(x) : x \in \mathbb{R}^n\}, \quad N_\alpha^* = N_\alpha(y_\alpha):$$

Դիցուք $\tilde{y}_\alpha$ կետը բավարարում է

$$N_\alpha(\tilde{y}_\alpha) \leq N_\alpha^* + \xi(\alpha) \quad (4.5)$$

անհավասարությանը: Տեղի ունի գուգամիտության հետևյալ թեորեմը:

Թեորեմ 4.2. Դիցուք $f(x)$ -ը անընդհատ ուռուցիկ ֆունկցիա է, և նրա մինիմումի կետերի բազմությունը դատարկ չէ: Նշանակենք այն M^* -ով:

Եթե

$$\lim_{\alpha \rightarrow 0} \frac{\xi(\alpha)}{\alpha} = 0,$$

ապա

$$\lim_{\alpha \rightarrow 0} y_\alpha = y_0,$$

որտեղ y_0 -ն այնպիսի վեկտոր է, որ

$$y_0 = \arg \min \{\Omega(x) : x \in M^*\}:$$

Ապացույց: Քանի որ $f(x) \geq f(y)$ ($y \in M^*$), ապա $N_\alpha(x) \rightarrow +\infty$, երբ $\|x\| \rightarrow +\infty$: Հետևաբար՝ $\{x \in \mathbb{R}^n / N_\alpha(x) \leq C\}$ մակարդակի բազմությունները սահմանափակ են: Ուստի (4.5) պայմանից կարելի է գտնել y_α վեկտորը:

Ունենք

$$\begin{aligned} f(y_0) &\leq f(\tilde{y}_\alpha) \leq f(\tilde{y}_\alpha) + \alpha \Omega(\tilde{y}_\alpha) \leq N_\alpha^* + \xi(\alpha) \leq \\ &\leq f(y_\alpha) + \alpha \Omega(y_\alpha) + \xi(\alpha) \leq f(y_0) + \Omega(y_0) + \xi(\alpha) \leq f(\tilde{y}_\alpha) + \alpha \Omega(y_0) + \xi(\alpha) : \end{aligned}$$

Այստեղից ստանում ենք

$$\Omega(\tilde{y}_\alpha) \leq \Omega(y_0) + \frac{\xi(\alpha)}{\alpha}, \quad (4.6)$$

$$f(y_0) \leq f(\tilde{y}_\alpha) \leq f(y_0) + \alpha \Omega(y_0) + \xi(\alpha):$$

Այս երկու անհավասարություններից հետևում է, որ $\tilde{y}_\alpha$ փոխվում է կոմպակտի վրա, և $f(\tilde{y}_\alpha) \rightarrow f(y_0)$: Այժմ ենթադրենք, որ $\tilde{y}_{\alpha_k} \rightarrow \bar{y}$: (4.6) անհավասարությունից կստանանք

$$\Omega(\tilde{y}_\alpha) \rightarrow \Omega(\bar{y}) \leq \Omega(y_0) :$$

Իսկ f ֆունկցիայի անընդհատությունից հետևում , որ

$$f(\tilde{y}_{\alpha_k}) \rightarrow f(\bar{y}) = f(y_0):$$

Հետևաբար՝ $\Omega(y_0) = \Omega(\bar{y})$: Բայց քանի որ Ω -ն ուժեղ ուռուցիկ ֆունկցիա է, ապա այն M^* փակ ուռուցիկ բազմության վրա կարող է հասնել մինիմումի միայն մեկ կետում (տես օրինակ՝ [50], թեորեմ 1.1.7): Այստեղից ստանում ենք, որ $\tilde{y}_\alpha \rightarrow y_0$:

Թվային օրինակի վրա մեկնաբանենք ռեգուլյարիզացիայի մեթոդը: Դիցուք

$$f(x) \equiv (w_2 - w_1 x)^2:$$

Որպես Ω ռեգուլյարիզատոր վերցնենք $\Omega(x) = |x|^2$ ֆունկցիան, որը գրականության մեջ կոչվում է L_2 ռեգուլյարիզատոր: Այս դեպքում կունենանք՝

$$N_\alpha(x) = f(x) + \alpha x^2:$$

Ածանցելով այս ֆունկցիան և հավասարեցնելով այն զրոյի՝ կստանանք

$$x_\alpha = \frac{w_1 \cdot w_2}{w_1^2 + \alpha}:$$

Ինչպես երևում է բանաձևից, երբ $\alpha \rightarrow 0$, ապա

$$x_\alpha \rightarrow \frac{w_2}{w_1},$$

որը $f(x)$ ֆունկցիայի մինիմումի կետն է:

Այժմ դիտարկենք, այսպես ասած, $L_1 = |x|$ ռեգուլյարիզատորի դեպքը: Ունենք

$$N_\alpha(x) = f(x) + \alpha|x|:$$

Դիցուք $x > 0$: Ածանցելով $N_\alpha(x)$ ֆունկցիան և այն հավասարեցնելով զրոյի՝ կստանանք

$$x_\alpha = \frac{2w_1 w_2 - \alpha}{2w_1^2}:$$

Իսկ եթե $x < 0$, ապա

$$x_\alpha = \frac{\alpha + 2w_1 w_2}{2w_1^2}:$$

Այսպիսով այս ռեգուլյարիզատորի դեպքում նույնպես

$$x_\alpha \rightarrow \frac{w_2}{w_1}, \text{ երբ } \alpha \rightarrow 0:$$

Նորից դիտարկենք գծային ռեգրեսիայի (4.1) խնդիրը: Կատարենք այստեղ ռեգուլյարիզացիա՝ որպես ռեգուլյարիզատոր վերցնելով $\Omega(\mathbf{w}) = \|\mathbf{w}\|^2$ ֆունկցիան: Նշանակենք $\mathbf{w}(\alpha)$ -ով

$$N_\alpha(\mathbf{w}) = Q(w) + \alpha\|\mathbf{w}\|^2$$

ֆունկցիայի մինիմումի կետը: Այդ դեպքում $\mathbf{w}(\alpha)$ -ն

$$X^T X \mathbf{w} - X^T Y + 2\alpha \mathbf{w} = 0$$

հավասարման արմատն է: Դժվար չէ ցույց տալ, որ $X^T X + 2\alpha E$ մատրիցը

չվերասերվող է ցանկացած $\alpha > 0$ թվի դեպքում: Հետևաբար՝

$$\mathbf{w}(\alpha) = (X^T X + 2\alpha E)^{-1} X^T Y:$$

Այսպիսով, (4.1) խնդրի լուծումը մոտարկվում է կայուն խնդիրների լուծումներով (խնդիրներ, որոնք ունեն միակ լուծումներ, և որոնք անընդհատ են կախված պարամետրերից):

4.4 Ժամանակային շարքեր և կանխատեսում

Պատահական մեծությունների $\{x_t, t \in N\}$ հաջորդականությունը կոչվում է ժամանակականային շարք: Ժամանակականային շարքը կոչվում է ստացիոնար, եթե

$$1 \quad E[x_t] = \text{const}:$$

$$2 \quad \text{cov}(x_t, x_{t+h}) = \gamma(h) \quad \forall t:$$

Սա նշանակում է, որ ստացիոնար պրոցեսի միջին արժեքը չի փոխվում ժամանակի ընթացքում: Իսկ ժամանակականային շարքի երկու անդամների կովարիացիան կախված է միայն այն բանից, թե որքան են նրանք իրարից հեռու ժամանակի առումով:

Դիտարկենք մի օրինակ: Դիցուք

$$x_t = \xi_1 \cos t + \xi_2 \sin t,$$

որտեղ ξ_1, ξ_2 պատահական մեծությունները իրարից անկախ են, բաշխված են միևնույն ձևով, ընդ որում՝ $p(\xi_1 = 1) = p(\xi_1 = -1) = \frac{1}{2}$: Այդ դեպքում

$$E[x_t] = 0, \quad \text{cov}(x_t, x_s) = \cos t \cos s D\xi_1 + \sin t \sin s D\xi_2 = \cos(t - s):$$

Ցույց տանք, որ $\gamma(h)$ ֆունկցիան գույգ է: Իրոք, ունենք

$$\gamma(-h) = \text{cov}(x_t, x_{t-h}) = \text{cov}(x_{t-h}, x_t) = \text{cov}(x_\theta, x_{\theta+h}) = \gamma(h):$$

Այժմ հաշվենք կորելացիոն գործակիցները՝

$$\rho(h) \equiv \text{corr}(x_t, x_{t+h}) = \frac{\text{cov}(x_t, x_{t+h})}{\sqrt{\text{Var}(x_t)\text{Var}(x_{t+h})}} = \frac{\gamma(h)}{\sqrt{\gamma(0)\gamma(0)}} = \frac{\gamma(h)}{\gamma(0)}:$$

Այսինքն՝ ժամանակային շարքի երկու անդամների կորելացիոն գործակիցը կախված է ժամանակային առումով նրանց հեռավորությունից: $\rho(h)$ ֆունկցիան կոչվում է ավտոկորելացիոն ֆունկցիա (autocorrelation function, ACF): ճշմարիտ է հետևյալ պնդումը:

Թեորեմ 4.3. Ստացիոնար $x_t, t \in N$ ժամանակականային շարքի համար

$$\lim_{h \rightarrow \infty} \rho(h) = 0:$$

Սա նշանակում է, որ եթե էլեմենտների հեռավորությունը անվերջ մեծանում է, ապա նրանց կորելացիոն գործակիցը ձգտում է զրոյի:

$\{u_t\}$, $t \in N$ շարքը կոչվում է սպիտակ աղմուկ (white noise), եթե

$$1 \quad E[u_t] = 0,$$

$$2 \quad Var(u_t) = \sigma^2,$$

$$3 \quad cov(u_t, u_{t+h}) = 0, \quad h \neq 0:$$

Այս դեպքում գրում են, որ $u_t \sim WN(0, \sigma^2)$: Դիտարկենք ստացիոնար հետևյալ մոդելը՝

$$x_t = \mu + \sum_{j=1}^p f_j x_{t-j} + u_t + \sum_{s=1}^q \theta_s u_{t-s},$$

$$u_t \sim WN(0, \sigma^2):$$

Այս մոդելը կոչվում է ARMA(p, q): Դիտարկենք հետևյալ օպերատորը՝ $L(x_t) = x_{t-1}$: Օպերատորային տեսքով մոդելը կունենա հետևյալ տեսքը՝

$$(1 - \sum_{j=1}^p f_j L^j) x_t = \mu + (1 + \sum_{s=1}^q \theta_s L^s) u_t:$$

Դիտարկենք հետևյալ բազմանդամները՝

$$F(z) = 1 - \sum_{j=1}^p f_j z^j = 1 - f_1 z - f_2 z^2 - \dots - f_p z^p,$$

$$\theta(z) = 1 + \sum_{s=1}^q \theta_s z^s = 1 + \theta_1 z + \dots + \theta_q z^q:$$

Այդ դեպքում մոդելը օպերատորային տեսքով ներկայացվում է հետևյալ հավասարությամբ՝

$$F(L)x_t = \mu + \theta(L)u_t:$$

Ճշմարիտ է հետևյալ պնդումը:

Թեորեմ 4.4. ARIMA մոդելը կլինի ստացիոնար այն և միայն այն դեպքում, երբ $F(z)$ բազմանդամի բոլոր արմատները մոդուլով մեծ են մեկից:

Օրինակ՝ դիտարկենք ARIMA(1, q) մոդելը՝

$$x_t = \mu + f x_{t-1} + u_t + \sum_{s=1}^q \theta_s u_{t-s}, \quad \theta_q \neq 0:$$

Այդ դեպքում $F(z) = 1 - fz$: Հետևաբար $|f| < 1$ պայմանի դեպքում մոդելը կլինի ստացիոնար: Հաշվենք $E[x_t]$ -ի արժեքը: Քանի որ $E[u_t] = 0$, $E[x_t] = const$, ապա

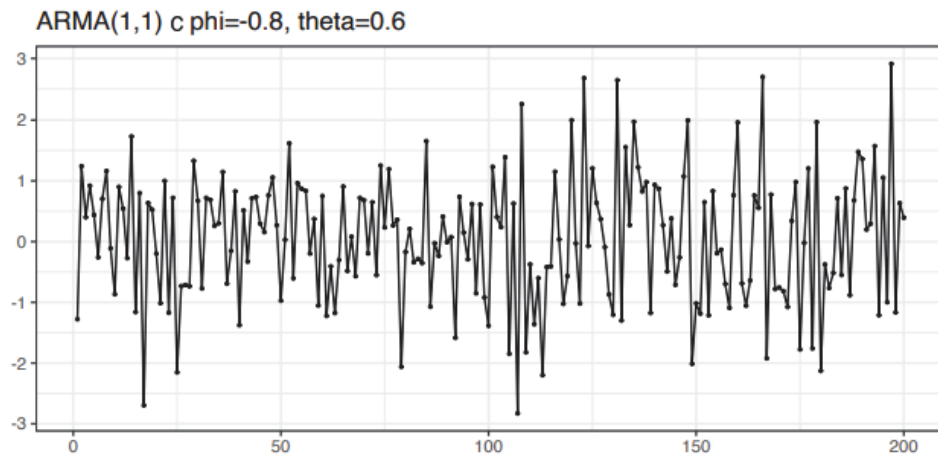
$$E[x_t] = \mu + \sum_{j=1}^p f_j E[x_{t-j}] + e u_t + \sum_{s=1}^q \theta_s E[u_{t-s}] =$$

$$\mu + \sum_{j=1}^p f_j E[x_t]:$$

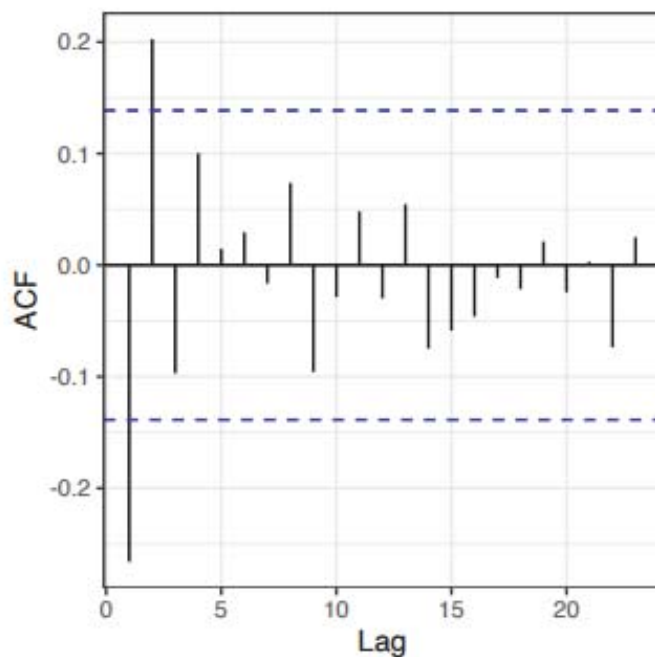
Հետևաբար՝

$$E[x_t] = \frac{\mu}{1 - \sum_{j=1}^p f_j} = \frac{\mu}{F(1)}:$$

Նկ 4.1-4.2 պատկերված են ARMA(1,1) մոդելը և ACF-ի գրաֆիկը:



Նկ. 4.1: ARMA(1,1)



Նկ. 4.2: ACF

Այժմ քննարկենք, թե ինչպես է կատարվում կանխատեսում այդ մոդելով: Ենթադրենք՝ հայտնի են (x_t, u_t) , $t = 1, 2, \dots, T$ պարամետրերը: Կատարենք կանխատեսումներ $T + \tau$, $\tau = 1, 2, \dots$ ժամանակների համար: Որպես չափանիշ

վերցնենք միջին քառակուսային շեղման ֆունկցիան՝

$$E[\hat{x}_{T+\tau} - x_{T+\tau}]^2 \rightarrow \min:$$

Եթե $\tau = 1$, ապա

$$x_{T+1} = \mu + f_1 x_T + \dots + f_p x_{T+1-p} + u_{T+1} + \theta_1 u_T + \dots + \theta_q u_{T+1-q}:$$

Այդ դեպքում կարելի ցույց տալ, որ օպտիմալ կանխատեսումը $T + 1$ պահին հետևյալն է՝

$$\hat{x}_{T+1} = \mu + f_1 x_T + \dots + f_p x_{T+1-p} + \theta_1 u_T + \dots + \theta_q u_{T+1-q}:$$

Նշենք, որ այստեղ բացակայում է u_{T+1} անդամը: $\tau = 2$ -ի համար ունենք

$$\begin{aligned} x_{T+2} = & \mu + f_1 x_{T+1} + f_2 x_T + \dots + f_p x_{T+2-p} + \\ & + u_{T+2} + \theta_1 u_{T+1} + \theta_2 u_T + \dots + \theta_q u_{T+2-q}: \end{aligned}$$

Այս դեպքում օպտիմալ կանխատեսումը կարտահայտվի հետևյալ բանաձևով՝

$$\hat{x}_{T+2} = \mu + f_1 \hat{x}_{T+1} + f_2 x_T + \dots + f_p x_{T+2-p} + \theta_2 u_T + \dots + \theta_q u_{T+2-q}:$$

Նկատենք, որ այստեղ բացակայում են u_{T+2} , u_{T+1} անդամները, իսկ x_{T+1} անդամը փոխարինվել է $\hat{x}_{T+1}$ կանխատեսումով: Անալոգիայով կատարվում են կանխատեսումներ $T + 3, T + 4, \dots$ ժամանակներում: Տեղի ունի հետևյալ թեորեմը:

Թեորեմ 4.5. $\lim_{\tau \rightarrow \infty} \hat{x}_{T+\tau} = E[x_t]$:

ԽՆԴԻՐՆԵՐ

- 1 Տրված է ուսուցանող հավաքածու (աղյուսակ 4.3): Նվազագույն քառակուսիների մեթոդով կառուցել $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$ տեսքի մոդել:
- 2 Կառուցել նույն տիպի մոդել ինչպես առաջին օրինակում՝ օգտագործելով $\Omega(x) = \|x\|^2$ ռեգուլյատորը:
- 3 . Դիցուք ունենք տվյալներ, որոնք ներկայացված են 4,4 աղյուսակում: $ARMA(1, 1)$ մոդելով կատարեք կանխատեսումներ $t, t + 1, t + 2$ պահերին:

$$x_t = 0,6x_{t-1} + u_t - 0,6u_{t-1}:$$

x	1	1	0	0	-1
y	4	4	0	2	6

Նկ. 4.3: Ռեգրեսիա

time	t	$t - 1$	$t - 2$
x	-0.5	1.2	-0.9
u	-0.2	0.3	-0.4

Նկ. 4.4: Տվյալներ

5

Ֆունկցիայի մինիմիզացիայի գրադիենտային և սուբգրադիենտային մեթոդներ

Գրադիենտային մեթոդը դասվում է դիֆերենցելի ֆունկցիաների մինիմիզացիայի թվային հիմնական մեթոդների շարքին: Այդ մեթոդի էությունը շատ պարզ է: Եթե f ֆունկցիան դիֆերենցելի է, ապա նրա հակագրադիենտը՝ $h = -f'(x)$ վեկտորը, յուրաքանչյուր x կետում ցույց է տալիս ֆունկցիայի **նվազման ուղղությունը**:

Դա հիմք է տալիս ենթադրելու, որ կամայական սկզբնական x^0 կետից սկսվող $x^{k+1} = x^k - \alpha_k f'(x^k)$ ($\alpha_k > 0$) ռեկուրենտ առնչությամբ կառուցված հաջորդականության անդամները մեծ k ինդեքսների դեպքում մոտ կլինեն f -ի մինիմումի կետին: Այստեղ այս ենթադրությունը հիմնավորվում է որոշ դասի ֆունկցիաների և α_k թվերի հատուկ եղանակներով ընտրությունների դեպքերում: Ֆունկցիայի մինիմիզացիայի բազմազան ալգորիթմների և նրանց պրակտիկ իրականացումներին կարելի է ծանոթանալ [13, 37-40, 43, 47] աշխատանքներում:

5.1 Ֆունկցիայի մինիմիզացիայի գրադիենտային եղանակների ընդհանուր նկարագրությունը

Սահմանում 5.1. h վեկտորը կոչվում է f ֆունկցիայի նվազման ուղղություն x կետում, եթե բավականաչափ փոքր դրական α թվերի համար տեղի ունի

$$f(x + \alpha h) < f(x)$$

անհավասարությունը:

Այլ խոսքով, h -ը այն վեկտորն է, որի ուղղությամբ բավականաչափ փոքր շարժվելիս կարելի է ֆունկցիայի արժեքները փոքրացնել:

Լեմմա 5.1. Դիցուք f -ը դիֆերենցելի է x կետում, և h -ն այնպիսի վեկտոր է, որ

տեղի ունի

$$(f'(x), h) < 0$$

անհավասարությունը: Այդ դեպքում h -ը f -ի նվազման ուղղություն է x կետում:

Ապացույց: Քանի որ f -ը դիֆերենցելի է, ապա

$$f(x + \alpha h) = f(x) + \alpha[(f'(x), h) + \frac{o(\alpha)}{\alpha}]:$$

Այստեղից բավականաչափ փոքր դրական α թվերի համար միջակ փակագծերում գրած արտահայտությունը դառնում է բացասական: Այնպես որ

$$f(x + \alpha h) < f(x):$$

Դիտողություն: Մասնավորապես, որպես նվազման ուղղություն կարելի է վերցնել $h = -f'(x)$, որը կոչվում է հակագրադիենտ: Կարելի է ցույց տալ, որ հակագրադիենտը տալիս է ֆունկցիայի ամենաարագ նվազման ուղղությունը:

Գրադիենտային վայրէջքի մեթոդները կառուցվում են հետևյալ կերպ. ընտրվում է կամայական x^0 կետ $\mathbb{R}^n$ տարածությունից և կառուցվում է

$$x^{k+1} = x^k + \alpha_k h^k, \quad k = 0, 1, 2, \dots \quad (5.1)$$

ոեկուրենտ առնչությունը: Այստեղ $h^k = -f'(x^k)$, իսկ α_k թվերը կոչվում են քայլեր, որոնք ընտրվում են որոշակի օրինաչափությամբ: Հայտնի են քայլի ընտրության մի քանի եղանակներ, որոնք ներկայացվում են ստորև:

1. Հաստատուն քայլի եղանակ: Այս դեպքում $\alpha_k \equiv \eta$: Այսինքն՝ բոլոր իտերացիաներում քայլը չի փոխվում՝

$$x^{k+1} = x^k - \eta f'(x^k):$$

2. Քայլի ընտրության ապրիորի եղանակը: Այս դեպքում α_k քայլերը դրական թվեր են, որոնք բավարարում են հետևյալ պայմանները.

$$\sum \alpha_k = +\infty, \quad \sum \alpha_k^2 < +\infty:$$

Մասնավորապես, այս պայմանները բավարարում են

$$\alpha_k = \frac{1}{k+1}, \quad \alpha_k = \frac{\ln(k+1)}{k+1}, \quad k = 0, 1, 2, \dots$$

թվային հաջորդականությունները:

3. Քայլի ընտրության ամենաարագ վայրէջքի եղանակը: Այս դեպքում $\alpha_k > 0$ քայլերը ընտրվում են հետևյալ կերպ. կազմում են $g(\alpha) \equiv f(x^k + \alpha h^k)$ ֆունկցիան և գտնում նրա մինիմումի կետը $(0, \infty)$ միջակայքի վրա: Այսինքն՝

$$\alpha_k \equiv \arg \min_{\alpha > 0} g(\alpha):$$

Որոշ դասի ֆունկցիաների համար կարելի է ստանալ անայիտիկ բանաձևեր α_k քայլերի որոշման համար, ինչպես հետևյալ օրինակում:

Օրինակ: Դիցուք $f(x) = 1/2(\mathbf{A}x, x) + (b, x)$, որտեղ $\mathbf{A}$ -ն $(n \times n)$ -չափանի սիմետրիկ, դրական որոշյալ մատրից է, իսկ b -ն n չափանի վեկտոր է: Կարելի է ցույց տալ, որ f -ը ուռուցիկ է, ունի միակ մինիմումի կետ $\mathbb{R}^n$ -ի վրա և նրա գրադիենտը որոշվում է հետևյալ բանաձևով. $f'(x) = \mathbf{A}x + b$: Այս դեպքում ունենք

$$\begin{aligned} g(\alpha) &\equiv f(x^k + \alpha h^k) = 1/2[\mathbf{A}(x^k + \alpha h^k), x^k + \alpha h^k] + \\ &+ (b, x^k + \alpha h^k) = \alpha^2[1/2(\mathbf{A}h^k, h^k)] + \\ &+ \alpha(\mathbf{A}x^k + b, h^k) + (1/2\mathbf{A}x^k + b, x^k): \end{aligned}$$

Դժվար չէ նկատել, որ ստացված քառակուսային եռանդամը հասնում է իր փոքրագույն արժեքին $\mathbb{R}^n$ -ի վրա հետևյալ

$$\alpha_k = -\frac{(\mathbf{A}x^k + b, h^k)}{(\mathbf{A}h^k, h^k)} = -\frac{(f'(x^k), h^k)}{(\mathbf{A}h^k, h^k)} \geq 0 \quad (5.2)$$

կետում: Հետևաբար՝

$$f(x^k + \alpha_k h^k) = \min_{\alpha \geq 0} f(x^k + \alpha h^k) = \min_{\alpha \in \mathbb{R}^n} f(x^k + \alpha h^k):$$

4. Քայլի ընտրության կիսման եղանակը: Ֆիքսում ենք որևէ դրական ε թիվ $(0, 1)$ միջակայքից և $\alpha = 1$ թվի համար ստուգում ենք

$$f(x^k - \alpha f'(x^k)) - f(x^k) \leq -\varepsilon \alpha \|f'(x^k)\|^2 \quad (5.3)$$

անհավասարությունը: Եթե այն տեղի ունի, ապա α_k -ն համարում ենք հավասար α -ի: Հակառակ դեպքում α -ն կիսում ենք և նորից ստուգում (5.3) անհավասարությունը և այսպես շարունակ: Եթե որևէ p -րդ քայլում բավարարվում է (5.3) անհավասարությունը, ապա $\alpha_k = \frac{1}{2^p}$:

ԽՆԴԻՐՆԵՐ

1. Գտնել հետևյալ ֆունկցիաների էքստրեմումի կետերը ամենաարագ վայրէջքի մեթոդով.

սկզբնական x^0 կետից սկսած կառուցել $\{x^k\}$ հաջորդականությունը (5.1) ռեկուրենտ առնչությամբ և քայլերի ընտրությունը կատարել (5.2) բանաձևով: Եթե $\|f'(x^k)\| < \varepsilon_0$, ապա պրոցեսն ավարտել և x^k -ն համարել ֆունկցիայի էքստրեմումի կետ:

Այստեղ $\varepsilon_0 > 0$ թիվը նախապես տրված ճշտությունն է:

ա) $f(x_1, x_2) = 5x_1^2 + 5x_2^2 - 6x_1x_2 \rightarrow \min, x^0 = (1, 1), \varepsilon_0 = 0.01$;

բ) $-3x_1^2 - 2x_2^2 + x_2 + 6x_2 - 15 \rightarrow \max, x^0 = (0, 1), \varepsilon_0 = 0.1$;

գ) $x_1^2 + x_2^2 + 2x_1x_2 + x_2 \rightarrow \min, x^0 = (1, 0), \varepsilon_0 = 0.01$:

2. Դիցուք f -ը երկու անգամ դիֆերենցելի է x կետում, և

$$(f'(x), h) = 0, (f''(x), h) < 0:$$

Ապացուցել, որ h -ը f ֆունկցիայի նվազման ուղղությունն է:

3. Դիցուք $H(x) = f''(x)$ հեսիանը դրական որոշյալ է և $f'(x) \neq 0$: Ապացուցել, որ

$$h = -H(x)^{-1}f'(x)$$

վեկտորը f ֆունկցիայի նվազման ուղղությունն է:

Առաջադրանք: Գրել ծրագիր, որը մինիմիզացնում է $f(x) = 1/2(\mathbf{A}x, x) + (b, x)$ ֆունկցիան $\mathbb{R}^n$ -ի վրա: Այստեղ $\mathbf{A}$ -ն $(n \times n)$ -չափանի սիմետրիկ, դրական որոշյալ մատրից է, իսկ b -ն n չափանի վեկտոր է: Մուտքի տվյալներն են $\mathbf{A}, b, x^0, n, \varepsilon, \varepsilon_0$ պարամետրերը: Այստեղ ε_0 -ն ճշտությունն է: Կառուցել $\{x^k\}$ հաջորդականությունը գրադիենտային իջեցման երեք մեթոդներով և համեմատել դրանք քայլերի քանակի տեսակյունից: Եթե $\|f'(x^k)\| < \varepsilon_0$, ապա պրոցեսն ավարտել և x^k -ն համարել մինիմումի կետ: Համեմատել ստացված արդյունքները $x^* = \mathbf{A}^{-1}b$ վեկտորի հետ, որը f -ի մինիմումի կետն է:

5.2 Գրադիենտային իջեցում հաստատուն քայլի դեպքում

Դիտարկենք հետևյալ խնդիրը՝

$$f(w) \rightarrow \min, w \in \{w : \|w\|_2 \leq B\}:$$

Քննարկենք գրադիենտային իջեցման եղանակը հաստատուն քայլի դեպքում՝

$$w^1 = 0, w^{t+1} = w^t - \eta f'(w^t), t = 1, 2, \dots:$$

Նշանակենք

$$\bar{w} = \frac{1}{T} \sum_{t=1}^T w^t:$$

Տեղի ունի գրադիենտային այս մեթոդի զուգամիտության հետևյալ գնահատականը:

Թեորեմ 5.1. Դիցուք

- f -ը ուռուցիկ է, դիֆերենցելի և ρ -լիպշիցյան ֆունկցիա է:
- $w^* = \arg \min_{w: \|w\|_2 \leq B} f(w)$:

- $\eta = \sqrt{B^2/(\rho^2 T)}$:

Այդ դեպքում ճշմարիտ է հետևյալ անհավասարությունը՝

$$f(\bar{w}) \leq f(w^*) + \frac{B\rho}{\sqrt{T}} :$$

Ապացույց: Նշանակելով $v_t = f'(w^t)$ ՝ կունենանք

$$\begin{aligned} (w^t - w^*, v_t) &= \frac{1}{\eta} (w^t - w^*, \eta v_t) = \\ &= \frac{1}{2\eta} (-||w^t - w^* - \eta v_t||^2 + ||w^t - w^*||^2 + \eta^2 ||v_t||^2) = \\ &= \frac{1}{2\eta} (-||w^{t+1} - w^*||^2 + ||w^t - w^*||^2) + \frac{\eta}{2} ||v_t||^2 : \end{aligned}$$

Գումարելով այս անհավասարությունը՝ կստանանք

$$\sum_{t=1}^T (w^t - w^*, v_t) = \frac{1}{2\eta} \sum_{t=1}^T (-||w^{t+1} - w^*||^2 + ||w^t - w^*||^2) + \frac{\eta}{2} \sum_{t=1}^T ||v_t||^2 :$$

Դժվար չէ նկատել, որ այս հավասարության աջ մասի առաջին գումարելիի մեջ կրճատվում են բոլոր անդամները, բացի առաջինից և վերջինից: Այնպես որ

$$\begin{aligned} \sum_{t=1}^T (w^t - w^*, v_t) &= \frac{1}{2\eta} (||w^1 - w^*||^2 - ||w^{T+1} - w^*||^2) + \frac{\eta}{2} \sum_{t=1}^T ||v_t||^2 \leq \\ &= \frac{1}{2\eta} (||w^1 - w^*||^2) + \frac{\eta}{2} \sum_{t=1}^T ||v_t||^2 = \\ &= \frac{1}{2\eta} ||w^1 - w^*||^2 + \frac{\eta}{2} \sum_{t=1}^T ||v_t||^2 : \end{aligned}$$

Այժմ ցույց տանք, որ $||v_t|| \leq \rho$: Դիցուք $u^t = w^t + v_t/||v_t||$: Այդ դեպքում f -ի ուղղորդիչությունից հետևում է, որ

$$f(u^t) - f(w^t) \geq (w^t, u^t - v_t) = ||v_t|| :$$

Մյուս կողմից, քանի որ f -ը ρ -լիպշիցյան է, ապա

$$f(u^t) - f(w^t) \leq \rho ||u^t - w^t|| :$$

Այս երկու անհավասարություններից ստանում ենք, որ $||v_t|| \leq \rho$: Հետևաբար՝

$$f(\bar{w}) - f(w^*) = f\left(\frac{1}{T} \sum_{t=1}^T w^t\right) - f(w^*) \leq$$

$$\begin{aligned}
&\leq \frac{1}{T} \sum_{t=1}^T f(w^t) - f(w^*) \leq \\
&\leq \frac{1}{T} \sum_{t=1}^T (f(w^t) - f(w^*)) \leq \\
&\frac{1}{T} \sum_{t=1}^T (w^t - w^*, w_t) \leq \\
&\frac{1}{T} \left(\frac{\|w^*\|}{2\eta} + T\eta\rho/2 \right) \leq \frac{B\rho}{\sqrt{T}}:
\end{aligned}$$

Վերջում, եթե պետք է հաշվել քայլերի այն քանակը, որը ϵ ճշտությամբ կվերականգնի ֆունկցիայի մինիմալ արժեքը, ապա քայլերի այդ T քանակը պետք է ընտրել այնպես, որ $\frac{B\rho}{\sqrt{T}} < \epsilon$: Այստեղից կստանանք $T \geq \frac{B^2\rho^2}{\epsilon^2}$: Թեորեմն ապացուցվեց:

5.3 Կոորդինատային իջեցման մեթոդը

Դիտարկենք հետևյալ ռեկուրենտ առնչությունը՝

$$w_{k+1} = w_k - \alpha_k \frac{\partial f}{\partial w^{i_k}}(w_k) e^{i_k}, \quad (5.4)$$

որտեղ e^{i_k} -ն միավոր վեկտոր է, որի $i_k \in \{1, 2, \dots, n\}$ տեղում գրված է մեկ: Ենթադրենք f ֆունկցիայի մասնակի ածանցյալները բավարարում են հետևյալ պայմանը՝

$$\left| \frac{\partial f}{\partial w^i}(w + \Delta w^i e^i) - \frac{\partial f}{\partial w^i}(w) \right| \leq L_i \|\Delta w^i\|: \quad (5.5)$$

Նշանակենք

$$\hat{L} = \max_{i \in [1:n]} L_i:$$

Թեորեմ 5.2. Դիցուք c հաստատունով f ուժեղ ուռուցիկ ֆունկցիան բավարարում է (5.5) պայմանը, իսկ $\{w_k\}$ հաջորդականությունը կառուցվում է (5.5) ռեկուրենտ առնչությամբ: Դիցուք $\alpha_k = \frac{1}{\sqrt{k}}$, իսկ i_k -ն պատահական մեծություն է՝ հավասարաչափ բաշխված $\{1, 2, \dots, n\}$ բազմության վրա: Այդ դեպքում տեղի ունի հետևյալ անհավասարությունը՝

$$E[f(w_{k+1})] - \min_{w \in \mathbb{R}^n} f(w) \leq \left(1 - \frac{c}{n\hat{L}}\right)^k (f(w_1) - \min_{w \in \mathbb{R}^n} f(w)):$$

Ապացույց: Քանի որ ֆունկցիայի մասնակի ածանցյալները բավարարում են Լիպշիցի պայմանը ըստ կոորդինատային ուղղությունների, ապա

$$f(w_{k+1}) \leq f(w_k) + \frac{\partial f}{\partial w^{i_k}}(w_k)(w_{k+1}^{i_k} - w_k^{i_k}) + \frac{\hat{L}}{2}(w_{k+1}^{i_k} - w_k^{i_k})^2:$$

Այստեղից, հաշվի առնելով, որ $\alpha_k = 1/\hat{L}$, կստանանք

$$f(w_{k+1}) - f(w_k) \leq -\frac{1}{\hat{L}}\left(\frac{\partial f}{\partial w^{i_k}}(w_k)\right)^2 + \frac{1}{2\hat{L}}\left(\frac{\partial f}{\partial w^{i_k}}(w_k)\right)^2 = -\frac{1}{2\hat{L}}\left(\frac{\partial f}{\partial w^{i_k}}(w_k)\right)^2:$$

Հետևաբար՝

$$\begin{aligned} E_{i_k}[f(w_{k+1})] - f(w_k) &\leq -\frac{1}{2\hat{L}} E_{i_k}\left[\left(\frac{\partial f}{\partial w^{i_k}}(w_k)\right)^2\right] = \\ &= -\frac{1}{2\hat{L}} \left(\frac{1}{n} \sum_{i=1}^n \left(\frac{\partial f}{\partial w^i}\right)^2\right) = -\frac{1}{2Ln} \|f'(w)\|^2: \end{aligned} \quad (5.6)$$

Այժմ ցուց տանք, որ

$$2c(f(w) - \min_{w \in \mathbb{R}^n} f(w)) \leq \|f'(w)\|^2, \quad \forall w \in \mathbb{R}^n: \quad (5.7)$$

Դիտարկենք հետևյալ ֆունկցիան՝

$$q(\bar{w}) = f(w) + (f'(w), \bar{w} - w) + \frac{c}{2} \|w - \bar{w}\|^2:$$

Այդ դեպքում

$$\bar{w}_* = w - \frac{1}{c} f'(w)$$

վեկտորը նրա մինիմումի կետն է, իսկ մինիմալ արժեքը կլինի՝

$$q(\bar{w}_*) = f(w) - \frac{1}{2c} \|f'(w)\|^2:$$

Այստեղից, կիրառելով ուժեղ ուռուցիկ ֆունկցիայի անհավասարությունը, կստանանք

$$\begin{aligned} \min_{w \in \mathbb{R}^n} f(w) &\geq f(w) + (f'(w), \bar{w}_* - w) + \frac{c}{2} \|w - \bar{w}_*\|^2 \geq \\ &= f(w) - \frac{1}{2c} \|f'(w)\|^2: \end{aligned}$$

(5.6) անհավասարության երկու մասից հանելով $\min_{w \in \mathbb{R}^n} f(w)$ թիվը և հաշվի առնելով (5.7) անհավասարությունը (լուծելով նաև ստացված ոեկուրենտ առնչությունը)՝ կստանանք թեորեմի եզրակացությունը:

Առաջադրանք: Իրականացնել կոորդինատային իջեցման մեթոդը

$$f(w) = \frac{1}{m} \sum_{i=1}^m \ln(1 + e^{y(w, x_i)}) + \lambda \sum_{i=1}^n |w_i|$$

ֆունկցիայի համար՝ կազմելով համապատասխան կոդը Python-ով:

5.4 Քայլի կիսման եղանակի գուգամիտության թեորեմը

Լեմմա 5.2. Դիցուք f ֆունկցիան բավարարում է հետևյալ պայմանները.

1) f -ի գրադիենտը բավարարում է L իպշիցի պայմանը, այսինքն՝ գոյություն ունի այնպիսի $L > 0$ հաստատուն, որ

$$\|f'(x) - f'(y)\| \leq L\|x - y\| \quad \forall x, y \in R^n:$$

2) f -ը ներքևից սահմանափակ է, այսինքն՝ գոյություն ունի այնպիսի $m > 0$ թիվ, որ տեղի ունի

$$f(x) \geq m \quad \forall x \in R^n$$

անհավասարությունը:

Այդ դեպքում քայլի կիսման եղանակով վերջավոր քայլերի ընթացքում ընտրվում է α_k -ն, և

$$\alpha_k > (1 - \varepsilon)/L > 0, \quad k = 0, 1, 2, \dots :$$

Ապացույց: Դիցուք $h^k = -f'(x^k)$: Համաձայն L ազրանժի միջին արժեքի վերաբերյալ թեորեմի՝ գոյություն ունի $\theta \in (0, 1)$ հաստատուն, այնպիսին, որ եթե $\alpha > (1 - \varepsilon)/L$, ապա

$$\begin{aligned} f(x^k + \alpha h^k) - f(x^k) &= (f'(x^k + \alpha \theta h^k), \alpha h^k) = \\ &= (f'(x^k + \alpha \theta h^k) - f'(x^k), \alpha h^k) + \alpha (f'(x^k), h^k) \leq \\ &\leq \|L\alpha \theta h^k\| \|\alpha h^k\| - \alpha \|f'(x^k)\|^2 \leq L\alpha^2 \|f'(x^k)\|^2 - \\ &- \alpha \|f'(x^k)\|^2 = -\alpha \|f'(x^k)\|^2 (1 - \alpha L) \leq -\varepsilon \alpha \|f'(x^k)\|^2: \end{aligned}$$

Թեորեմ 5.3. Դիցուք f ֆունկցիան բավարարում է լեմմա 5.2-ի բոլոր պայմանները և $\{x^k\}$ -ն կիսման մեթոդով կառուցված հաջորդականությունն է:

Այդ դեպքում $f'(x^k) \rightarrow 0$, երբ $k \rightarrow \infty$:

Ապացույց: Ունենք

$$f(x^{k+1}) - f(x^k) \leq -\varepsilon \alpha \|f'(x^k)\|^2: \quad (5.8)$$

Այստեղից հետևում է, որ $f(x^{k+1}) \leq f(x^k)$: Այսինքն՝ $\{f(x^k)\}$ հաջորդականությունը մոնոտոն նվազող է և ներքևից սահմանափակ է, հետևաբար հաջորդականությունը զուգամետ է, այսինքն՝

$$f(x^k) - f(x^{k+1}) \rightarrow 0:$$

(5.8)-ից հետևում է, որ

$$\|f'(x^k)\|^2 \leq \frac{f(x^k) - f(x^{k+1})}{\varepsilon \alpha_k}:$$

Այստեղից՝ քանի որ, ըստ լեմմա 5.2-ի, $\alpha_k > (1 - \varepsilon)/L > 0$, ապա $f'(x^k) \rightarrow 0$:

Թեորեմ 5.4. Ենթադրենք գոյություն ունեն այնպիսի դրական $D > 0$, $d > 0$ հաստատուններ, որ

$$D\|h\|^2 \geq (f''(x)h, h) \geq d\|h\|^2 \quad \forall x, h \in \mathbb{R}^n: \quad (5.9)$$

Այդ դեպքում կհաման եղանակով կառուցված $\{x^k\}$ հաջորդականությունը գումարիտում է f -ի միակ x^* մինիմումի կետին երկրաչափական պրոգրեսիայի արագությամբ:

Այսինքն՝ գոյություն ունեն $C > 0$ և $q \in (0, 1)$ հաստատուններ, այնպիսիք, որ

$$\|x^k - x^*\| \leq Cq^k, \quad k = 0, 1, \dots :$$

Ապացույց: Քանի որ $f'(x^*) = 0$, ապա, ըստ Թեյլորի բանաձևի,

$$f(x) - f(x^*) = \frac{1}{2}(f''(x^* + \theta(x - x^*))(x - x^*), x - x^*),$$

որտեղ $\theta \in (0, 1)$: Այստեղից, հաշվի առնելով (4.5) անհավասարությունը, ստանում ենք

$$\frac{d}{2}\|x - x^*\|^2 \leq f(x) - f(x^*) \leq \frac{D}{2}\|x - x^*\|^2: \quad (5.10)$$

Հաշվի առնելով այս պայմանը և ներկայացնելով f ֆունկցիան Թեյլորի բանաձևի տեսքով x կետում՝ կունենանք

$$\begin{aligned} f(x^*) - f(x) &= (f'(x), x^* - x) + \\ &+ 1/2(f''(x + \theta(x^* - x))(x^* - x), x^* - x) \geq \\ &\geq -\|f'(x)\|\|x - x^*\| + d/2\|x - x^*\|^2: \end{aligned}$$

Այստեղից կստանանք

$$f(x) - f(x^*) \leq \|f'(x)\|\|x - x^*\| - d/2\|x - x^*\|^2: \quad (5.11)$$

Հաշվի առնելով նաև (5.10) անհավասարության ձախ մասը՝ (5.11)-ից կստանանք

$$\frac{d}{2}\|x - x^*\|^2 \leq \|f'(x)\|\|x - x^*\| - d/2\|x - x^*\|^2:$$

Հետևաբար

$$\|x - x^*\| \leq \frac{\|f'(x)\|}{d}: \quad (5.12)$$

Օգտվելով (5.10)-(5.12) անհավասարություններից՝ ստանում ենք

$$f(x) - f(x^*) \leq \frac{\|f'(x)\|^2}{d} - d/D(f(x) - f(x^*)):$$

Այստեղից

$$\|f'(x)\|^2 \geq d(1 + d/D)(f(x) - f(x^*)): \quad (5.13)$$

Կիրառելով (5.13) անհավասարությունը կհսման մերձողի

$$f(x^{k+1}) - f(x^k) \leq -\varepsilon \alpha_k \|f'(x^k)\|^2$$

անհավասարությունում՝ կստանանք

$$f(x^{k+1}) - f(x^*) \leq [1 - \varepsilon \alpha_k d(1 + d/D)](f(x^k) - f(x^*)):$$

Այստեղից, հաշվի առնելով $\alpha_k > \bar{\alpha} \equiv (1 - \varepsilon)/L > 0$ անհավասարությունը կստանանք

$$f(x^k) - f(x^*) \leq (f(x^0) - f(x^*))\bar{q}^k, \quad (5.14)$$

որտեղ

$$\bar{q} = (1 - \varepsilon)\bar{\alpha}d(1 + d/D):$$

Վերջապես, օգտվելով (5.10) և (5.14) անհավասարությունից, կստանանք

$$\|x^k - x^*\| \leq \sqrt{\frac{2}{d}} \sqrt{f(x^k) - f(x^*)} \leq Cq^k,$$

որտեղ

$$C = \sqrt{\frac{2}{d}} \sqrt{f(x^0) - f(x^*)}, \quad q = \sqrt{\bar{q}}:$$

Խնդիր: Ցույց տալ, որ

$$f(x_1, x_2, \dots, x_n) = \lambda^{-1} \ln \left(\sum_{i=1}^n \exp(\lambda x_i) \right), \quad \lambda > 0$$

ֆունկցիան բավարարում է

$$0 \leq (f''(x)h, h) \leq D\|h\|^2 \quad \forall x, h \in \mathbb{R}^n$$

պայմանը, որտեղ D -ին ինչ-որ դրական հաստատուն է:

5.5 Ապրիորի մերձողի գուգամիտությունը

Թեորեմ 5.5. Դիցուք f -ը ուռուցիկ դիֆերենցելի ֆունկցիա է՝ որոշված $\mathbb{R}^n$ -ի վրա, և M^* -ը նրա մինիմումի կետերի բազմությունն է $\mathbb{R}^n$ -ի վրա: Ենթադրենք $M^* \neq \emptyset$: Դիցուք $\{x^k\}$ -ն հետևյալ ռեկուրենտ առնչությամբ կառուցված հաջորդականությունն է՝

$$x^0 \in \mathbb{R}^n, \quad x^{k+1} = x^k - \alpha_k \frac{f'(x^k)}{\|f'(x^k)\|}, \quad k = 0, 1, 2, \dots, \quad (5.15)$$

որտեղ α_k -երը դրական թվեր են՝ բավարարող

$$\sum \alpha_k = +\infty, \quad \sum \alpha_k^2 < +\infty \quad (5.16)$$

պայմանները:

Այդ դեպքում

$$x^k \rightarrow \bar{x} \in M^* :$$

Այսինքն՝ $\{x^k\}$ հաջորդականությունը զուգամիտում է f -ի որևէ $\bar{x}$ մինիմումի կետի:

Ապացույց: Նշանակենք $v^k = f'(x^k)$: Վերցնենք որևէ $x^* \in M^*$ կետ և հաշվենք նրա հեռավորությունը $\{x^k\}$ հաջորդականության անդամներից:

Ունենք

$$\|x^{k+1} - x^*\|^2 = \|x^k - x^*\|^2 + \frac{2\alpha_k}{\|v^k\|} (v^k, x^* - x^k) + \alpha_k^2 : \quad (5.17)$$

Քանի որ x^* -ը f -ի մինիմումի կետ է M -ի վրա, ապա

$$(v^k, x^* - x^k) \leq f(x^*) - f(x^k) \leq 0 :$$

Հաշվի առնելով այս պայմանը և (5.17)-ը՝ կստանանք

$$\|x^m - x^*\|^2 \leq \|x^0 - x^*\|^2 + \sum_{i=0}^{m-1} \alpha_i^2 : \quad (5.18)$$

Քանի որ $\sum \alpha_k^2$ շարքը զուգամետ է, ապա (5.18) անհավասարությունից հետևում է, որ $\{x^k\}$ -ն սահմանափակ հաջորդականություն է: Այստեղից հետևում է, որ սահմանափակ կլինի նաև $\{v^k\}$ հաջորդականությունը: Այսինքն՝ գոյություն ունի $C > 0$ թիվ, այնպիսին, որ

$$\|v^k\| \leq C : \quad (5.19)$$

Ցույց տանք, որ գոյություն ունի ինդեքսների այնպիսի $\{k_s\}$ ենթահաջորդականություն, որ

$$(v^{k_s}, x^* - x^{k_s}) \rightarrow 0, \quad \text{երբ } k_s \rightarrow \infty : \quad (5.20)$$

Ենթադրենք հակառակը: Այդ դեպքում գոյություն ունի այնպիսի $N > 0$ թիվ, որ ինչ-որ K համարից սկսած տեղի ունի

$$(f'(x^k), x^* - x^k) < -N < 0, \quad \forall k > K \quad (5.21)$$

անհավասարությունը:

Օգտվելով (5.17)-(5.21) անհավասարություններից՝ կստանանք

$$\|x^{k+1} - x^*\|^2 \leq \|x^0 - x^*\|^2 - \frac{2N}{C} \sum_{i=0}^k \alpha_i + \sum_{i=0}^k \alpha_i^2 : \quad (5.22)$$

Այս անհավասարության մեջ անցնելով սահմանի, երբ $k \rightarrow \infty$ ՝ կստանանք, որ նրա աջ մասը ձգտում է $-\infty$, իսկ ձախ մասը ոչ բացասական թիվ է, ինչը հակասություն

Է: Այժմ $\{x_{k_s}\}$ ենթահաջորդականության կետերի համար կիրառելով ուռուցիկ ֆունկցիայի հիմնական անհավասարությունը, կունենանք՝

$$0 \geq f(x^*) - f(x^{k_s}) \geq (v^{k_s}, x^* - x^{k_s}):$$

Այստեղ անցնելով սահմանի, երբ $k_s \rightarrow \infty$ և հաշվի առնելով նաև (5.20)-ը՝ կստանանք

$$f(x^{k_s}) \rightarrow f(x^*) = \min_{x \in M} f(x):$$

Քանի որ $\{x^{k_s}\}$ -ը սահմանափակ հաջորդականություն է, ապա նրանից կարելի է անջատել գուգամետ ենթահաջորդականություն: Ընդհանրությունը չխախտելով՝ ենթադրենք որ $x^{k_s} \rightarrow \bar{x}$: Այստեղից, օգտվելով f -ի անընդհատությունից, կստանանք

$$f(x^{k_s}) \rightarrow f(\bar{x}) = f(x^*) = \min_{x \in M} f(x),$$

այսինքն՝ $\bar{x} \in M^*$: Ցույց տանք, որ $x^k \rightarrow \bar{x}$: Քանի որ $x^{k_s} \rightarrow \bar{x}$, ապա կամայական $\varepsilon > 0$ թվի համար գոյություն ունի այնպիսի K համար, որ երբ $k_s > K$, ապա տեղի ունի

$$\|x^{k_s} - \bar{x}\|^2 < \frac{\varepsilon}{2} \quad (5.23)$$

անհավասարությունը: Կարող ենք նաև համարել, որ ցանկացած p համարի համար տեղի ունի

$$\sum_{i=k_s}^{k_s+p-1} \alpha_i^2 < \frac{\varepsilon}{2} \quad (5.24)$$

անհավասարությունը: Օգտվելով (5.17) և (5.23)-(5.24) անհավասարություններից՝ ստանում ենք, որ երբ $k_s > K$, ապա ցանկացած p բնական թվի համար տեղի ունի հետևյալ անհավասարությունը՝

$$\|x^{k_s+p} - \bar{x}\|^2 \leq \|x^{k_s} - \bar{x}\|^2 + \sum_{i=k_s}^{k_s+p-1} \alpha_i^2 < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon:$$

Իսկ սա նշանակում է, որ

$$x^k \rightarrow \bar{x}:$$

ԽՆԴԻՐՆԵՐ

Ապրիորի մեթոդով լուծել հետևյալ խնդիրները:

Վերցնել $x^0 = (1, 1)$, $\alpha_k = 1/k + 1$, $k = 0, 1, 2, \dots$: Կատարել իտերացիայի երկու քայլ:

1. $3x_1^2 + x_2^2 + 11x_2 + 3x_1 \rightarrow \min:$
2. $-2x_1^2 - x_2^2 + 8x_1 + 6x_2 - 25 \rightarrow \max:$

5.6 Գրադիենտի պրոյեկտման մեթոդը

Դիցուք $M \subseteq \mathbb{R}^n$ -ը փակ ուռուցիկ բազմություն է: Հիշեցնենք, որ $\Pi_M(a)$ -ով նշանակվում է a վեկտորի **պրոյեկցիան** M բազմության վրա: Այսինքն՝ $\Pi_M(a)$ -ն M բազմության ամենամոտիկ կետն է a վեկտորից:

Ճշմարիտ է հետևյալ պնդումը (տե՛ս օրինակ՝ [22], լեմմա 2.1, էջ 225):

Լեմմա 5.3. Π_M օպերատորը բավարարում է L Լիպշիցի պայմանը $L = 1$ հաստատունով, այսինքն՝

$$\|\Pi_M(x) - \Pi_M(y)\| \leq \|x - y\| \quad \forall x, y \in \mathbb{R}^n : \quad (5.25)$$

Դիտարկենք մաթեմատիկական ծրագրավորման հետևյալ խնդիրը՝

$$f(x) \rightarrow \min, \quad x \in M:$$

Այստեղ $f(x)$ -ը դիֆերենցելի ֆունկցիա է, իսկ M -ը ուռուցիկ բազմություն է: Գրադիենտի պրոյեկտման մեթոդով այս խնդրի լուծման համար կառուցվում է $\{x_k\}$ հաջորդականություն հետևյալ ռեկուրենտ առնչությամբ՝

$$x^0 \in M, \quad x^{k+1} = \Pi_M(x^k - \alpha_k f'(x^k)), \quad k = 0, 1, 2, \dots:$$

Ուսումնասիրենք այն պայմանները, որոնց դեպքում այս հաջորդականությունը կգումգամիտի f -ի որևէ մինիմումի կետի:

Լեմմա 5.4. Դիցուք f -ը θ հաստատունով ուժեղ ուռուցիկ ֆունկցիա է M ուռուցիկ բազմության վրա: Այդ դեպքում տեղի ունի հետևյալ անհավասարությունը.

$$(f'(x) - f'(y), x - y) \geq 2\theta \|x - y\|^2 \quad \forall x, y \in M : \quad (5.26)$$

Ապացույց: Ըստ f ֆունկցիայի ուժեղ ուռուցիկության սահմանման կամայական $x, y \in M$ կետերի համար ունենք՝

$$f(x) - f(y) \geq (f'(x), x - y) + \theta \|x - y\|^2,$$

$$f(y) - f(x) \geq (f'(y), y - x) + \theta \|x - y\|^2:$$

Գումարելով այս երկու անհավասարությունները՝ կստանանք (5.26) անհավասարությունը:

Հենվելով այս արդյունքների վրա՝ ապացուցենք պրոյեկտման մեթոդի գումգամիտության հետևյալ թեորեմը:

Թեորեմ 5.6. Դիցուք f -ը θ հաստատունով ուժեղ ուռուցիկ ֆունկցիա է M փակ ուռուցիկ բազմության վրա: Ենթադրենք նաև, որ f' գրադիենտը բավարարում է L Լիպշիցի պայմանը M բազմության վրա $L > 0$ հաստատունով, այսինքն՝

$$\|f'(x) - f'(y)\| \leq L \|x - y\| \quad \forall x, y \in M:$$

Այդ դեպքում, եթե պրոյեկման մեթոդում որպես α_k քայլեր վերցնենք միևնույն հաստատունը $(0, \frac{4\theta}{L^2})$ միջակայքից, ապա $\{x^k\}$ հաջորդականությունը կգումարի f -ի միակ մինիմումի կետին երկրաչափական պրոգրեսիայի արագությամբ:

Ապացույց: Դիտարկենք հետևյալ օպերատորը $\mathbf{A}_\alpha : M \rightarrow M$,

$$\mathbf{A}_\alpha(x) \equiv \Pi_M(x - \alpha f'(x)):$$

Ցույց տանք, որ այն սեղմող օպերատոր է: Օգտվելով (5.25) – (5.26) անհավասարություններից՝ կունենանք

$$\begin{aligned} \|\mathbf{A}_\alpha(x) - \mathbf{A}_\alpha(y)\|^2 &= \|\Pi_M(x - \alpha f'(x)) - \Pi_M(y - \alpha f'(y))\|^2 \leq \\ &\leq \|x - \alpha f'(x) - y + \alpha f'(y)\|^2 \leq \|x - y + \alpha(f'(y) - f'(x))\|^2 = \\ &= \|x - y\|^2 + 2\alpha(x - y, f'(y) - f'(x)) + \alpha^2\|f'(x) - f'(y)\|^2 = \\ &\leq \|x - y\|^2(1 - 4\theta\alpha + \alpha^2 L^2) : \end{aligned} \quad (5.27)$$

Ընտրենք α այպես, որ

$$q \equiv \sqrt{1 - 4\theta\alpha + \alpha^2 L^2}$$

թիվը փոքր լինի մեկից: Եթե $\alpha \in (0, \frac{4\theta}{L^2})$, ապա $q < 1$ և հետևաբար $\mathbf{A}_\alpha$ օպերատորը կլինի սեղմող: Դիցուք x^* -ը նրա անշարժ կետն է M բազմության վրա, այսինքն՝

$$\Pi_M(x^* - \alpha f'(x^*)) = x^* : \quad (5.28)$$

Ցույց տանք, որ այդ անշարժ կետը f -ի մինիմումի կետն է: Իրոք, քանի որ պրոյեկտման օպերատորը բավարարում է

$$(\Pi_M(a) - a, x - \Pi_M(a)) \geq 0 \quad \forall a \in \mathbb{R}^n, \quad \forall x \in M \quad (5.29)$$

անհավասարությունը, ապա այստեղ տեղադրելով $a = x^* - \alpha f'(x^*)$, $\Pi_M(a) = x^*$ և հաշվի առնելով (5.28)-ը՝ կստանանք

$$\begin{aligned} (x^* - x^* + \alpha f'(x^*), x - x^*) &\geq 0 \Rightarrow \\ \Rightarrow (f'(x^*), x - x^*) &\geq 0 \quad \forall x \in M : \end{aligned} \quad (5.30)$$

Վերջապես, հաշվի առնելով նաև ուռուցիկ ֆունկցիայի հիմնական անհավասարությունը, (5.30)-ից կունենանք

$$f(x) - f(x^*) \geq (f'(x^*), x - x^*) \geq 0 \quad \forall x \in M,$$

այսինքն՝ x^* -ը f -ի մինիմումի կետն է:

Բացի դրանից՝ (5.27)-ից հետևում է, որ

$$\|x^{k+1} - x^*\| = \|\mathbf{A}_\alpha x^k - \mathbf{A}_\alpha x^*\| \leq q \|x^k - x^*\| \leq \dots$$

$$\leq q^{k+1} \|x^0 - x^*\|:$$

Իսկ սա նշանակում է, որ $\{x^k\}$ հաջորդականությունը զուգամիտում է x^* կետին երկրաչափական պրոգրեսիայի արագությամբ:

Թեորեմ 5.7. Դիցուք $f(x)$ - ուռուցիկ դիֆերենցելի ֆունկցիա է M ուռուցիկ կոմպակտ բազմության վրա: Կառուցենք x^k հաջորդականությունը պրոյեկտման մեթոդով՝ յուրաքանչյուր k -րդ իտերացիայում α_k քայլը ընտրելով հետևյալ բանաձևով՝

$$\alpha_k = \frac{D}{G\sqrt{k}}:$$

Այդ դեպքում տեղի ունի հետևյալ անհավասարությունը՝

$$f(\hat{x}_N) - \min_{x \in M} f(x) \leq \frac{3}{2} DG \frac{1}{\sqrt{N}},$$

որտեղ

$$\bar{x}_N = \frac{1}{N}(x^1 + x^2 + \dots + x^N),$$

D -ն M բազմության տրամագիծն է, իսկ G հաստատունը այնպիսին է, որ

$$\|f'(x)\| \leq G \quad \forall x \in M:$$

Ապացույց: Դիցուք x^* -ը մինիմումի կետն է: Այդ դեպքում կունենանք

$$\begin{aligned} \|x^{k+1} - x^*\|^2 &\leq \|\Pi_M(x^k - \alpha_k f'(x^k)) - x^*\|^2 \leq \\ &\|x^k - \alpha_k f'(x^k) - x^*\|^2 = \\ &= \|x^k - x^*\|^2 + \alpha_k^2 \|f'(x^k)\|^2 - 2\alpha_k (f'(x^k), x^k - x^*): \end{aligned}$$

Այստեղից, հաշվի առնելով նաև, որ f -ը ուռուցիկ է՝ կստանանք

$$\begin{aligned} 2\left(\sum_{k=1}^N (f(x^k) - f(x^*))\right) &\leq 2\sum_{k=1}^N (f'(x^k), x^k - x^*) \leq \\ &\sum_{k=1}^N \frac{\|x^k - x^*\|^2 - \|x^{k+1} - x^*\|^2}{\alpha_k} + G^2 \sum_{k=1}^N \alpha_k \leq \\ &\leq \sum_{k=1}^N \|x_k - x^*\|^2 \left(\frac{1}{\alpha_k} - \frac{1}{\alpha_{k-1}}\right) + G^2 \sum_{k=1}^N \alpha_k \leq \\ &D^2 \frac{1}{\alpha_N} + G^2 \sum_{k=1}^N \alpha_k \leq 3DG\sqrt{N}: \end{aligned}$$

Հետևաբար

$$f(\bar{x}_N) - f(x^*) \leq \frac{1}{N} \sum_{k=1}^N (f(x^k) - f(x^*)) \leq \frac{3}{2} DG \frac{1}{\sqrt{N}}:$$

1. Դիցուք $a \in \mathbb{R}^n$ կամայական կետ է: Ապացուցել, որ եթե

ա) $M = \{x \in \mathbb{R}^n / \|x - x^0\| \leq r\}$,
 ապա

$$\Pi_M(a) = x^0 + \frac{a - x^0}{\|a - x^0\|} r;$$

բ) եթե $M = \{x \in \mathbb{R}^n / b_j \leq x_j \leq c_j, j \in [1 : n]\}$,
 ապա

$$(\Pi_M(a))_j = \begin{cases} b_j, & \text{եթե } a_j < b_j \\ a_j, & \text{եթե } b_j \leq a_j \leq c_j \\ c_j, & \text{եթե } a_j > c_j; \end{cases}$$

գ) եթե $M = \{x \in \mathbb{R}^n / x_j \geq 0, j \in [1 : n]\}$,
 ապա

$$\Pi_M(a) = (\max(0, a_1), \dots, \max(0, a_n)):$$

Առաջադրանք: Կազմել ծրագիր, որը գրադիենտի պրոյեկտման մեթոդով կիրառանա

$$f(x) = 1/2(\mathbf{A}x, x) + (b, x)$$

քառակուսային ֆունկցիայի մինիմիզացիան գնդի և գուգահեռանիստի վրա: Որպես պրոյեկտման օպերատորներ վերցնել վերևում նշված բանաձևերը: L հաստատունը վերցնել $\mathbf{A}$ մատրիցի նորմը, ուժեղ ուռուցիկության θ հաստատունը վերցնել մատրիցի մինիմալ սեփական արժեքը, իսկ

$$\alpha_k = \frac{2\theta}{L^2}:$$

Կանգառի քայլ համարել $\|x^{k+1} - x^k\| < \varepsilon_0$ պայմանը, որտեղ $\varepsilon_0 > 0$ նախապես տրված ճշտություն է:

5.7 Գծայնացման կամ պայմանական գրադիենտի մեթոդը

Դիտարկենք պայմանական օպտիմիզացիայի ընդհանուր խնդիրը.

$$f(x) \rightarrow \min, x \in M: \quad (5.31)$$

Ենթադրենք՝ f -ը դիֆերենցելի ուռուցիկ ֆունկցիա է, իսկ $M \subset \mathbb{R}^n$ -ը կոմպակտ ուռուցիկ բազմություն է:

Տանք գծայնացման մեթոդի համառոտ նկարագրությունը: Ընտրում ենք կամայական x^0 կետ M բազմությունից և կառուցում ենք $\{x_k\}$ հաջորդականությունը

$$x^{k+1} = x^k + \alpha_k h^k, \quad k = 0, 1, \dots$$

ռեկուրենտ առնչությամբ: Այստեղ h^k վեկտորը f -ի այնպիսի նվազման ուղղություն է $x^k \in M$ կետում, որ բավականաչափ փոքր α_k քայլերի դեպքում $x_{k+1} \in M$: h^k վեկտորի որոշման համար k -րդ քայլում լուծում ենք հետևյալ միջանկյալ խնդիրը.

$$(f'(x^k), x - x^k) \rightarrow \min, \quad x \in M:$$

Ենթադրենք $\bar{x}_k$ -ն այդ խնդրի որևէ լուծում է: Նշանակենք

$$\eta_k = (f'(x^k), \bar{x}^k - x^k), \quad h^k = \bar{x}^k - x^k:$$

h^k վեկտորը կոչվում է պայմանական հակագրադիենտ: Ակնհայտ է, որ $\eta_k \leq 0$: Իրոք,

$$\eta_k = \min_{x \in M} (f'(x^k), x - x^k) \leq (f'(x^k), x^k - x^k) \leq 0:$$

Եթե $\eta_k < 0$, ապա ընտրում ենք α_k քայլը կիսման մեթոդով: Վերցնում ենք $\alpha = 1$ և ստուգում

$$f(x^k + \alpha h^k) - f(x^k) \leq \alpha \eta_k \quad (5.32)$$

անհավասարությունը: Եթե այն տեղի ունի է, ապա համարում ենք $\alpha_k = 1$, հակառակ դեպքում α -ն կիսում ենք և նորից ստուգում նշված անհավասարությունը և այսպես շարունակ: Երբ առաջին անգամ տեղի ունենա (5.32) անհավասարությունը, ապա այդ α -ն համարվում է α_k -ի արժեք, և ցիկլը ավարտվում է: Այնուհետև կառուցում ենք x^{k+1} կետը ռեկուրենտ առնչությամբ.

$$x^{k+1} = x^k + \alpha_k h^k:$$

Թեորեմ 5.8. Դիցուք

ա) $M \subset \mathbb{R}^n$ -ը ուռուցիկ կոմպակտ է;

բ) $f(x)$ -ը դիֆերենցելի է և ուռուցիկ M -ի վրա;

գ) $f'(x)$ գրադիենտը M բազմության վրա բավարարում է L իպոչիցի պայմանը:

Այդ դեպքում՝

- 1) եթե որևէ k -րդ քայլում $\eta_k = 0$, ապա x^k -ն (5.31) խնդրի լուծումն է, և այգործիքմն ավարտվում է:

2) եթե $\eta_k < 0$, $k = 0, 1, 2, \dots$, ապա

$$f(x^k) \rightarrow \min_{x \in M} f(x):$$

Ապացույց: Դիցուք $\eta_k = 0$: Ըստ ուռուցիկ ֆունկցիայի հիմնական անհավասարության՝ ունենք

$$f(x) - f(x^k) \geq (f'(x^k), x - x^k) \geq \eta_k = 0 \quad \forall x \in M,$$

այսինքն՝ x^k -ն (5.31) խնդրի լուծումն է:

Յույց տանք, որ h^k ուղղությամբ բավականաչափ փոքր քայլ կատարելիս մնում ենք M բազմության մեջ: Իրոք՝ քանի որ M -ը ուռուցիկ է, ապա

$$x^k + \alpha h^k = x^k + \alpha(\bar{x}_k - x^k) = (1 - \alpha)x^k + \alpha\bar{x}_k \in M, \quad \forall \alpha \in [0, 1]:$$

Քանի որ f -ի գրադիենտը M կոմպակտի վրա բավարարում է Լիպշիցի պայմանը, և $x_k \in M$, ապա կարելի է ցույց տալ, որ վերջավոր քայլերից հետո (5.32) անհավասարությունը տեղի ունի, և հետևաբար α_k քայլը ընտրվում է: Միաժամանակ գոյություն ունի այնպիսի $\bar{\alpha} > 0$ թիվ, որ $\alpha_k > \bar{\alpha} > 0$, $k = 0, 1, \dots$:

Այժմ ապացուցենք, որ

$$f(x^k) \rightarrow \min_{x \in M} f(x):$$

Համաձայն (5.32) անհավասարության, ունենք

$$f(x^{k+1}) \leq f(x^k) + \varepsilon \alpha_k \eta_k: \quad (5.33)$$

Գումարելով (5.33) անհավասարությունները $k \in [0 : m - 1]$ ինդեքսների համար՝ կստանանք

$$\min_{x \in M} f(x) - f(x^0) \leq f(x^m) - f(x^0) \leq \varepsilon \sum_{k=0}^{m-1} \alpha_k \eta_k:$$

Այստեղից հետևում է, որ բացասական անդամներով $\sum \alpha_k \eta_k$ շարքը գուգամետ է: Հետևաբար նրա ընդհանուր անդամը ձգտում է զրոյի, այսինքն՝

$$\alpha_k \eta_k \rightarrow 0: \quad (5.34)$$

Քանի որ $\alpha_k > \bar{\alpha} > 0$, ապա (5.34)-ից հետևում է, որ $\eta_k \rightarrow 0$: Այստեղից, ընտրելով $x^{k_j} \rightarrow x^* \in M$ գուգամետ ենթահաջորդականությունը և օգտվելով ուռուցիկ ֆունկցիայի հիմնական անհավասարությունից, կստանանք

$$f(x) - f(x^{k_j}) \geq (f'(x^{k_j}), x - x^{k_j}) \geq \eta_{k_j} \quad \forall x \in M: \quad (5.35)$$

Քանի որ $\eta_{k_j} \rightarrow 0$ և $x^{k_j} \rightarrow x^*$, ապա (5.35) անհավասարությունում անցնելով սահմանի՝ կստանանք

$$f(x) - f(x^*) \geq 0 \quad \forall x \in M,$$

այսինքն՝ x^* -ը f -ի մինիմումի կետն է M բազմության վրա:

Այսպիսով ապացուցվեց, որ $\{f(x^k)\}$ հաջորդականության $\{f(x^{k_j})\}$ ենթահաջորդականությունը զուգամիտում է $\min_{x \in M} f(x)$: Մյուս կողմից, քանի որ $\{f(x^k)\}$ հաջորդականությունը մոնոտոն նվազող է, ապա այն նույնպես կզուգամիտի $\min_{x \in M} f(x)$ -ին:

Գրականության մեջ այս մեթոդը կոչվում է Ֆրանկ-Վուլֆի մեթոդ [29]: Այս մեթոդը մեկնաբանենք մի թվային օրինակի միջոցով: Դիցուք ունենք հետևյալ խնդիրը՝

$$f(x) \equiv f(x_1, x_2) = -x_1^2 + x_1x_2 - 2x_2^2 + 4x_1 + 6x_2 \rightarrow \max,$$

$$\begin{cases} x_1 + x_2 \leq 4, \\ x_1 + 2x_2 \geq 2, \\ x_1, x_2 \geq 0: \end{cases}$$

Դիցուք $x_0 = (3, 1)$: Նկարագրենք Ֆրանկ-Վուլֆի մեթոդը այս խնդրի լուծման համար:

- Գտնում ենք նպատակային f ֆունկցիայի և նրա գրադիենտի արժեքները սկզբնական x^0 կետում՝ $f(x^0) = 10$, $f'(x^0) = (-1, 5)$:
- Գրաֆիկական եղանակով լուծելով գծային ծրագրավորման

$$\begin{cases} -x_1 + 5x_2 \rightarrow \max, \\ x_1 + x_2 \leq 4, \\ x_1 + 2x_2 \geq 2, \\ x_1, x_2 \geq 0 \end{cases}$$

խնդիրը՝ ստանում ենք $\bar{x}^0 = (0, 4)$ (տե՛ս նկ. 5.1):

- Հաջորդ x^1 կետը գտնում ենք հետևյալ բանաձևով՝

$$x^1 = x^0 + \alpha_0(\bar{x}^0 - x^0) = (3 - 3\alpha_0, 1 + 3\alpha_0):$$

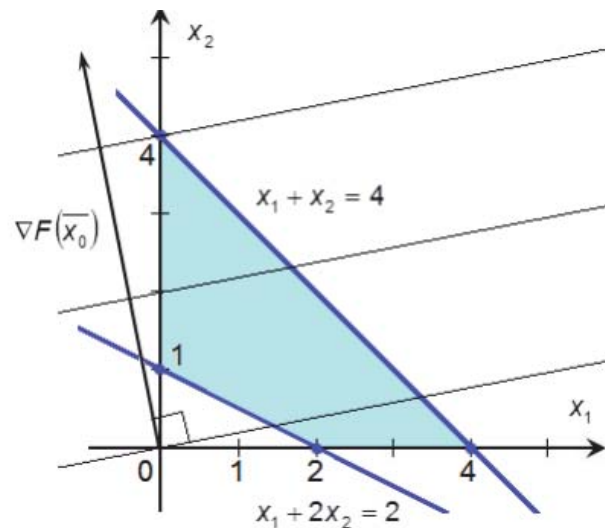
Գտնում ենք f ֆունկցիայի արժեքը այդ կետում՝

$$h(\alpha_0) \equiv f(x^1) = -36\alpha_0^2 + 18\alpha_0 + 10:$$

Դժվար չէ նկատել, որ $\alpha_0 = 0.25$ -ն h -ի մաքսիմումի կետն է: Ուստի

$$x^1 = (2.25, 1.75), \quad f(x^1) = 12.25, \quad f'(x^1) = (1.25, 1.25):$$

Քանի որ x^1 կետը գտնվում է $x_1 + x + 2 = 4$ ուղղի վրա, և $f'(x^1)$ գրադիենտն ուղղահայաց է այդ ուղղին, ապա այդ կետը օպտիմալ է:



Նկ. 5.1: Ֆրանկ-Վուլֆի մեթոդ

ԽՆԴԻՐՆԵՐ

1. Ապացուցել պնդումները հետևյալ խնդրի համար.

$$(c, x) \rightarrow \min, x \in M:$$

ա) Եթե $M = \{x \in \mathbb{R}^n / \|x - x^0\| \leq r\}$,

այսինքն

$$x^* = x^0 + \frac{c}{\|c\|} r$$

վեկտորը խնդրի լուծումն է:

բ) Եթե $M = \{x \in \mathbb{R}^n / a_j \leq x_j \leq b_j, j \in [1 : n]\}$,

այսինքն

$$x_j^* = \begin{cases} a_j, & \text{եթե } c_j \geq 0, \\ b_j, & \text{եթե } c_j < 0 \end{cases}$$

կոորդինատներով վեկտորը խնդրի լուծումն է:

2. Գտնել $f(x) = x_1^2 + x_2^2$ ֆունկցիայի պայմանական հակագրադիենտը $x = (2, 3)$ կետում $M = \{x \in \mathbb{R}^2 / x_1 + x_2 \leq 6, x_1 \geq 0, x_2 \geq 0\}$ բազմության վրա:

Առաջադրանք: Կազմել ծրագիր, որը իրականացնում է

$$f(x) = 1/2(\mathbf{A}x, x) + (b, x)$$

ֆունկցիայի մինիմիզացիան $M = \{x \in \mathbb{R}^n / \mathbf{C}x \leq d, x \geq 0\}$ բազմության վրա պայմանական գրադիենտի մեթոդով: Այստեղ $\mathbf{A}$ -ն $(n \times n)$ չափանի սիմետրիկ դրական որոշյալ մատրից է, իսկ $\mathbf{C}$ -ն $(m \times n)$ -չափանի մատրից է: k -րդ քայլում օգտագործելով սիմպլեքս ալգորիթը, ստանալ

$$\eta_k = \min_{x \in M} (f'(x^k), x - x^k)$$

խնդրի որևէ լուծում: Ալգորիթմի կանգառի համար ընդունել $|\eta_k| < \varepsilon_0$ պայմանը, որտեղ $\varepsilon_0 > 0$ նախապես տրված ճշտություն է: Եթե նշված պայմանը կատարվում է, ապա x^k վեկտորը համարել խնդրի լուծում և ավարտել ալգորիթմը: Սկզբնական $x^0 \in M$ կետի ընտրությունը նույնպես կատարել սիմպլեքս ալգորիթմով:

5.8 Ստոխաստիկ գրադիենտի մեթոդը հաստատուն քայլի դեպքում

Այս մեթոդի յուրաքանչյուր քայլում շարժման ուղղությունը ընտրվում է պատահականորեն՝ միայն այն պայմանով, որ այդ պատահական վեկտորի մաթեմատիկական սպասումը համընկնի ֆունկցիայի գրադիենտի հետ: Ստոխաստիկ գրադիենտի մեթոդը հետևյալն է:

- $w^1 = 0$:
- Դիցուք v_t -ն այնպիսի պատահական վեկտոր է, որ

$$E[v_t | w^t] = f'(w^t):$$

- $w^{t+1} = w^t - \eta v_t$:
- $\bar{w} = (1/T) \sum_{t=1}^T w^t$:

Տեղի ունի հետևյալ թեորեմը:

Թեորեմ 5.9. Դիցուք

- f -ը ուռուցիկ է, դիֆերենցելի և ρ -լիպշիցյան ֆունկցիա է:
- $w^* = \arg \min_{w: \|w\| \leq B} f(w)$:
- $\eta = \sqrt{B^2 / (\rho^2 T)}$:
- 1-հավանականությամբ տեղի ունեն

$$\|v_t\| \leq B, \quad \forall t$$

անհավասարությունները:

Այդ դեպքում

$$E[f(\bar{w})] - f(w^*) \leq \frac{B\rho}{\sqrt{T}}:$$

Ապացույց: Դիցուք $v_{1:t}$ -ն $v_1, v_2, \dots, v_t$ հաջորդականությունն է: Թեորեմ 5.1-ում ցույց տրվեցին հետևյալ անհավասարությունները՝

$$f(\bar{w}) - f(w^*) \leq \frac{1}{T} \sum_{t=1}^T (f(w^t) - f(w^*)),$$

$$\frac{1}{T} \sum_{t=1}^T (w^t - w^*, v_t) \leq \frac{B\rho}{\sqrt{T}} (w^t - w^*, v_t):$$

Վերցնելով այս անհավասարությունների երկու մասերի մաթեմատիկական սպասումները՝ կստանանք

$$E_{v_{1:T}}[f(\bar{w}) - f(w^*)] \leq E_{v_{1:T}} \left[\frac{1}{T} \sum_{t=1}^T (f(w^t) - f(w^*)) \right],$$

$$E_{v_{1:T}} \left[\frac{1}{T} \sum_{t=1}^T (w^t - w^*, v_t) \right] \leq E_{v_{1:T}} \left[\frac{B\rho}{\sqrt{T}} (w^t - w^*, v_t) \right]:$$

Մնաց ցույց տալ, որ

$$E_{v_{1:T}} \left[\frac{1}{T} \sum_{t=1}^T f(\bar{w}) - f(w^*) \right] \leq E_{v_{1:T}} \left[\frac{1}{T} \sum_{t=1}^T (w^t - w^*, v_t) \right]: \quad (5.36)$$

Մաթեմատիկական սպասման գծային հատկության համաձայն՝ ունենք

$$E_{v_{1:T}} \left[\frac{1}{T} \sum_{t=1}^T (w^t - w^*, v_t) \right] = \frac{1}{T} \sum_{t=1}^T E_{v_{1:T}}[(w^t - w^*, v_t)]:$$

Այժմ հիշենք մաթեմատիկական սպասման հետևյալ հատկությունը՝ α և β պատահական մեծությունների և g ֆունկցիայի համար տեղի ունի

$$E_\alpha[g(\alpha)] = E_\beta E_\alpha[g(\alpha)|\beta]$$

հավասարությունը: Դիցուք $\alpha = v_{1:t}$, $\beta = v_{1:t-1}$: Այդ դեպքում կունենանք

$$\begin{aligned} E_{v_{1:T}}[(w^t - w^*, v_t)] &= E_{v_{1:t}}[(w^t - w^*, v_t)] = \\ &= E_{v_{1:t-1}} E_{v_{1:t}}[w^t - w^*, v_t | v_{1:t-1}]: \end{aligned}$$

Եթե $v_{1:t-1}$ մեծությունը հայտնի է, ապա w^t վեկտորը արդեն պատահական մեծություն չէ, և հետևաբար՝

$$E_{v_{1:t-1}} E_{v_{1:t}}[w^t - w^*, v_t | v_{1:t-1}] = E_{v_{1:t-1}}(w^t - w^*, E_{v_t}[v_t | v_{1:t-1}]):$$

Քանի որ w^t կախված է միայն $v_{1:t-1}$ հաջորդականությունից, և $E_{v_t}[v_t | w^t] = f'(w^t)$, ապա

$$E_{v_t}[v_t | v_{1:t-1}] = f'(w^t):$$

Այսպիսով ցույց տվեցինք, որ

$$E_{v_{1:T}}[(w^t - w^*, v_t)] \geq E_{v_{1:t-1}}[f(w^t) - f(w^*)] = E_{v_{1:T}}[f(w^t) - f(w^*)]:$$

Գումարելով այս անհավասարությունը և այնուհետև բաժանելով T -ի վրա, հաշվի

առնելով նաև մաթեմատիկական սպասման գծային հատկությունը, կստանանք (5.36) անհավասարությունը: Թեորեմն ապացուցվեց:

5.9 AdaGrad (adaptive gradient) ալգորիթմը

Այս մեթոդը նման է ստոխաստիկ գրադիենտային մեթոդին, բայց միայն այն տարբերությամբ, որ քայլը իտերացիայից իտերացիա փոխվում է: *AdaGrad* ալգորիթմը կառուցում է x_t հաջորդականությունը հետևյալ ռեկուրենտ առնչությամբ՝

$$y_{t+1} = x_t - \eta G_t^{-1} f'(x_t), \quad x_{t+1} = \arg \min_{x \in M} \|y_{t+1} - x\|_{G_t}^2,$$

որտեղ

$$G_t = \left(\sum_{\tau=1}^t f'(x_\tau)(f'(x_\tau))^T \right)^{\frac{1}{2}}:$$

Նշանակենք

$$\|x\|_A^2 = x^T A x, \quad D = \max_{u, v \in M} \|u - v\|:$$

Տեղի ունի հետևյալ գնահատականը:

Թեորեմ 5.10. *Դիցուք x^* -ը f ուռուցիկ ֆունկցիայի մինիմումի կետն է M ուռուցիկ կոմպակտ բազմության վրա: Այդ դեպքում $\eta = D$ արժեքի դեպքում *AdaGrad*-ի T -րդ իտերացիայի համար ճշմարիտ է հետևյալ անհավասարությունը՝*

$$f\left(\frac{1}{T} \sum_{t=1}^T x_t\right) - f(x^*) \leq \frac{2D}{T} \text{Tr}(G_T),$$

որտեղ

$$\text{Tr}(G_T) \equiv \sum_i g_{ii}:$$

Ապացույց: Ունենք

$$y_{t+1} - x^* = x_t - x^* - \eta G_t^{-1} f'(x_t):$$

Հետևաբար՝

$$G_t(y_{t+1} - x^*) = G_t(x_t - x^*) - \eta f'(x_t):$$

Այս երկու հավասարություններից ստանում ենք՝

$$\begin{aligned} (y_{t+1} - x^*)^T G_t (y_{t+1} - x^*) &= \\ &= (x_t - x^*)^T G_t (x_t - x^*) - 2\eta (f'(x_t), x_t - x^*) + \eta^2 (f'(x_t))^T G_t^{-1} f'(x_t): \end{aligned}$$

Մյուս կողմից, քանի որ x_{t+1} -ը y_{t+1} կետի պրոյեկցիան է ուռուցիկ կոմպակտի վրա

ըստ G_t նորմի, ապա

$$(y_{t+1} - x^*)^T G_t (y_{t+1} - x^*) = \|y_{t+1} - x^*\|^2 \geq \|x^{t+1} - x^*\|^2:$$

Ուստի

$$(f'(x_t), x_t - x^*) \leq \frac{\eta}{2} (f'(x_t))^T G_t^{-1} f'(x_t) + \frac{1}{2\eta} (\|x_t - x^*\|_{G_t}^2 - \|x_{t+1} - x^*\|_{G_t}^2):$$

Փոխարելով այս անհավասարությունները $t = 1, 2, \dots, T$ արժեքների համար՝ կստանանք

$$\begin{aligned} \sum_{t=1}^T (f'(x_t), x_t - x^*) &\leq \\ \frac{\eta}{2} \sum_{t=1}^T (f'(x_t))^T G_t^{-1} f'(x_t) + \frac{1}{2\eta} \|x_1 - x^*\|_{G_0}^2 + \\ + \frac{1}{2\eta} \sum_{t=1}^T (\|x_t - x^*\|_{G_t}^2 - \|x_{t+1} - x^*\|_{G_{t-1}}^2) - \frac{1}{2\eta} \|x_{T+1} - x^*\|_{G_T}^2 &\leq \\ \frac{\eta}{2} \sum_{t=1}^T (f'(x_t))^T G_t f'(x_t) + \frac{1}{2\eta} \sum_{t=1}^T (x_t - x^*)^T (G_t - G_{t-1}) (x_t - x^*) &: \end{aligned}$$

Համաձայն լեմաներ 5.12. 5.1[15]-ի՝ տեղի ունեն հետևյալ գնահատականները՝

$$\begin{aligned} \sum_{t=1}^T (f'(x_t))^T G_t^{-1} f'(x_t) &\leq 2Tr(G_T), \\ \sum_{t=1}^T (x_t - x^*)^T (G_t - G_{t-1}) (x_t - x^*) &\leq D^2 Tr(G_T) : \end{aligned}$$

Նկատի ունենալով այս անհավասարությունները և f ֆունկցիայի ուռուցիկությանը հիմնական $f(x_t) - f(x^*) \leq (f'(x_t), x_t - x^*)$ անհավասարությունը՝ կստանանք՝

$$\frac{1}{T} \sum_{t=1}^T (f(x_t) - f(x^*)) \leq \frac{D^2 Tr(G_T)}{2\eta T} + \frac{\eta Tr(G_T)}{T}:$$

Այստեղից, հաշվի առնելով Յենսենի

$$f\left(\frac{1}{T} \sum_{t=1}^T x_t\right) \leq \frac{1}{T} \sum_{t=1}^T f(x_t)$$

անհավասարությունը, $\eta = D$ արժեքի համար, կունենանք

$$f\left(\frac{1}{T} \sum_{t=1}^T x_t\right) - f(x^*) \leq \frac{2D}{T} Tr(G_T):$$

Թերեմն ապացուցվեց:

Երբ f ֆունկցիան ուժեղ ուռուցիկ է, ապա *AdaGrad* ալգորիթով կարելի գտնել ֆունկցիայի մինիմումի կետը ամբողջ տարածության վրա: Այդ դեպքում կոորդինատներով այս ալգորիթմը կունենա հետևյալ տեսքը՝

$$x_{t+1}^i = x_t^i - \frac{\eta}{\sqrt{g_{ii}}} \frac{\partial f}{\partial x_i}, \quad g_{ii} = \sum_{\tau=1}^t \left(\frac{\partial f}{\partial x_i}(x_\tau) \right)^2, \quad i = 1, 2, \dots, n:$$

Դիտարկենք թվային մի օրինակ և կատարենք *AdaGrad*-ի մեկ խտրացիա:

Դիցուք

$$f(x_1, x_2) = ((x_1 + x_2 a) - b)^2, \quad a = 0.39, \quad b = 9.84:$$

Որպես սկզբնական մոտավորություն վերցնենք $(x_1, x_2) = (0, 0)$ կետը, $\eta = 5$: Այդ դեպքում $f'(0, 0) = (-19.68, -7.68)$: Հետևաբար՝

$$G_1 = \sum_{\tau=1}^1 f'(0)(f'(0))^T = \begin{pmatrix} -19.68 \\ -7.68 \end{pmatrix} (-19.68, -7.68) = \begin{pmatrix} 387.30 & 151.14 \\ 151.14 & 58.98 \end{pmatrix}:$$

Ուստի հաջորդ խտրացիայում ալգորիթը կգնարացնի հետևյալ կետը՝

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} - \begin{pmatrix} \frac{5}{\sqrt{387.30}} \\ \frac{5}{\sqrt{58.98}} \end{pmatrix} \odot \begin{pmatrix} -19.68 \\ -7.68 \end{pmatrix} = \begin{pmatrix} 5 \\ 5 \end{pmatrix},$$

որտեղ $\odot$ նշանը Ադամարի իմաստով արտադրյալ է: Այն $A(m \times n)$ և $B(m \times n)$ մատրիցների միջև սահմանվում է հետևյալ կերպ՝

$$(A \odot B)_{ij} = (A)_{ij}(B)_{ij}:$$

5.10 Գրադիենտային արագացված մեթոդներ (Gradient decent with momentum and Adam method)

Այս պարագրաֆում քնարկվող մեթոդը առաջարկվել է Բ. Պոլյակի [40] և Յու. Նեստերովի կողմից [15], որոնք սովորական գրադիենտային մեթոդի մոդիֆիկացիաներ են, սակայն ապահովում են գույքամիտության ավելի բարձր արագություն:

Նախ հիշենք որոշակի գնահատականներ գրադիենտային եղանակի գույքամիտության արագության վերաբերյալ: Ճշմարիտ է հետևյալ պնդումը:

Թեորեմ 5.11. Դիցուք f -ը β -ոդորկ, և ուռուցիկ ֆունկցիա է, և x^* -ը նրա մինիմումի կետն է $\mathbb{R}^n$ -ի վրա: Դիցուք կատարվում է գրադիենտային իջեցում հետևյալ բանաձևով՝

$$x_{t+1} = x_t - \frac{1}{\beta} f'(x_t):$$

Այդ դեպքում տեղի ունի հետևյալ անհավասարությունը՝

$$f(x_t) - f(x^*) \leq \frac{2\beta \|x_0 - x^*\|^2}{t + 4}:$$

Սա նշանակում է, որ գուգամիտության արագությունը $O(1/t)$ կարգի է:

Ավելի ուժեղ գնահատական տեղի ունի, եթե ֆունկցիայից պահանջվի նաև ուժեղ ուռուցիկություն:

Թեորեմ 5.12. Դիցուք f -ը β -ողորկ և μ -ուժեղ ուռուցիկ ֆունկցիա է, և x_t կառուցվում է գրադիենտային իջեցման եղանակով՝ $\alpha_t = 1/\beta$ հաստատուն քայլով: Ենթադրենք նաև $0 < \mu < \beta$:

Այդ դեպքում տեղի ունի հետևյալ անհավասարությունը՝

$$f(x_t) - f(x^*) \leq \frac{\beta}{2} \left(1 - \frac{\mu}{\beta}\right)^t \|x_0 - x^*\|^2:$$

Այս գնահատականից հետևում է, որ որքան փոքր է μ/β հարաբերությունը, այնքան գուգամիտության արագությունը փոքր է:

Դիտարկենք քառակուսային ֆունկցիայի դեպքը՝

$$f(x_1, x_2) = \frac{1}{2}(\lambda_1 x_1^2 + \lambda_2 x_2^2),$$

որտեղ $0 < \lambda_1 < \lambda_2$: Այս դեպքում $\lambda_1 = \mu$, $\lambda_2 = \beta$: Կիրառենք գրադիենտային իջեցում՝

$$(x_{t+1,1}, x_{t+1,2}) = x_t - \frac{1}{\lambda_2} f'(x_t) = \left((1 - \frac{\lambda_1}{\lambda_2})x_{t,1}, 0\right):$$

Այսպիսով, եթե λ_1/λ_2 հարաբերությունը շատ փոքր է, ապա, համաձայն վերը նշված գնահատականի, $x_{t,1}$ -ն դանդաղ կգուգամիտի զրոյի:

Այժմ դիտարկենք գրադիենտային իջեցման մեթոդի մի մոդիֆիկացիա, որը կոչվում է **Պոյակի մոմենտների մեթոդ (Momentum method)**:

Այս դեպքում x_t հաջորդականությունը կառուցվում է հետևյալ ռեկուրենտ առնչություններով՝

$$m_{t+1} = \gamma_t m_t + (1 - \gamma_t) f'(x_t),$$

$$x_{t+1} = x_t - \alpha_t m_{t+1}:$$

Այս բանաձևերից կարելի է գրակացնել, որ x_t կետից շարժումը կատարվում է ոչ թե $-f'(x_t)$ հակագրադիենտի ուղղությամբ, այլ $-m_{t+1}$ վեկտորի ուղղությամբ, որը $-f'(x_0), -f'(x_1), \dots, -f'(x_t)$ հակագրադիենտների միջինն է:

Ցույց տանք, որ x_t վեկտորը կարելի է ներկայացնել հետևյալ տեսքով՝

$$x_{t+1} = x_t - \tilde{\alpha}_t f'(x_t) + \tilde{\gamma}_t (x_t - x_{t-1}),$$

որտեղ

$$\tilde{\alpha}_t = \alpha_t (1 - \gamma_t), \quad \tilde{\gamma}_t = \frac{\alpha_t \gamma_t}{\alpha_{t-1}}:$$

Իրոք, իտերացիոն երկրորդ բանաձևից ունենք

$$m_{t+1} = \frac{x_t - x_{t+1}}{\alpha_t}, t = 1, 2, \dots :$$

Այստեղից՝

$$m_t = \frac{x_{t-1} - x_t}{\alpha_{t-1}}, t = 0, 1, 2, \dots :$$

Տեղադրելով m_t -ի այս արժեքը իտերացիոն առաջին բանաձևում՝ կստանանք

$$\frac{x_t - x_{t+1}}{\alpha_t} = \gamma_t \left(\frac{x_{t-1} - x_t}{\alpha_{t-1}} \right) + (1 - \gamma_t) f'(x_t) :$$

Հետևաբար՝

$$x_{t+1} = x_t - \alpha_t (1 - \gamma_t) f'(x_t) + \frac{\alpha_t \gamma_t}{\alpha_{t-1}} (x_t - x_{t-1}) :$$

Դիցուք ունենք քառակուսային

$$f(x) = \frac{1}{2} (Ax, x) + (b, x)$$

Ֆունկցիան, որտեղ A -ն սիմետրիկ, դրական որոշյալ մատրից է, իսկ b -ն վեկտոր է։ Կառուցենք $\{x_t\}$ հաջորդականությունը *Momentum*-ի հետևյալ բանաձևերով՝

$$m_{t+1} = \gamma m_t + (1 - \gamma) f'(x_t),$$

$$x_{t+1} = x_t - \alpha m_{t+1}, m_0 = f'(x_0),$$

$$\alpha = \frac{1}{\sqrt{\lambda_{\min} \lambda_{\max}}}, \gamma = \left(\frac{\sqrt{\lambda_{\max}} - \sqrt{\lambda_{\min}}}{\sqrt{\lambda_{\max}} + \sqrt{\lambda_{\min}}} \right)^2,$$

որտեղ $\lambda_{\min}, \lambda_{\max}$ թվերը համապատասխանաբար A մատրիցի մինիմալ և մաքսիմալ սեփական արժեքներն են։

Տեղի ունի գուգամիտության հետևյալ գնահատականը։

Թեորեմ 5.13. Գոյությունի ունի $C > 0$ հաստատուն, այնպիսին, որ

$$f(x_t) - f(x^*) \leq Ct^2 \gamma^t \|x_0 - x^*\|^2 :$$

Դիցուք $\lambda_{\min} < \lambda_{\max}$ ։ Այդ դեպքում

$$\left(\frac{\sqrt{\lambda_{\max}} - \sqrt{\lambda_{\min}}}{\sqrt{\lambda_{\max}} + \sqrt{\lambda_{\min}}} \right)^2 \approx 1 - 4 \sqrt{\frac{\lambda_{\min}}{\lambda_{\max}}} < 1 - \frac{\lambda_{\min}}{\lambda_{\max}} :$$

Հետևաբար *Momentum*-ը մեծ արագություն ունի, քան ստանդարտ գրադիենտային իջեցումը։ Օրինակ՝ եթե $\lambda_{\min}/\lambda_{\max} = 0.01$, ապա

$$1 - \frac{\lambda_{\min}}{\lambda_{\max}} = 0.99,$$

$$1 - 4\sqrt{\frac{\lambda_{min}}{\lambda_{max}}} = 0.6 :$$

Ուստի ստանդարտ դեպքում զուգամիտությունը դանդաղ է, իսկ *Momentum*-ի դեպքում՝ արագ:

Այժմ քննարկենք *Momentum* ալգորիթմի ևս մի եղանակ, որը կոչվում է **Նեստերովի *Momentum* ալգորիթմ**:

Դիցուք f -ը ուռուցիկ և β -ողորկ ֆունկցիա է: Դիցուք $\{x_t\}$, $\{y_t\}$ հաջորդականությունները կառուցվում են հետևյալ ռեկուրենտ առնչություններով՝

$$y_{t+1} = x_t - \frac{1}{\beta} f'(x_t),$$

$$x_{t+1} = (1 - \gamma_t) y_{t+1} + \gamma_t y_t,$$

որտեղ

$$\gamma_t = \frac{1 - \lambda_t}{\lambda_{t+1}},$$

$$\lambda_0 = 0, \lambda_t = \frac{1 + \sqrt{1 + 4\lambda_{t-1}^2}}{2} :$$

Տեղի ունի հետևյալ գնահատականը:

Թեորեմ 5.14. Դիցուք x^* -ը f -ի մինիմումի կետն է $\mathbb{R}^n$ -ի վրա: Այդ դեպքում

$$f(y_t) - f(x^*) \leq \frac{2\beta \|x_1 - x^*\|^2}{(t-1)^2} :$$

Ապացույց: Հաշվի առնելով f ֆունկցիայի β -ողորկությունը և ուռուցիկությունը, ցանկացած x , y կետերի համար տեղի ունի հետևյալ անհավասարությունը՝

$$\begin{aligned} f(x - \frac{1}{\beta} f'(x)) - f(y) &\leq \\ f(x - \frac{1}{\beta} f'(x)) - f(x) + (f'(x), x - y) &\leq \\ (f'(x), x - \frac{1}{\beta} f'(x) - x) + \frac{\beta}{2} \|x - \frac{1}{\beta} f'(x) - x\|^2 + (f'(x), x - y) &= \\ -\frac{1}{2\beta} \|f'(x)\|^2 + (f'(x), x - y) : \end{aligned}$$

Տեղադրելով այս անհավասարության մեջ $x = x_t$, $y = y_t$ ՝ կստանանք

$$\begin{aligned} f(y_{t+1}) - f(y_t) &\leq -\frac{1}{2\beta} \|f'(x_t)\|^2 + (f'(x_t), x_t - y_t) = \\ -\frac{\beta}{2} \|y_{t+1} - x_t\|^2 - \beta(y_{t+1} - x_t, x_t - y_t) : \end{aligned} \tag{5.37}$$

Նմանապես տեղադրելով $x = x_t$, $y = x^*$ ՝ կստանանք

$$f(y_{t+1}) - f(x^*) \leq -\frac{\beta}{2} \|y_{t+1} - x_t\|^2 - \beta(y_{t+1} - x_t, x_t - x^*) : \quad (5.38)$$

Նշանակենք $\delta_t = f(y_t) - f(x^*)$: Այժմ բազմապատկելով (5.37) անհավասարությունը $(\lambda_t - 1)$ -ով և գումարելով (5.38)-ի հետ՝ կստանանք

$$\begin{aligned} & \lambda_t \delta_{t+1} - (\lambda_t - 1) \delta_t \leq \\ & \leq -\frac{\beta}{2} \lambda_t \|y_{t+1} - x_t\|^2 - \beta(y_{t+1} - x_t, \lambda_t x_t - (\lambda_t - 1)y_t - x^*) : \end{aligned}$$

Բազմապատկենք այս անհավասարությունը λ_t -ով և հաշվի առնենք, որ $\lambda_{t-1}^2 = \lambda_t^2 - \lambda_t$, կստանանք

$$\begin{aligned} & \lambda_t^2 \delta_{t+1} - \lambda_{t-1}^2 \delta_t \leq \\ & -\frac{\beta}{2} \left(\|\lambda_t(y_{t+1} - x_t)\|^2 + 2\lambda_t(y_{t+1} - x_t, \lambda_t x_t - (\lambda_t - 1)y_t - x^*) \right) : \end{aligned} \quad (5.39)$$

Հեշտ է նկատել, որ

$$\begin{aligned} & \left(\|\lambda_t(y_{t+1} - x_t)\|^2 + 2\lambda_t(y_{t+1} - x_t, \lambda_t x_t - (\lambda_t - 1)y_t - x^*) \right) = \\ & = \|\lambda_t y_{t+1} - (\lambda_t - 1)y_t - x^*\|^2 - \|\lambda_t x_t - (\lambda_t - 1)y_t - x^*\|^2 : \end{aligned} \quad (5.40)$$

Ըստ սահմանման՝ ունենք նաև

$$x_{t+1} = \lambda_t y_{t+1} - (\lambda_t - 1)y_t : \quad (5.41)$$

Նշանակելով $u_t = \lambda_t x_t - (\lambda_t - 1)y_t - x^*$ և հաշվի առնելով (5.39)-(5.41) անհավասարությունները՝ կստանանք

$$\lambda_t^2 \delta_{t+1} - \lambda_{t-1}^2 \delta_t \leq \frac{\beta}{2} (\|u_t\|^2 - \|u_{t+1}\|^2) :$$

Գումարելով այս անհավասարությունները $t = 1, 2, \dots, T$ արժեքների համար՝ կստանանք

$$\delta_T \leq \frac{\beta}{2\lambda_{T-1}^2} \|u_1\|^2 : \quad (5.42)$$

Հեշտ է տեսնել, որ $\lambda_T \geq T/2$: Իրոք, ունենք

$$\begin{aligned} \lambda_t &= \frac{1 + \sqrt{1 + 4\lambda_{t-1}^2}}{2} \geq \\ & \frac{1}{2} + \frac{1}{2} 4\lambda_{t-1}^2 = \frac{1}{2} + \lambda_{t-1} : \end{aligned}$$

Այստեղից, քանի որ $\lambda_0 = 0$, ապա $\lambda_T \geq T/2$:

Այստեղից և (5.42) անհավասարությունից անմիջապես բխում է պահանջվող գնահատականը: Թեորեմն ապացուցվեց:

Այժմ քննարկենք Պոլյակի մոմենտների մեթոդի մի մեկնարկություն:

Ոչ պայմանական օպտիմիզացիայի $f(x) \rightarrow \min$ խնդրի գրադիենտային եղանակը կարող ենք մեկնաբանել դիֆերենցիալ հավասարումների տերմիններով: Դիտարկենք դիֆերենցիալ $p \frac{dx}{dt} + f'(x) = 0$, ($p > 0$) հավասարումը: Այս հավասարման լուծման պարզագույն տարբերական սխեման Էյլերի հետևյալ բացահայտ սխեման է՝

$$p \frac{x_{n+1} - x_n}{h} + f'(x_n) = 0:$$

Այստեղից

$$x_{n+1} = x_n - \frac{h}{p} f'(x_n),$$

որը օպտիմիզացիոն խնդրի գրադիենտային իջեցման եղանակն է:

Այժմ դիտարկենք երկրորդ կարգի հետևյալ դիֆերենցիալ հավասարումը՝

$$m \frac{d^2 x}{dt^2} + p \frac{dx}{dt} + f'(x) = 0:$$

որը նկարագրում է m զանգվածով գնդիկի շարժման հավասարումը f' պոտենցիալով դաշտում սահքի ուժի առկայությամբ: Էներգիայի կորստի պատճառով գնդիկը ժամանակի ընթացքում կհայտնվի մինիմում պոտենցիալով կետում: Այս դիֆերենցիալ հավասարման վերջավոր տարբերական սխեման հետևյալն է՝

$$m \frac{x_{n+2} - 2x_n + x_{n-1}}{h^2} + p \frac{x_n - x_{n-1}}{h} + f'(x_n) = 0:$$

Պարզ ձևափոխություններից հետո կստանանք

$$x_{n+1} = x_n - \alpha f'(x_n) + \gamma(x_n - x_{n-1})$$

տեսքի ռեկուրենտ առնչությունը, որը համընկնում է Պոյակի մոմենտների մեթոդի հետ: Վերջինս գրականության մեջ երբեմն կոչվում է ծանր գնդիկի մեթոդ:

Վերջում ներկայացնենք վերևում նկարագրված մոմենտների մեթոդի ժամանակակից տարբերակը, որը ներկայումս ներդրային ցանցերի ուսուցման հիմնական ալգորիթմներից է: Այն գրականության մեջ կոչվում է Ադամի ալգորիթմ (Adam Method): Այս մեթոդի գուգամիտության պայմանների հետ կարելի է ծանոթանալ [2] աշխատանքում: Ալգորիթմի կոդը (pseudocode) ներկայացված է: Այն բաղկացած է հետևյալ իտերացիաներից.

- η : **Learning rate**
- $\beta_1, \beta_2 \in (0, 1)$: Exponential decay rate the moment estimate
- $f(x)$: **Cost function with parameters** w
- x_0 : **Initial parameter vector**
- $m_0 := 0, v_0 = 0, t = 0$ (Initialize timestep)
- **while** w not converged **do**
 $t := t + 1$

```

 $g_t = f'(x_{t-1})$ 
 $m_t := \beta_1 m_{t-1} + (1 - \beta_1) g_t$ 
 $v_t := \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$ 
 $\hat{m}_t := m_t / (1 - \beta_1^t)$ 
 $\hat{v}_t := v_t / (1 - \beta_2^t)$ 
 $x_t := x_{t-1} - \eta \hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon)$ 
end while
return  $x_t$  (Resulting parameters)

```

Առաջադրանք: Կազմել ծրագիր (կոդ Python-ով), որը կիրականացնի $f(x) = \frac{1}{2}(Ax, x) + (b, x)$ ֆունկցիայի մինիմիզացիան *AdaGrad*, *Momentum* և *Adam* ալգորիթմներով ամբողջ տարածության վրա: Այստեղ $A(n \times n)$ -ը մատրիցը սիմետրիկ է և դրական որոշյալ, իսկ b -ն վեկտոր է:

5.11 Ստոխաստիկ գրադիենտի մեթոդի կիրառությունը ռեալ ռիսկի մինիմիզացիայի խնդրում

Մինչև այժմ մենք դիտարկեցինք ստոխաստիկ գրադիենտի մեթոդը ուռուցիկ ֆունկցիաների համար: Այժմ այն կիրառենք մեքենայական ուսուցման խնդիրներում: Մեքենայական ուսուցման հիմնական խնդիրն է մինիմիզացնել ռեալ ռիսկի ֆունկցիան՝

$$L_D(w) \equiv E_{z \sim D}[l(w, z)] \rightarrow \min, \|w\| \leq B:$$

Սակայն քանի որ D բաշխումը հայտնի չէ, ապա չենք կարող հաշվել $(L_D(w))'$ գրադիենտը: Բայց ստոխաստիկ գրադիենտի մեթոդում պետք գտնել պատահական վեկտոր, որի մաթեմատիկական սպասումը համընկնի $(L_D(w))'$ գրադիենտի հետ: Այժմ ցույց տանք, որ հեշտությամբ կարելի է կառուցել այդպիսի վեկտոր: Դիտարկենք այն դեպքը, երբ կորստի l ֆունկցիան դիֆերենցելի է: Դա նշանակում է, որ L_D ֆունկցիան նույնպես դիֆերենցելի է: v_t պատահական վեկտորը կառուցում ենք հետևյալ կերպ: Սկզբից ֆիքսում ենք պատահական վեկտոր $z \sim D$: Այնուհետև որպես v_t վեկտոր վերցնում ենք $l(w, z)$ ֆունկցիայի գրադիենտը ըստ w -ի w^t կետում՝ $v_t = (l(z, w^t))'$: Այդ դեպքում կունենանք

$$E[v_t | w^t] = E_{z \sim D}[(l(w^t, z))'] = (E_{z \sim D}[l(w^t, z)])' = (L_D(w^t))':$$

Հետևաբար՝ ստոխաստիկ գրադիենտի ալգորիթմը ռեալ ռիսկի մինիմիզացիայի խնդրում հետևյալն է.

- $w^1 = 0$:
- Ֆիքսում ենք կամայական $z \sim D$ -ից և նշանակում $v_t = (l(w^t, z))'$:

- $w^{t+1} = w^t - \eta v_t$, որտեղ η -ն հաստատուն է:
- $\bar{w} = (1/T) \sum_{t=1}^T w_t$:

Կիրառելով թեորեմ 5.8-ը՝ ստանում ենք հետևյալ պնդումը:

Թեորեմ 5.15. Դիցուք տեղի ունեն թեորեմ 5.8-ի ենթադրությունները: Այդ դեպքում ցանկացած ϵ -ի համար ստոխաստիկ գրադիենտի մեթոդով $L_D(w)$ -ի մինիմիզացիայի խնդրում՝ կատարելով $T \geq B^2 \rho^2 / \epsilon^2$ քայլ՝ ստանում ենք, որ

$$E[L_D(\bar{w})] \leq \min_{w \in H} L_D(w) + \epsilon:$$

5.12 Ստոխաստիկ գրադիենտի մեթոդի կիրառությունը ուռուցիկ և β -ոդորկ ֆունկցիաների մինիմիզացիայի խնդիրներում

Թեորեմ 5.16. Դիցուք բոլոր z պարամետրերի համար կորստի $l(., z)$ ֆունկցիան ուռուցիկ է, β -ոդորկ է և ոչ բացասական: Այդ դեպքում կիրառելով ստոխաստիկ գրադիենտային մեթոդը $L_D(w)$ -ի մինիմիզացիայի խնդրում՝ ստանում ենք հետևյալ գնահատականը՝

$$E[L_D(\bar{w})] \leq \frac{1}{1 - \eta\beta} \left(L_D(w^*) + \frac{\|w^*\|^2}{2\eta T} \right),$$

որտեղ w^* կամայական ֆիքսված վեկտոր է:

Ապացույց: Հիշենք, որ β -ոդորկ ֆունկցիաների համար տեղի ունի հետևյալ անհավասարությունը՝

$$\|f'(w)\|^2 \leq 2\beta f(w):$$

Դիցուք $z_1, z_2, \dots, z_T$ այն նմուշներն են, որոնք մասնակցում են ստոխաստիկ գրադիենտի ալգորիթմում: Նշանակենք $f_t(.) = l(., z_t)$: Դժվար չէ նկատել, որ $v_t = f'_t(w^t)$: Քանի որ f_t ֆունկցիան ուռուցիկ է, ապա

$$f_t(w^t) - f_t(w^*) \leq (v_t, w^t - w^*):$$

Գումարելով այս անհավասարությունները, հաշվի առնելով նաև թեորեմ 5.1-ում ստացված անհավասարությունը՝ ստանում ենք

$$\begin{aligned} \sum_{t=1}^T (f_t(w^t) - f_t(w^*)) &\leq \sum_{t=1}^T (v_t, w^t - w^*) \leq \\ &\leq \frac{\|w^*\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^T \|v_t\|^2 \leq \end{aligned}$$

$$\leq \frac{\|w^*\|^2}{2\eta} + \eta\beta \sum_{t=1}^T f_t(w^t):$$

Բաժանելով այս անհավասարությունը T -ի վրա և կատարելով խմբավորում՝ կստանանք

$$\frac{1}{T} \sum_{t=1}^T f_t(w^t) \leq \frac{1}{1-\eta\beta} \left(\frac{1}{T} \sum_{t=1}^T f_t(w^*) + \frac{\|w^*\|^2}{2\eta T} \right):$$

Վերցնենք այս անհավասարության երկու մասերի մաթեմատիկական սպասումները ըստ $z_1, z_2, \dots, z_T$ պարամետրերի: Կիրառելով թեորեմ 5.8-ի դատողությունները, հաշվի առնելով նաև, որ $E[f_t(w^*)] = L_D(w^*)$ ՝ կստանանք

$$E\left[\frac{1}{T} \sum_{t=1}^T f_t(w^t)\right] = \frac{1}{T} \sum_{t=1}^T L_D(w^t) \geq E[L_D(\bar{w})]:$$

Թերեմն ապացուցվեց:

Հետևանք: Եթե ենթադրենք նաև, որ

$$l(0, z) \leq 1, \quad \eta = \frac{1}{\beta(1+3/\epsilon)},$$

որտեղ ϵ ցանկացած դրական թիվ է, ապա կատարելով $T \geq \frac{12\beta^2\beta}{\epsilon^2}$ քայլ՝ կստանանք հետևյալ գնահատականը՝

$$E[L_D(\bar{w})] \leq \min_{w \in H} L_D(w) + \epsilon:$$

5.13 Ստոխաստիկ գրադիենտի պրոյեկտման մեթոդներ և կիրառություններ

Ինչպես արդեն նշել ենք ներածական մասում, մեքենայական ուսուցման հիմնական խնդիրներից է տրված $S = (x_i, y_i)$, $i = 1, 2, \dots, l$ հավաքածուի օգնությամբ գտնել w պարամետրի այնպիսի արժեք, որը մինիմիզացնում է էմպիրկ ռիսկի

$$Q(a_S) = \frac{1}{l} \sum_{i=1}^l L(a(x_i, w), x_i)$$

ֆունկցիան ըստ w պարամետրի: Այն դեպքում, երբ L -ը դիֆերենցելի է, ապա կարելի է օգտագործել գրադիենտային իջեցման մեթոդներ այդ պարամետրի որոնման համար: Սակայն, եթե S հավաքածուն մեծ ծավալի է, ապա յուրաքանչյուր քայլում պետք է հաշվել $Q'(w^k)$ գրադիենտը, որը կախված է հավաքածուի ամեն մի տարրից.

$$\frac{\partial Q}{\partial w_i}(w^k) = \frac{1}{l} \sum_{j=1}^l \frac{\partial L(a(x_j, w^k), x_j)}{\partial w_i}, \quad i = 1, 2, \dots, l:$$

Այս հաշվողական բարդությունից խուսափելու համար որպես նվազման ուղղություն պատահականորեն ընտրում ենք $-L'(a(x_j, w^k), x_j)$, $j = 1, 2, \dots, l$ հակագրադիենտներից որևէ մեկը, որը Q ֆունկցիայի ստոխաստիկ հակագրադիենտն է:

Այս պարագրաֆում կքննարկենք ստոխաստիկ գրադիենտի պրոյեկտման մեթոդներ, ինչպես նաև դրանցում քայլի ընտրության եղանակները և գուգամիտության հարցեր:

Դիտարկենք հետևյալ խնդիրը՝

$$f_l(w) \equiv \frac{1}{l} \sum_{i=1}^l L_i(w) \rightarrow \min, \quad w \in M,$$

որտեղ M -ը փակ ուռուցիկ բազմություն է, իսկ L_i , $i = 1, 2, \dots, l$ -ը դիֆերենցելի ուռուցիկ ֆունկցիաներ են:

Նկարագրենք գրադիենտային իջեցման այն մեթոդը, որը գեներացնում է $\{w_i\}_{i=1}^l$ վեկտորների հաջորդականություն այնպես, որ

$$\frac{1}{l} \left(\sum_{i=1}^l L_i(w_i) - \min_{w \in M} \sum_{i=1}^l L_i(w) \right) \rightarrow 0, \quad \text{երբ } l \rightarrow \infty:$$

Այդ ընթացակարգը հետևյալն է.

- Input: convex set M , $x_1 \in M$, step sizes $\{\eta_i\}$
- for $i = 1$ to l do
- Play w_i and observe cost $L_i(w)$
- Update

$$y_{i+1} = x_i - \eta_i L'_i(w_i), \quad w_{i+1} = \Pi_M(y_{i+1}),$$

- end for.

Ճշմարիտ է հետևյալ պնդումը:

Թեորեմ 5.17. Դիցուք

1) L_i , $i = 1, 2, \dots$, ֆունկցիաները ուժեղ ուռուցիկ են $\alpha > 0$ հաստատունով՝ M ուռուցիկ բազմության վրա:

2 Գոյություն ունի այնպիսի $G > 0$ հաստատուն, որ

$$\|L'_i(w)\| \leq G, \quad \forall w \in M, \quad i = 1, 2, \dots$$

3) $\eta_i = \frac{1}{\alpha i}$, $i = 1, 2, \dots$:

Այդ դեպքում

$$\frac{1}{l} \left(\sum_{i=1}^l L_i(w_i) - \min_{w \in M} \sum_{i=1}^l L_i(w) \right) \leq \frac{G^2}{2\alpha l} (1 + \ln l):$$

Ապացույց: Քանի որ ըստ ենթադրության L_i ֆունկցիաները ուժեղ ուռուցիկ են միևնույն $\alpha > 0$ հաստատունով, ապա ցանկացած $w^* \in M$ կետի համար ունենք

$$2(L_i(w_i) - L_i(w^*)) \leq 2(L'_i(w_i), w_i - w^*) - \alpha \|w_i - w^*\|^2:$$

Ունենք նաև

$$\|w_{i+1} - w^*\|^2 = \|\Pi_M(w_i - \eta_i L'_i(w_i)) - w^*\|^2 \leq \|(w_i - \eta_i L'_i(w_i)) - w^*\|^2:$$

Հետևաբար՝

$$\|w_{i+1} - w^*\|^2 \leq \|w_i - w^*\|^2 + \eta_i^2 \|L'_i(w_i)\|^2 - 2\eta_i (L'_i(w_i), w_i - w^*):$$

Այստեղից՝

$$2(L'_i(w_i), w_i - w^*) \leq \frac{\|w_i - w^*\|^2 - \|w_{i+1} - w^*\|^2}{\eta_i} + \eta_i G^2:$$

Գումարելով այս անհավասարությունները և հաշվի առնելով, որ $\eta_i = \frac{1}{\alpha i}$, $\frac{1}{\eta_0} \equiv 0$, կստանանք

$$\begin{aligned} 2 \sum_{i=1}^l (L_i(w_i) - L_i(w^*)) &\leq \\ &\leq \sum_{i=1}^l \|w_i - w^*\|^2 \left(\frac{1}{\eta_i} - \frac{1}{\eta_{i-1}} - \alpha \right) + G^2 \sum_{i=1}^l \eta_i = 0 + G^2 \sum_{i=1}^l \frac{1}{\alpha i} \leq \\ &\leq \frac{G^2}{\alpha} (1 + \ln l): \end{aligned}$$

Այստեղից՝

$$\frac{1}{l} \left(\sum_{i=1}^l L_i(w_i) - \min_{w^* \in M} f_l(w^*) \right) \leq \frac{G^2}{2\alpha l} (1 + \ln l) \rightarrow 0:$$

Հետևանք: Եթե

$$L_i(w) = L(w), \quad i = 1, 2, \dots,$$

ապա

$$L(\bar{w}_l) - \min_{w \in M} L(w) \rightarrow 0, \text{ երբ } l \rightarrow \infty,$$

որտեղ

$$\bar{w}_l = \frac{1}{l} (w_1 + w_2 + \dots + w_l):$$

Մեկնաբանենք այս թեորեմը պարզ օրինակի վրա:

Օրինակ 1: Դիցուք $L_i(w) = (w - y_i)^2$, $w \in [0, 1]$, $y_i \in [0, 1]$: Նշանակենք

$$R_l(w) = \sum_{i=1}^l (w_i - y_i)^2 - \sum_{i=1}^l (w - y_i)^2:$$

Ունենք

$$L'_i(w) = 2(w - y_i):$$

Հետևաբար՝ $|l'_i(w)| \leq 2, \forall i, l''_i(w) = 2$: Ուստի թեորեմ 5.17-ի պայմանները կատարվում են $G = 2, \alpha = 2$ պարամետրերով: Կատարելով գրադիենտային իջեցում $\eta_i = 1/(\alpha i), i = 1, 2, \dots$ քայլերով՝ կստանանք

$$\begin{aligned} w_{i+1} &= \Pi_M(w_i - \eta_i L_i(w_i)) = \\ &= \Pi_M(w_i - \frac{1}{i}(w_i - y_i)) = (1 - \frac{1}{i})w_i + \frac{1}{i}y_i = \\ &= \frac{i-1}{i}(\frac{i-2}{i-1}w_{i-1} + \frac{1}{i-1}y_{i-1}) + \frac{1}{i}y_i = \\ &= \frac{i-2}{i}w_{i-1} + \frac{1}{i}y_{i-1} + \frac{1}{i}y_i = \frac{1}{i} \sum_{j=1}^i y_j: \end{aligned}$$

Այստեղից, համաձայն թեորեմ 5.17-ի՝ կստանանք

$$R_l(w) \leq \frac{G^2}{2\alpha}(1 + \ln l) = 1 + \ln l:$$

Օրինակ 2: Վերցնենք, այսպես ասած, Էմպիրիկ ռիսկի ֆունկցիան լոգիստիկ ռեգրեսիայում՝ $L_i(w) = \ln(1 + e^{<w, x_i>y} + \frac{\alpha}{2}\|w\|^2)$: Ցույց տանք, որ եթե M -ը կոմպակտ ուռուցիկ բազմություն է, իսկ նմուշների $X \equiv \{x_i\}_{i=1}^\infty$ բազմությունը սահմանափակ է, ապա թեորեմ 5.17-ի պայմանները բավարարված են: Իրոք, ունենք

$$L'_i(w) = \frac{-x_i y e^{-<w, x_i>y}}{1 + e^{-<w, x_i>y}} + \alpha w:$$

Այս բանաձևից և վերը նշված պայմաններից հետևում է, որ գոյություն ունի մի այնպիսի $G > 0$ հաստատուն, որ $\|L'_i(w)\| \leq G, \forall w \in M, i = 1, 2, \dots$: Իսկ ուժեղ ուռուցիկության պայմանը ակնհայտ է: Լոգիստիկ ռեգրեսիայի վերաբերյալ մանրամասն քննարկում կկատարվի հավանականային դասակարգման բաժնում:

Թեորեմ 5.18. Դիցուք M -ը ուռուցիկ կոմպակտ բազմություն է $g(w)$ -ը $f(w)$ ուռուցիկ դիֆերենցելի ֆունկցիայի ստոխաստիկ գրադիենտն է w կետում՝

$$E[g(w)] = f'(w)$$

և

$$E[\|g(w)\|^2] \leq G^2:$$

Կառուցենք w_k հաջորդականությունը ստոխաստիկ գրադիենտի պրոյեկտման մեթոդով՝

$$w_{k+1} = \Pi_M(w_k - \alpha_k g_k), (g_k = g(w^k)), \alpha_k = \frac{D}{G\sqrt{k}}:$$

Այդ դեպքում

$$E[f(\bar{w}_T)] - \min_{w \in M} f(w) \leq \frac{3GD}{2\sqrt{T}},$$

որտեղ

$$\bar{w}_T = \frac{1}{T}(w_1 + w_2 + \dots + w_T), D = \text{diam}(M):$$

Ապացույց: Թեորեմի ապացույցը անալոգ է թեորեմ 5.7- ապացույցին: Դիցուք w_* -ը f ֆունկցիայի մինիմումի կետն է M բազմության վրա: Այդ դեպքում կունենանք՝

$$\begin{aligned} ||w_{k+1} - w_*||^2 &\leq ||\Pi_M(w_k - \alpha_k g_k) - w_*||^2 \leq \\ &||w_k - \alpha_k g_k - w_*||^2 = \\ &= ||w_k - w_*||^2 + \alpha_k^2 ||g_k||^2 - 2\alpha_k (g_k, w_k - w_*): \end{aligned}$$

Այստեղից, հաշվի առնելով f -ի ուռուցիկությունը և $E[g_k|w_k] = f'(w_k)$, $E||g_k||^2 \leq G^2$ պայմանները, կստանանք

$$\begin{aligned} 2\left(\sum_{k=1}^T E[f(w_k) - f(w_*)|w_k]\right) &\leq 2\sum_{k=1}^T E[(g_k, w_k - w_*)] \leq \\ E\left[\sum_{k=1}^T \frac{||w_k - w_*||^2 - ||w_{k+1} - w_*||^2}{\alpha_k} |w_k\right] &+ G^2 \sum_{k=1}^T \alpha_k \leq \\ &\leq \left[\sum_{k=1}^T ||w_k - w_*||^2 \left(\frac{1}{\alpha_k} - \frac{1}{\alpha_{k+1}}\right)\right] + G^2 \sum_{k=1}^T \alpha_k \leq \\ D^2 \frac{1}{\alpha_T} + G^2 \sum_{k=1}^T \alpha_k &\leq 3DG\sqrt{T}: \end{aligned}$$

Հետևաբար՝

$$E[f(\bar{w}_T)] - f(x^*) \leq \frac{1}{T} \sum_{k=1}^T (f(w^k) - f(w_*)) \leq \frac{3}{2} DG \frac{1}{\sqrt{T}}:$$

Նորից դիտարկենք պայմանական օպտիմիզացիայի ընդհանուր խնդիրը՝

$$f(w) \rightarrow \min, w \in M:$$

Կառուցենք $\{w_k\}$ հաջորդականությունը հետևյալ ռեկուրենտ առնչությամբ՝

$$w_{k+1} = \Pi_M(w_k - \alpha_k g_k), w_0 \in M:$$

Ենթադրենք՝ w_* -ը խնդրի լուծումն է: Ճշմարիտ է հետևյալ թեորեմը:

Թեորեմ 5.19. Դիցուք g_k պատահական վեկտորը բավարարում է հետևյալ պայմանները՝

- M -ը ուռուցիկ կոմպակտ բազմություն է:
- $E[g_k|w_k] = f'(w_k)$, $\forall k$:
- Դիցուք գոյություն ունեն այնպիսի σ , θ , L դրական թվեր, որ

$$E[||g_k - f'(w_k)||^2 | w_k] \leq \sigma^2,$$

$$f(u) - f(v) \geq (f'(v), u - v) + \frac{L}{2} ||u - v||^2, u, v \in M,$$

$$f(u) - f(v) \geq (f'(v), u - v) + \frac{\theta}{2} \|u - v\|^2, \quad x, y \in M:$$

Դիցուք

$$\alpha_k \equiv \alpha < 1/2L, \quad \|f'(w)\|^2 \leq G, \quad \forall w \in M:$$

Այդ դեպքում տեղի ունի հետևյալ անհավասարությունը՝

$$E[\|w_k - w_*\|^2] \leq (1 - \alpha\theta)^k \|w_0 - w_*\|^2 + \frac{\alpha(\sigma^2 + 2G)}{\theta}:$$

Ապացույց: Հաշվի առնելով, որ $w_{k+1} = \Pi_M(w_k - \alpha g_k)$, կունենանք

$$\begin{aligned} \|w_{k+1} - w_*\|^2 &= \|\Pi_M(w_k - \alpha g_k) - \Pi_M(w_*)\|^2 \leq \\ &\leq \|w_k - \alpha g_k - w_*\|^2 = \|w_k - w_*\|^2 - 2\alpha(w_k - w_*, g_k) + \alpha^2 \|g_k\|^2: \end{aligned}$$

Վերցնենք այս հավասարության երկու մասի պայմանական $E[., x_k]$ մաթեմատիկական սպասումը. կստանանք՝

$$E[\|w_{k+1} - w_*\|^2 | w_k] = \|w_k - w_*\|^2 - 2\alpha(w_k - w_*, f'(w_k)) + \alpha^2 E[\|g_k\|^2 | w_k]:$$

Այժմ գնահատենք $E[\|g_k\|^2 | w_k]$ մաթեմատիկական սպասումը: Ունենք

$$\begin{aligned} E[\|g_k\|^2 | w_k] &= E[\|f'(w_k) + g_k - f'(w_k)\|^2] = \|f'(w_k)\|^2 + 2E[(f'(w_k), g_k - f'(w_k)) | w_k] + \\ &+ E[\|g_k - f'(w_k)\|^2 | w_k] = \|f'(w_k)\|^2 + E[\|g_k - f'(w_k)\|^2 | w_k] \leq \|f'(w_k)\|^2 + \sigma^2: \end{aligned}$$

Որպեսզի գնահատենք $\|f'(w_k)\|^2$ արտահայտության արժեքը վերևից, օգտվենք հետևյալ անհավասարությունից՝

$$\|f'(u) - f'(v)\|^2 \leq 2L(f(u) - f(v) - (f'(v), u - v)) \quad (\text{տե՛ս, [39], թեորեմ 2.1.5}):$$

Այս անհավասարության մեջ տեղադրենք $x = w_k$, $y = w_*$ և հաշվի առնելով, որ $(f'(w_*), w_k - w_*) \geq 0$ ՝ կստանանք՝

$$E[\|g_k\|^2 | w_k] \leq 4L(f(w_k) - f(w_*)) + \sigma^2 + 2\|f'(w_*)\|^2:$$

Տեղադրելով այս գնահատականը $E[\|w_{k+1} - w_*\|^2 | w_k]$ արտահայտության մեջ՝ կստանանք

$$E[\|w_{k+1} - w_*\|^2 | w_k] \leq \|w_k - w_*\|^2 - 2\alpha(w_k - w_*, f'(w_k)) + 4\alpha^2 L(f(w_k) - f(w_*)) + \alpha^2(\sigma^2 + 2\|f'(w_*)\|^2):$$

Այժմ գնահատենք այս անհավասարության աջ մասի սկայար արտադրյալը: Քանի որ f ֆունկցիան ուժեղ ուռուցիկ է, ապա

$$f(w_*) - f(w_k) \geq (f'(w_k), w_* - w_k) + \frac{\theta}{2} \|w_k - w_*\|^2:$$

Այստեղից հետևում է, որ

$$(f'(w_k), w_k - w_*) \geq f(w_k) - f(w_*) + \frac{\theta}{2} \|w_k - w_*\|^2:$$

Հետևաբար կունենանք՝

$$E[\|w_{k+1} - w_*\|^2 | w_k] \leq (1 - \alpha\theta) \|w_k - w_*\|^2 - 2\alpha(1 - 2\alpha L)(f(w_k) - f(w_*)) + \alpha^2(\sigma^2 + 2\|f'(w_*)\|^2):$$

Եթե $\alpha < 1/2L$, ապա վերևի անհավասարությունից կստանանք

$$E_{w_k}[E[\|w_{k+1} - w_*\|^2 | w_k]] \leq (1 - \theta\alpha)E[\|w_k - w_*\|^2] + \alpha^2(\sigma^2 + 2\|f'(w_*)\|^2):$$

Բայց հայտնի է, որ

$$E_{w_k}[E[\|w_{k+1} - w_*\|^2 | w_k]] = E[\|w_{k+1} - w_*\|^2]:$$

Այսպիսով՝

$$\begin{aligned} E[\|w_{k+1} - w_*\|^2] &\leq (1 - \alpha\theta)^{k+1} \|w_0 - w_*\|^2 + \alpha^2(\sigma^2 + \|f'(w_*)\|^2) \sum_{i=0}^k (1 - \alpha\theta)^i \leq \\ &\leq (1 - \alpha\theta)^k \|w_0 - w_*\|^k + \frac{\alpha(\sigma^2 + 2G)}{\theta}: \end{aligned}$$

Այժմ ենթադրենք, որ կատարում ենք T քայլեր: Վերցնենք բոլոր քայլերում քայլի հաստատուն՝ $\alpha_i = 1/(2L + \sqrt{T})$, $i = 1, 2, \dots, T$: Այդ դեպքում կունենանք

$$\begin{aligned} E[\|w_T - w_*\|^2] &\leq \left(1 - \frac{1}{(2L + \sqrt{T})\theta}\right)^T \|w_0 - w_*\|^2 + \frac{\sigma^2 + 2G}{\theta(2L + \sqrt{T})} \leq \\ &\|w_0 - w_*\|^2 \exp\left(-\frac{\theta\sqrt{T}}{1 + 2L}\right) + \frac{\sigma^2 + 2G}{2L + \sqrt{T}}: \end{aligned}$$

Այժմ դիտարկենք քայլի ընտրության ևս մի եղանակ՝

$$\alpha_k = \frac{2}{\alpha(k+1)}:$$

Ցույց տանք, որ

$$E[\|w_k - w_*\|^2] \leq \frac{R_1}{k+1}, \quad k > \frac{4L}{\theta},$$

որտեղ

$$R_1 = \max\{\|w_0 - w_*\|^2, (\sigma^2 + 2G)/\theta^2\}:$$

Նշանակենք $\epsilon_k = E[\|w_k - w_*\|^2]$ և ինդուկցիայով ցույց տանք, որ

$$\epsilon_k \leq \frac{R_1}{k+1}:$$

$k = 0$ -ի դեպքում անհավասարությունը ակնհայտ է: Ենթադրենք այն ճիշտ է $k = m$ -ի

համար: Ապացուցենք այն $k = m + 1$ -ի համար: Իրոք, ունենք

$$\begin{aligned}\epsilon_{m+1} &\leq (1 - \theta\alpha_m)\epsilon_m + \alpha_m(\sigma^2 + 2G) \leq \\ &(1 - \frac{2}{m+1})\epsilon_m + \frac{\sigma^2 + 2G}{(\theta m)^2} \leq \\ &(1 - \frac{2}{m+1})\frac{R_1}{m+1} + \frac{R_1}{(m+1)^2} = \frac{m}{(m+1)^2}R_1 \leq \frac{R_1}{m+2}:\end{aligned}$$

Հետևանք: Ենթադրենք տեղի ունեն թեորեմ 5.19-ի բոլոր պայմանները և

$$\alpha_k = \frac{2}{\alpha(1+k)}:$$

Այդ դեպքում

$$w_k \xrightarrow{\mathbb{P}} w_*:$$

Այսինքն՝ $\{w_k\}$ հաջորդականությունը $P = 1$ հավանականությամբ գոյգամիտում է w_* -ի:

Իրոք, համաձայն Մարկովի անհավասարության, ունենք

$$P\left(\|w_k - w_*\| \geq \epsilon\right) = P\left(\|w_k - w_*\|^2 \geq \epsilon^2\right) \leq \frac{E[\|w_k - w_*\|^2]}{\epsilon^2} \leq$$

$$\frac{R_1}{\epsilon^2(k+1)} \rightarrow 0, \text{ երբ } \rightarrow \infty:$$

Օրինակ 3: Դիցուք

$$f(w) = \frac{1}{l} \sum_{i=1}^l l_i(w):$$

Այս դեպքում որպես պատահական g_k վեկտոր (ստոխաստիկ գրադիենտ) վերցնենք կամայական $L'_{\xi_k}(w_k)$ գրադիենտը k -րդ քայլում, որտեղ ξ_k պատահական մեծությունն է՝ հավասարաչափ բաշխված $\{1, 2, \dots, l\}$ բազմության վրա: Այդ դեպքում թեորեմ 5.12-ի բոլոր պայմանները կկատարվեն, եթե պահանջենք, որ բոլոր ֆունկցիաները ուժեղ ուռուցիկ են, և նրանց գրադիենտները բավարարում են Լիպշիցի պայմանը M ուռուցիկ կոմպակտի վրա:

Այժմ այս օրինակում դիտարկված ֆունկցիայի նկատմամբ կիրառենք ստոխաստիկ գրադիենտի պրոյեկտման մեթոդը փոփոխական քայլերի դեպքում՝

$$w_{k+1} = \Pi_M(w_k - \alpha_k l'_{\xi_k}(w_k)):$$

Տեղի ունի հետևյալ թեորեմը:

Թեորեմ 5.20. Դիցուք

- M -ը ուռուցիկ կոմպակտ բազմություն է:
- $L_i, \quad i = 1, 2, \dots, l$ ֆունկցիաները θ -ուժեղ ուռուցիկ ֆունկցիաներ են M կոմպակտի վրա::

- $\alpha_k = 1/(\theta(k+1))$:

Այդ դեպքում տեղի ունի հետևյալ գնահատականը՝

$$E[||w_k - w_*||^2] \leq \frac{R_2}{k+1},$$

որտեղ R_2 -ը ինչ-որ դրական հաստատուն է:

Ապացույց: Քանի որ l_i ֆունկցիաները ուժեղ ուռուցիկ են θ հաստատունով, ապա այդպիսին է նաև f ֆունկցիան, և հետևաբար՝

$$f(w_*) - f(w_k) \geq (f'(w_k), w_* - w_k) + \frac{\theta}{2} ||w_k - w_*||^2,$$

$$f(w_k) - f(w_*) \geq (f'(w_*), w_k - w_*) + \frac{\theta}{2} ||w_k - w_*||^2:$$

Ունենք նաև $(f'(w_*), w_k - w_*) \geq 0$ (քանի որ w_* -ը f -ի մինիմումի կետն է M բազմության վրա): Այստեղից, գումարելով այս անհավասարությունները, կստանանք

$$\theta ||w_k - w_*||^2 \leq (f'(w_k), w_k - w_*):$$

Հաշվի առնելով այս անհավասարությունը և $E_{\xi_k}[l'_{\xi_k}(w_k)] = f'(w_k)$ պայմանը՝

$$\begin{aligned} E[||w_{k+1} - w_*||^2 | \xi_k] &\leq E[||w_k - \alpha_k l'_{\xi_k}(w_k) - w_*||^2] = \\ &||w_k - w_*||^2 + 2\alpha_k (E_{\xi_k}[(l'_{\xi_k}(w_k), w_* - w_k)] + \alpha_k^2 E_{\xi_k}[||l'_{\xi_k}(w_k)||^2] \leq \\ &||w_k - w_*||^2 + 2\alpha_k (f'(w_k), w_* - w_k) + \alpha_k^2 C \leq \\ &(1 - 2\theta\alpha_k) ||w_k - w_*||^2 + \alpha_k^2 C, \end{aligned}$$

որտեղ $C \geq ||l'(w)||$, $w \in M$: Այժմ վերցնելով այս անհավասարության երկու կողմերի մաթեմատիկական սպասումները ըստ ξ_k պատահական մեծության՝ կստանանք

$$E[||w_k - w_*||^2] \leq (1 - 2\theta\alpha_k) E[||w_k - w_*||^2] + \alpha_k^2 C:$$

Նշանակենք $\epsilon_k = E[||w_k - w_*||^2]$ և ինդուկցիայով ցույց տանք, որ

$$\epsilon_k \leq \frac{R_2}{k+1},$$

որտեղ $R_2 = \max\{||w_0 - w_*||^2, C/\theta^2\}$: $k = 0$ -ի դեպքում անհավասարությունը ակնհայտ է: Ենթադրենք այն ճիշտ է $k = m$ -ի համար: Ապացուցենք այն $k = m+1$ -ի համար: Իրոք, ունենք

$$\begin{aligned} \epsilon_{m+1} &\leq (1 - 2\theta\alpha_m)\epsilon_m + \alpha_m C \leq \\ &(1 - \frac{2}{m+1})\epsilon_m + \frac{C}{(\theta m)^2} \leq \\ &(1 - \frac{2}{m+1})\frac{R_2}{m+1} + \frac{R_2}{(m+1)^2} = \frac{m}{(m+1)^2} R_2 \leq \frac{R_2}{m+2}: \end{aligned}$$

Այժմ գծային ռեգրեսիայի խնդրում նկարագրենք գրադիենտային և ստոխաստիկ գրադիենտային մեթոդները և կատարենք համեմատություն: Ունենք

$$Q = \frac{1}{l} \sum_{i=1}^l ((\mathbf{w}, \mathbf{x}_i) - y_i)^2 \equiv$$

$$\frac{1}{l} \sum_{i=1}^l (w_0 + w_1 x_{1i} + w_2 x_{2i} + \dots + w_n x_{ni} - y_i)^2:$$

Գրադիենտային եղանակի դեպքում ունենք հետևյալ ընթացակարգը.

- $w_0, w_1, \dots, w_n$ կշիռներին տրվում է նախնական արժեքներ:
- Հաշվելով Q ֆունկցիայի մասնակի ածանցյալները ըստ այդ կշիռների՝ կատանանք

$$\frac{\partial Q}{\partial w_0} = \frac{2}{l} \sum_{i=1}^l (w_0 + w_1 x_{1i} + w_2 x_{2i} + \dots + w_n x_{ni} - y_i),$$

$$\frac{\partial Q}{\partial w_k} = \frac{2}{l} \sum_{i=1}^l (w_0 + w_1 x_{1i} + \dots + w_n x_{ni} - y_i) x_{ki}, \quad k = 1, 2, \dots, n:$$

- Կշիռների թարմացված հաշվարկ կատարվում է գրադիենտային իջեցման մեթոդով (հաստանում h քայլի դեպքում) հետևյալ բանաձևերով՝

$$w_i^t = w_i^{t-1} - h \frac{\partial Q}{\partial w_i}, \quad i = 0, 1, \dots, n:$$

- Հաշվում ենք միջին քառակուսային շեղումը t -րդ քայլում՝

$$Q^t = \frac{1}{l} \sum_{i=1}^n (w_0^t + w_1^t x_{1i} + w_2^t x_{2i} + \dots + w_n^t x_{ni} - y_i)^2:$$

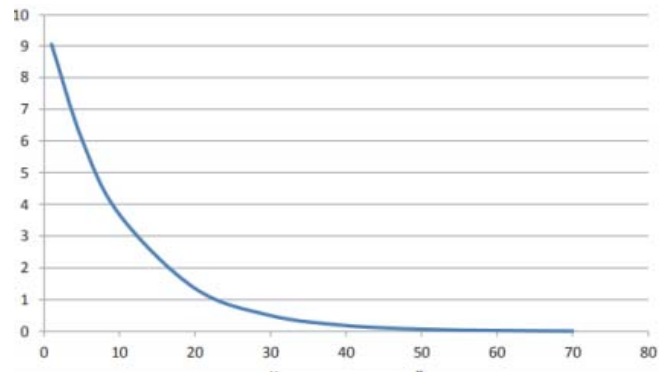
- Ստուգում ենք

$$|Q^t - Q^{t-1}| < \epsilon$$

կանգառի պայմանը, որտեղ ϵ -ը նախապես տրված ճշտություն է: Եթե այն կատարվում է, ապա ցիկլը ավարտվում է, հակառակ դեպքում այգորիթմը շարունակվում է:

Նկար 5.2-ում նկարագրված է սխալանքի նվազման գրաֆիկը գրադիենտային մեթոդի դեպքում: Ինչպես երևում է գծագրից սկզբնական իտերացիաների ընթացքում կատարվում է սխալանքի կտրուկ նվազում: Իսկ 40 իտերացիայից հետո սխալանքի նվազումը արդեն աննշան է: Այս մեթոդի հիմնական թերությունն այն է, որ յուրաքանչյուր իտերացիայի ընթացքում պետք է հաշվել Q ֆունկցիոնալի գրադիենտը ամբողջ հավաքածուի վրա, որը կարող է մեծ չափերի լինել: Այդ թերությունը վերացնում են՝ օգտագործելով ստոխաստիկ գրադիենտային մեթոդը: Սկզբում հավաքածուից վերցվում է կամայական օբյեկտ, իսկ հաջորդ

իտերացիայում պատահականորեն ընտրվում է ուրիշը: Նկարագրենք այս ավգորիթմը հաստատուն քայլերի դեպքում:



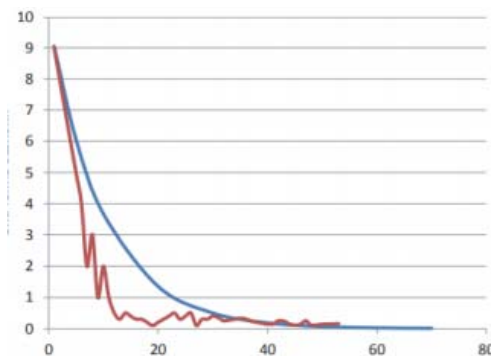
Նկ. 5.2: Գրադիենտային մեթոդը գծային ռեգրեսիայի դեպքում

- $w_0, w_1, \dots, w_n$ կշիռներին տրվում է պատահական սկզբնական արժեքներ:
- հավաքածուից պատահականորեն ընտրվում է $x_i = (x_{1i}, \dots, x_{ni})$ օբյեկտ և հաշվվում են հետևյալ մասնակի ածանցյալները՝

$$\frac{\partial Q}{\partial w_0} = w_0 + w_1 x_{1i} + \dots + w_n x_{ni} - y_i,$$

$$\frac{\partial Q}{\partial w_j} = (w_0 + w_1 x_{1i} + \dots + w_n x_{ni} - y_i) x_{ji}:$$

Այնուհետև կշիռների թարմացումը կատարվում է վերը նշված բանաձևերով:



Նկ. 5.3: Ստոխաստիկ գրադիենտային մեթոդը գծային ռեգրեսիայի դեպքում

Ինչպես երևում է գծագիր 5.3-ից, ստոխաստիկ գրադիենտային մեթոդը զուգամիտում է ավելի արագ, քան սովորականը: Սակայն զուգամիտությունը ոչ կայուն է: Դա կապված է այն բանի հետ, որ իտերացիայից իտերացիա կատարվում է օբյեկտի փոփոխություն:

5.14 Ստոխաստիկ սուբգրադիենտային մեթոդներ

Նորից դիտարկենք α -ուռուցիկ f ֆունկցիայի մինիմիզացիայի խնդիրը M ուռուցիկ բազմության վրա.

$$f(w) \rightarrow \min, w \in M:$$

Նկարագրենք այս խնդրի լուծման ստոխաստիկ սուբգրադիենտային իջեցման հետևյալ ալգորիթը.

- $w_1 \in M$:
- Դիցուք v_t -ն պատահական վեկտոր է, այնպիսին, որ

$$E[v_t | w_t] \in \partial f(w_t):$$

- $w_{t+1} = \Pi_M(w_t - \eta_t v_t)$, $\eta_t = 1/(\alpha t)$:

Թեորեմ 5.21. Դիցուք w^* -ը f α -ուռուցիկ ֆունկցիայի մինիմումի կետն է M -ի վրա, և $E[||v_t||^2] \leq \rho^2$: Այդ դեպքում

$$E[f(\bar{w}_T)] - f(w^*) \leq \frac{\rho^2}{2\alpha T} (1 + \ln T),$$

որտեղ $\bar{w}_T = \frac{1}{T}(w_1 + w_2 + \dots + w_T)$:

Ապացույց: Նշանակենք՝

$$g_t = E[v_t | w_t]:$$

Քանի որ f α -ուժեղ ուռուցիկ է, իսկ $g_t \in \partial f(w_t)$, ապա

$$(w_t - w^*, g_t) \geq f(w_t) - f(w^*) + \frac{\alpha}{2} ||w_t - w^*||^2:$$

Ճշմարիտ է նաև հետևյալ անհավասարությունը՝

$$\begin{aligned} ||w_t - w^*||^2 - ||w_{t+1} - w^*||^2 &\geq ||w_t - w^*||^2 - ||(w_t - \eta_t v_t) - w^*||^2 = \\ &= 2\eta_t (w_t - w^*, v_t) - \eta_t^2 ||v_t||^2: \end{aligned}$$

Վերցնելով այս անհավասարության երկու մասերի մաթեմատիկական սպասումները և հաշվի առնելով, որ $E[||v_t||^2] \leq \rho^2$, կստանանք՝

$$(w_t - w^*, g_t) \leq \frac{E[||w_t - w^*||^2 - ||w_{t+1} - w^*||^2]}{2\eta_t} + \frac{\eta_t}{2} \rho^2:$$

Այստեղից, հաշվի առնելով f ֆունկցիայի α -ուժեղ ուռուցիկության պայմանը՝

կստանանք

$$\begin{aligned} & \sum_{t=1}^T (E[f(w_t) - f(w^*)]) \leq \\ & \leq E\left[\sum_{t=1}^T \left(\frac{\|w_t - w^*\|^2 - \|w_{t+1} - w^*\|^2}{2\eta_t} - \frac{\alpha}{2}\|w_t - w^*\|^2\right) + \frac{\rho^2}{2} \sum_{t=1}^T \eta_t\right] : \end{aligned}$$

Դժվար չէ նկատել, որ անհավասարության աջ մասի առաջին գումարելին հավասար է $-\alpha T \|w_{T+1} - w^*\|^2 \leq 0$: Մյուս կողմից, քանի որ $\eta_t = 1/(\alpha t)$, ապա

$$\sum_{t=1}^T (E[f(w_t)] - f(w^*)) \leq \frac{\rho^2}{2\alpha} \sum_{t=1}^T \frac{1}{t} \leq \frac{\rho^2}{2\alpha} (1 + \ln T) :$$

Բաժանելով այս անհավասարության երկու կողմերը T -ի և օգտվելով Յենսենի անհավասարությունից՝ կստանանք թեորեմի պնդումը: Այժմ գնահատենք, թե քանի իտերացիայից հետո տեղի կունենա

$$E[f(\bar{w})] \leq \min_w f(w) + \epsilon$$

անհավասարությունը: Դրա համար բավական է պահանջել, որ

$$\frac{\rho^2}{2\alpha T} (1 + \ln T) \leq \epsilon :$$

Այսինքն՝

$$T \geq \frac{\rho^2}{2\alpha\epsilon} + \frac{\rho^2}{2\alpha\epsilon} \ln T :$$

Համաձայն լեմմա 2.3-ի՝ $x \geq a \ln x + b$, ($a \geq 1$ $b > 0$) անհավասարությունը տեղի ունենալու բավարար պայման է $x \geq 4a \ln(2a) + 2b$ անհավասարությունը: Մեր օրինակում

$$a = b = \frac{\rho^2}{2\alpha\epsilon} :$$

Ենթադրենք, որ

$$\frac{\rho^2}{2\alpha\epsilon} \geq 1 \equiv \epsilon \leq \frac{\rho^2}{2\alpha} :$$

Հետևաբար, եթե

$$T \geq 4a \ln(2a) = \frac{2\rho^2}{\alpha\epsilon} \ln \frac{\rho^2}{\alpha\epsilon} + \frac{\rho^2}{\alpha\epsilon},$$

ապա կստանանք նշված ϵ ճշտությունը: Այսպիսով այս ալգորիթմով կառուցվում է ϵ -օպտիմալ պրեդիկտոր:

Վերջում նշենք, որ եթե օբյեկտների քանակը (l թիվը) մեծ չէ, ապա կարելի է պարզապես կատարել ստոխաստիկ սուգրադիենտի պրոյեկտում: Օրինակ՝ դիտարկենք գծային ռեգրեսիայի խնդիրը l մաթրիկ կորստի հետևյալ ֆունկցիոնալով՝

$$Q_1 = \frac{1}{l} \sum_{i=1}^l |(w, x_i) - y_i| = \frac{1}{l} \|Aw - y\|_1 \rightarrow \min, \quad w \in M = \{w : \|w\|_{L_p} \leq r\} : \quad p = 2, \quad p = \infty :$$

Այստեղ $A^{l \times n}$ մատրից է, $x_i \in \mathbb{R}^n$, $y \in \mathbb{R}^l$: Ուռուցիկ ֆունկցիայի սուբդի-ֆերենցիայի բաժնում ցույց է տրված, որ

$$g \equiv \frac{1}{l} \sum_{i \in I_1(w)} \text{sign}((w, x_i) - y_i) x_i^T \in \partial Q_1(w), \quad I_1(w) = \{i : (w, x_i) \neq y_i\}:$$

Հետևաբար ստոխաստիկ սուբգրադիենտի իջեցման մեթոդը կիրառելով այս դեպքում՝ կստանանք

$$w_{t+1} = \Pi_M(w_t - \frac{\eta_t}{l} \text{sign}((x_i, w_t)) - y_{\xi_t}) x_{\xi_t}^T,$$

որտեղ ξ_t պատահական մեծություն է, որը հավասար հավանականությամբ արժեքներ է ընդունում $I_1(w_t)$ ինդեքսների բազմությունից:

L_2 նորմի դեպքում պրոյեկտման օպերատորը ունի հետևյալ տեսքը՝

$$\Pi_M(z) = \frac{z}{\max\{1, \|z\|_2/r\}}:$$

Իսկ L_∞ նորմի դեպքում պրոյեկտման օպերատորը հետևյալն է՝

$$(\Pi_M(z))_i = \begin{cases} r, & z_i \geq 1, \\ z_i, & z_i \in (-1, 1), \quad i = 1, 2, \dots, n, \\ -r, & z_i \leq -1: \end{cases}$$

5.15 L_1 և L_2 կարգավորիչների մասին

Դիցուք $f(w)$ -ը ուռուցիկ ֆունկցիա է՝ որոշված $\mathbb{R}^n$ -ի վրա: Դիտարկենք հետևյալ երկու խնդիրները՝

$$f(w) \rightarrow \min, \quad \|w\|_2^2 \leq \gamma$$

և

$$f(w) + \alpha \|w\|_2^2 \rightarrow \min:$$

Թեորեմ 5.22. Այս երկու խնդիրները համարժեք են հետևյալ իմաստով:

- Եթե w^* -ը առաջին խնդրի լուծումն է, ապա գոյություն ունի այնպիսի $\alpha > 0$, որ այն այդ պարամետրով երկրորդ խնդրի լուծումն է:
- Եթե w^* -ը երկրորդ խնդրի լուծումն է, ապա այն առաջինի լուծումն է γ պարամետրի ինչ-որ արժեքի դեպքում:

Ապացույց: Առաջին պնդումը հետևում է Կուն-Թակերի թեորեմից:

Դիցուք w^* -ն երկրորդ խնդրի լուծումն է, որ միշտ գոյություն ունի քանի որ

$$f(w) + \alpha \|w\|^2$$

ֆունկցիան ուժեղ ուռուցիկ է, և հետևաբար ունի միակ մինիմում $\mathbb{R}^n$ -ի վրա:

Վերցնենք $\gamma = \|w^*\|^2$. Այդ դեպքում եթե $\|w\| \leq \gamma$, ապա

$$f(w) + \alpha(\|w\|^2 - \gamma) \geq f(w^*) + \alpha(\|w^*\|^2 - \gamma) = f(w^*):$$

Հետևաբար՝

$$f(w) \geq f(w^*), \quad \|w\|^2 \leq \gamma:$$

Ռեգրեսիայի խնդիրներում սովորաբար դրվում են սահմանափակումներ կշռային վեկտորի վրա:

Այժմ քննարկենք հետևյալ երկու խնդիրները:

$$Q_2 \equiv \frac{1}{l} \sum_{i=1}^l (w_0 + w_1 x_{1i} + w_2 x_{2i} + \dots + w_n x_{ni} - y_i)^2 \rightarrow \min_{w_0, w}, \quad \|w\|_2 \equiv w_1^2 + \dots + w_n^2 \leq m,$$

$$Q_2 \rightarrow \min_{w_0, w}, \quad \|w\|_1 \equiv |w_1| + \dots + |w_n| \leq m:$$

Առաջինը կոչվում է *Ridge* (L_2), իսկ երկրորդը՝ *Lasso* (L_1) ռեգրեսիայի խնդիրներ:

Դիտարկենք երկրորդ խնդիրը մանրամասնորեն: Նշանակենք $L_1(w) \equiv \|w\|_1 = |w_1| + |w_2| + \dots + |w_n|$: Այս ֆունկցիան գրականության մեջ կոչվում է L_1 ռեգուլյատոր: Համաձայն Կուն-Թակերի թեորեմի՝ գոյություն ունի այնպիսի $\alpha \geq 0$ թիվ, որ

$$N_\alpha(w_0, w) \equiv Q_2(w_0, w) + \alpha L_1(w)$$

ֆունկցիայի մինիմումի կետը ամբողջ տարածության վրա սկզբնական խնդրի օպտիմալ լուծումն է: Քանի որ N_α ֆունկցիան ուռուցիկ է, ապա մինիմումի կետում տեղի ունեն Կուն-Թակերի թեորեմի անհրաժեշտ և բավարար պայմանները $w_0, w_1, \dots, w_n$ փոփոխականների նկատմամբ: Քանի որ L_1 ռեգուլյատորը կախված չէ w_0 փոփոխականից, ապա օպտիմալության պայմանն այդ պարամետրի նկատմամբ ունի հետևյալ տեսքը՝

$$\sum_{i=1}^l (w_0 + w_1 x_{i1} + \dots + w_n x_{in} - y_i) = 0:$$

Իսկ այդ պայմանները $w_1, \dots, w_n$ փոփոխականների նկատմամբ հետևյալն են՝

$$\frac{\partial N_\alpha}{\partial w_j} = 0, \quad \text{եթե } w_j \neq 0, \quad j = 1, 2, \dots, n,$$

$$0 \in \partial_{w_j} N_\alpha(w_0, w), \quad \text{եթե } w_j = 0, \quad j = 1, 2, \dots, n:$$

Քննարկենք այս երկու պայմանները մանրամասնորեն:

Նշանակենք $\langle w, x_i \rangle \equiv w_0 + w_1 x_{1i} + \dots + w_n x_{ni}$: Համաձայն այս նշանակման՝ կունենանք՝

$$\frac{\partial Q}{\partial w_j} = \frac{2}{l} x_j (\langle w, x_j \rangle - y_j) = -\frac{2}{l} x_j (y_j - \langle w^j, x^j \rangle + w_j x_j) =$$

$$= -\frac{2}{l}x_j(y_j - \langle w^j, x^j \rangle) + \frac{2}{l}w_j||x_j||^2,$$

որտեղ $\langle w^j, x^j \rangle$ սիմվոլով նշանակված է սկալյար արտադրյալը առանց j -րդ կոմպոնենտի: Նշանակենք

$$\rho_j = \frac{2}{l}x_j(y_j - \langle w^j, x^j \rangle), \quad z_j = \frac{2}{l}||x_j||^2:$$

Վերջնականում կստանանք հետևյալ անհրաժեշտ և բավարար պայմանը մինիմումի կետի համար՝

$$0 \in \partial_{w_j} N_\alpha(w_0, w) = -\rho_j + w_j z_j + \begin{cases} -\alpha, & w_j < 0, \\ [-\alpha, \alpha], & w_j = 0, \quad j = 1, 2, \dots, n \\ \alpha, & w_j > 0: \end{cases}$$

Քննարկենք այս դեպքերը առանձին-առանձին:

- $w_j < 0$ դեպքում ունենք

$$-\rho_j + w_j z_j - \alpha = 0:$$

Այստեղից՝

$$w_j = \frac{\rho_j - \alpha}{z_j}:$$

- $w_j = 0$ դեպքում կունենանք՝

$$0 \in -\rho_j + w_j z_j + [-\alpha, \alpha]:$$

Հետևաբար՝

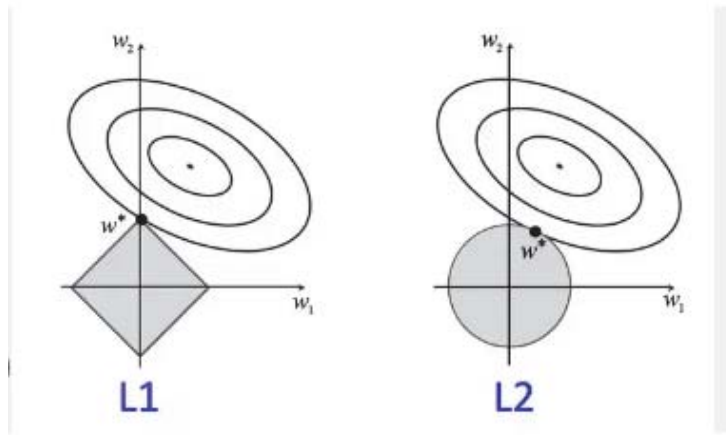
$$-\alpha \leq \rho_j \leq \alpha:$$

- $w_j > 0$ դեպքում

$$w_j = \frac{\rho_j + \alpha}{z_j}:$$

Ինչպես երևում է վերը նկարագրած ալգորիթմից, որոշ կշիռներ *Lasso* ռեգրեսիայի դեպքում հնարավոր է գրոյանան: Այս փաստը կարելի է մեկնաբանել երկրաչափորեն:

Երկրաչափորեն լուծել նշված խնդիրները նշանակում է գտնել նպատակային ֆունկցիայի մակարդակի գծերի և տրված բազմության շոշափման կետը: Ինտուիտիվորեն պարզ է, որ *Lasso* ռեգրեսիայի դեպքում այդ կետը, ամենայն հավանականությամբ, կտեղակայվի շեղանկյան գագաթում, մինչդեռ *Ridge* ռեգրեսիայի դեպքում դա տեղի է ունենում շատ հազվադեպ: Իսկ դա նշանակում է՝ կշիռներից մեկը հավասարվում է գրոյի (տե՛ս նկ.5.4):



Նկ. 5.4: Lasso, Ridge ռեգրեսիաները

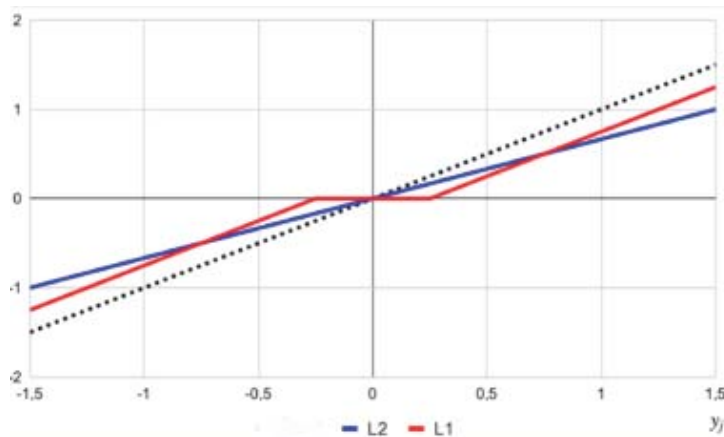
Համեմատենք այս երկու ռեգրեսիաները պարզ մոդելի վրա: Դիցուք $x_i = e_i = (0, \dots, 1, \dots, 0)$, $i = 1, 2, \dots, n$: Այդ դեպքում

$$w_j = \begin{cases} y_j, & \text{ռեգրեսիայի բացակայության դեպքում} \\ \frac{y_j}{1+\alpha}, & \text{Ridge ռեգրեսիայի դեպքում} \end{cases} :$$

Իսկ *Lasso* ռեգրեսիայի դեպքում

$$w_j = \begin{cases} y_j - \alpha/2, & \text{եթե } y_j > \alpha/2, \\ y_j + \alpha/2, & \text{եթե } y_j < -\alpha/2, \\ 0, & \text{եթե, } |y_j| \leq \alpha/2: \end{cases} :$$

Այս կախվածությունները ներկայացված են գրաֆիկորեն (նկ.5 5): Այս պարզ մոդելից երևում է՝ *Ridge* ռեգրեսիայի դեպքում կշիռները փոքրանում են:



Նկ. 5.5: Ridge, Lasso ռեգրեսիաներ պարզագույն մոդելի դեպքում

Այժմ դիտարկենք L_2 ռեգուլյարիզատորը միջին կորստի մինիմիզացիայի հետևյալ ընդհանուր խնդրում՝

$$\min_w \{f(w) \equiv \frac{\alpha}{2} \|w\|^2 + \frac{1}{l} \sum_{i=1}^l l(w, z_i)\} :$$

Խնդիրը լուծենք ստոխաստիկ սուրգրադիենտային իջեցման հետևյալ ալգորիթով.

- $w^0 = 0$,
- կատարում ենք թարմացում հետևյալ բանաձևով՝

$$w^{t+1} = w_t - \frac{1}{\alpha t}(\alpha w^t + v_t),$$

որտեղ $v_t \in \partial l(w^t, x)$, իսկ z -ը պատահական վեկտոր է, որը հավասար հավանականությամբ ընդունում է $\{z_1, z_2, \dots, z_l\}$ արժեքները:

Թեորեմ 5.23. Դիցուք $L(w, z_i)$, $i = 1, 2, \dots, l$ ուռուցիկ ֆունկցիաները, քստ w -ի, բավարարում են L իպոթիզի պայմանը ρ հաստատունով: Այդ դեպքում տեղի ունի հետևյալ անհավասարությունը՝

$$E[f(\bar{w}^T)] - \min_w f(w) \leq \frac{4\rho^2}{\alpha T}(1 + \ln T), \quad \bar{w}^T = \frac{1}{T}(w^1 + w^2 + \dots + w^T):$$

Ապացույց: Ունենք

$$\begin{aligned} w^{t+1} &= w^t - \frac{1}{\lambda t}(\lambda w^t + v_t) = (1 - \frac{1}{t})w^t - \frac{1}{\lambda t}v_t = \\ &= \frac{t-1}{t}w^t - \frac{1}{\lambda t}v_t = \frac{t-1}{t}\left(\frac{t-2}{t-1}w^{t-1} - \frac{1}{\lambda(t-1)}v_{t-1}\right) - \frac{1}{\lambda t}v_t = -\frac{1}{\lambda t}\sum_{i=1}^t v_i: \end{aligned}$$

Քանի որ ենթադրել ենք, որ կորստի $l(w, x_i)$ ֆունկցիաները բավարարում են L իպոթիզի պայմանը ρ հաստատունով, ապա $\|v_t\| \leq \rho$: Հետևաբար վերևի հավասարությունից հետևում է, որ

$$\|\lambda w^t + v_t\| \leq 2\rho:$$

Այստեղից, հաշվի առնելով թեորեմ 5.20-ի անհավասարությունը՝ կստանանք

$$E[f(\bar{w}^T)] - \min_w f(w) \leq \frac{4\rho^2}{\lambda T}(1 + \ln T):$$

Եթե որպես կորստի ֆունկցիա վերցնենք $l(w, z_i) = |(w, z_i) - y_i|$, ապա թեորեմի բոլոր պայմանները տեղի կունենան:

Վերջում նշենք, որ *Ridge* ռեգրեսիայի դեպքում օպտիմալ $\hat{w} = (\hat{w}_0, \hat{w})$ վեկտորի կառուցման համար կարելի է կատարել հաստատուն քայլով ստոխաստիկ գրադիենտային իջեցում հետևյալ ընթացակարգով.

- $w_0, w_1, \dots, w_n$ կշիռներին տալիս ենք սկզբնական պատահական արժեքներ:
- t -րդ քայլում հավաքածուից պատահականորեն ընտրում ենք որևէ օբյեկտ և

հաշվում ենք հետևյալ արտահայտությունները`

$$A(w, x_i) \equiv w_0 + w_1 x_{i1} + \dots + w_n x_{in} - y_i,$$

$$B(w, x_i) = (w_0 + w_1 x_{i1} + \dots + w_n x_{in} - y_i) x_{ji};$$

- կատարվում է կշիռների թարմացում հետևյալ բանաձևերով`

$$w_0^t = w_0^{t-1} - hA(w^{t-1}, x_i),$$

$$w_j^t = (1 - h\alpha)w_j^{t-1} - hB(w^{t-1}, x_i) \quad j = 1, 2, \dots, n :$$

5.16 Օպտիմալ պրեդիկտորների գոյության պրոբլեմը ուռուցիկ խնդիրներում

Դիտարկենք գծային ռեգրեսիայի խնդիրը մասնավոր դեպքում, երբ օբյեկտների վեկտորը ունի մեկ չափողականություն`

$$l(w, (x, y)) = (wx - y)^2:$$

Դիցուք ունենք $S = \{z_1, z_2, \dots, z_m\}$ հավաքածու: Նշանակենք`

$$A(S) = \arg \min_w L_S(w), \text{ որտեղ } L_S(w) = \frac{1}{l} \sum_{i=1}^l l(w, z_i):$$

Ենթադրենք` գոյություն ունի օպտիմալ պրեդիկտոր: Դա նշանակում է, որ գոյություն ունեն ալգորիթմ A և ֆունկցիա $m(.,.)$, այնպիսին, որ կամայական D բաշխման դեպքում և ցանկացած $\epsilon, \delta \in (0, 1)$ թվերի համար տեղի ունի հետևյալ պայմանը. եթե A ալգորիթմը մուտքում ստանում է $l \geq m(\epsilon, \delta)$ երկարության S հավաքածու, ապա այն վերադարձնում է $\hat{w} = A(S)$ հիպոթեզ, այնպիսին, որ

$$P(L_D(\hat{w}) - \min_w L_D(w) \leq \epsilon) > 1 - \delta:$$

Դիցուք

$$\epsilon = 1/100, \delta = 1/2, l \geq m(\delta, \epsilon), \theta = \frac{100/99}{2l}:$$

Դիտարկենք երկու բաշխումներ: Առաջին`  $D_1$  բաշխումը կենտրոնացված է առաջին երկու օբյեկտների վրա`

$$z_1 = (1, 0), \quad z_2 = (\theta, -1):$$

Առաջին օբյեկտի հանդես գալու հավանականությունը թող լինի θ , իսկ երկրորդինը` $1 - \theta$: Ենթադրենք նաև, որ երկրորդ դեպքում D_2 բաշխումը կենտրոնացված է երկրորդ z_2 օբյեկտի վրա: Նկատենք, որ երկու բաշխման

դեպքերում էլ հավանականությունը այն բանի, որ l երկարության հավաքածուի մեջ բոլոր օբյեկտները երկրորդ տիպի են՝ մեծ կամ հավասար են 0.99-ից: Դասակնհայտ է D_2 բաշխման դեպքում: Իսկ D_1 բաշխման դեպքում այդ պատահարի հավանականությունը հավասար է

$$(1 - \theta)^l \geq e^{-2l\theta} = 0.99:$$

Ենթադրենք ունենք A ալգորիթմ, որը ստանում է օբյեկտների հավաքածու, որը բաղկացած է l օբիեկտներից, որոնցից յուրաքանչյուրը հավասար է $(\theta, -1)$: Դիցուք ալգորիթմի էլքում ստացվում է $\hat{w}$ հիպոթեզ: Եթե $\hat{w} < -1/(2\theta)$, ապա D_1 բաշխման պարագայում կունենանք

$$L_{D_1}(\hat{w}) \geq \theta(\hat{w})^2 \geq 1/(4\theta):$$

Մյուս կողմից՝

$$\min_w L_{D_1}(w) \leq L_{D_1}(0) = (1 - \theta):$$

Հետևաբար

$$L_{D_1}(\hat{w}) - \min_w L_{D_1}(w) \geq \frac{1}{4\theta} - (1 - \theta) > \epsilon:$$

Այսինքն՝ A ալգորիթմը ձախողվում է D_1 բաշխման դեպքում:

Եթե $\hat{w} \geq -1/(2\hat{w})$ ապա

$$L_{D_2}(\hat{w}) \geq 1/4, \min_w L_{D_2}(w) = 0:$$

Սա նշանակում է, որ A ալգորիթմը ձախողվում է D_2 բաշխման դեպքում: Արդյունքում եզրակացնում ենք, որ քննարկվող դեպքում օպտիմալ արեդիկտոր չունի:

Դիտարկենք նույն խնդիրը, բայց այս անգամ հիպոթեզների բազմության վրա դնենք սահմանափակում: Դիցուք

$$H = \{w : \|w\| \leq 1\}:$$

Դիտարկենք երկու բաշխումներ՝ D_1 և D_2 , որոնք կենտրոնացված են երկու օբյեկտների վրա՝

$$z_1 = (1/\theta, 0), \quad z_2 = (1, -1):$$

Եթե ալգորիթմը ստանալով երկրորդ տիպի l օբյեկտներ, վերադարձնում է

$$\hat{w} < -1/2,$$

ապա D_1 բաշխման դեպքում կունենանք

$$L_{D_1}(\hat{w}) - \min_{w \in H} L_{D_1}(w) \geq$$

$$\theta(\hat{w}/\theta)^2 - L_{D_1}(0) \geq$$

$$1/(4\theta) - (1 - \theta) > \epsilon:$$

Նմանապես, եթե $\hat{w} \geq -1/2$, ապա D_2 բաշխման դեպքում կունենանք

$$L_{D_2}(\hat{w}) - \min_{w \in H} L_{D_2}(w) \geq (-1/2 + 1)^2 - 0 > \epsilon:$$

Քննարկված այս օրինակները ցույց են տալիս, որ կորստի ֆունկցիայի ուռուցիկության և հիպոթեզների բազմության սահմանափակության և ուռուցիկության պայմանները բավարար չեն օպտիմալ հիպոթեզի գոյության համար: Անհրաժեշտ են նաև լրացուցիչ պայմաններ: Դրանք կքննարկվեն հաջորդ բաժիններում:

5.17 Տիխոնովի ռեգուլյարիզացիա: Ալգորիթմների կայունություն և պրեդիկտորի օպտիմալություն

Նախորդ պարագրաֆում տեսանք, որ եթե կորստի ֆունկցիան միայն ուռուցիկ է, իսկ հիպոթեզների բազմությունը ուռուցիկ է և սահմանափակ, ապա նման խնդիրներում ընդհանրապես ասած օպտիմալ պրեդիկտոր գոյություն չունի: Այս բաժնում ցույց է տրվում, որ եթե կորստի ֆունկցիային ավելացնենք Տիխոնովի $\lambda||x||^2$ տեսքի ռեգուլյարիզատորը (կարգավորիչը), ապա ստացված խնդիրը դառնում է կայուն, և օպտիմալ պրեդիկտոր գոյություն կունենա: Նախ տանք կայուն ալգորիթմի գաղափարը և հետո կապացուցենք, որ նման ալգորիթմներով կառուցվում են օպտիմալ պրեդիկտորներ: Այդ գաղափարը քննարկենք ուռուցիկ կորստի ֆունկցիաների դեպքի համար: Դիցուք ունենք որևէ փորձնական հավաքածու՝ $S = \{z_1, z_2, \dots, z_{i-1}, z_i, z_{i+1}, \dots, z_l\} \sim D^l$ և մի լրացուցիչ օրինակ՝ z' : Նշանակենք

$$S_i = \{z_1, \dots, z_{i-1}, z', z_{i+1}, \dots, z_l\} \sim D^l,$$

$$A(S) = \arg \min_w (L_S(w) + \lambda ||w||^2),$$

որտեղ

$$L_S(w) = \frac{1}{l} \sum_{i=1}^l l(w, z_i):$$

Կասենք, որ A ալգորիթմը **կայուն** է, եթե ցանկացած $\epsilon > 0$ թվի համար գոյություն ունի այնպիսի $N(\epsilon)$ համար, որ

$$l(A(S_i), z_i) - l(A(S), z_i) \leq \epsilon, \quad l \geq N(\epsilon), \quad \forall i:$$

Սա նշանակում է, որ հավաքածուի որևէ մի տվյալի փոփոխության դեպքում կորուստը «աննշան» է փոխվում:

Մեր նպատակն է ցույց տալ, որ եթե $l(w, z_i)$ կորստի ֆունկցիան ուռուցիկ է և

լիպշիցյան, կամ այն ոչ բացասական է, ուռուցիկ է և β -ողորկ, ապա A ալգորիթմը կայուն է, իսկ $A(S)$ հիպոթեզը՝ օպտիմալ, երբ հիպոթեզների H բազմությունը լինի սահմանափակ:

Ուռուցիկ ֆունկցիայի բաժնում ցույց է տրված, որ $f_S(w) = L_S(w) + \lambda||w||^2$ ֆունկցիան 2λ -ուժեղ ուռուցիկ է: Հետևաբար՝

$$f_S(v) - f_S(A(S)) \geq \lambda||v - A(S)||^2 \quad \forall v:$$

Մյուս կողմից ցանկացած v, u վեկտորների և i ինդեքսի համար տեղի ունի հետևյալ հավասարությունը՝

$$\begin{aligned} f_S(v) - f_S(u) &= L_S(v) + \lambda||v||^2 - (L_S(u) + \lambda||u||^2) = \\ &= L_{S_i}(v) + \lambda||v||^2 - (L_{S_i}(u) + \lambda||u||^2) + \\ &\quad + \frac{l(v, z_i) - l(u, z_i)}{l} + \frac{l(u, z') - l(v, z')}{l}: \end{aligned}$$

Մասնավորապես, եթե վերցնենք $v = A(S_i)$, $u = A(S)$ և օգտվենք այն փաստից, որ v -ն $L_{S_i}(w) + \lambda||w||^2$ ֆունկցիայի մինիմումի կետն է, կստանանք՝

$$\begin{aligned} f_S(A(S_i)) - f_S(A(S)) &\leq \frac{l(A(S_i), z_i) - l(A(S), z_i)}{l} + \\ &\quad + \frac{l(A(S), z') - l(A(S_i), z')}{l}: \end{aligned}$$

Հետևաբար՝

$$\begin{aligned} \lambda||A(S_i) - A(S)||^2 &\leq \frac{l(A(S_i), z_i) - l(A(S), z_i)}{l} + \\ &\quad + \frac{l(A(S), z') - l(A(S_i), z')}{l}: \end{aligned} \quad (5.43)$$

Այժմ ենթադրենք, որ $l(w, z_i)$ ֆունկցիաները ρ -լիպշիցյան են:

Հետևաբար՝

$$\begin{aligned} l(A(S_i), z_i) - l(A(S), z_i) &\leq \rho||A(S_i) - A(S)||, \\ l(A(S_i), z') - l(A(S), z') &\leq \rho||A(S_i) - A(S)||: \end{aligned}$$

Հաշվի առնելով այս անհավասարությունները վերևի անհավասարությունից՝ կստանանք

$$\lambda||A(S_i) - A(S)||^2 \leq \frac{2\rho||A(S_i) - A(S)||}{l}:$$

Այստեղից՝

$$||A(S_i) - A(S)|| \leq \frac{2\rho}{\lambda l}:$$

Տեղադրելով այս գնահատականը

$$l(A(S_i), z_i) - l(A(S), z_i) \leq \rho||A(S_i) - A(S)||$$

անհավասարության մեջ՝ կստանանք՝

$$l(A(S_i), z_i) - l(A(S), z_i) \leq \frac{2\rho^2}{\lambda l} :$$

Այստեղից հետևում է, որ

$$E_S[L_D(A(S)) - L_S(A(S))] \leq \frac{2\rho^2}{\lambda l} : \quad (5.44)$$

Իրոք, ունենք

$$E_S[L_D(A(S))] = E_{S,z'}[l(A(S), z')] - E_{S,z'}[l(A(S_i), z_i)] :$$

Մյուս կողմից ունենք

$$E_S[L_S(A(S))] = E_{S,i}[l(A(S), z_i)] :$$

Այսպիսով՝

$$\begin{aligned} E_S[L_D(A(S)) - L_S(A(S))] &= \\ &= E_{S,z',i}[l(A(S_i), z_i) - l(A(S), z_i)] : \end{aligned}$$

Ուստի 5.44-ը տեղի ունի: Ուրեմն բավականաչափ մեծ l -երի դեպքում

$$E_{S \sim D'}[L_D(A(S)) - L_S(A(S))]$$

տարբերությունը փոքր է: Այդ դեպքում ասում են՝ A ալգորիթմը **կայուն է միջին** իմաստով:

Այժմ դիտարկենք այն դեպքը, երբ կորստի $l(w, z_i)$ ֆունկցիան ոչ բացասական β -ողորկ ֆունկցիա է: Արդեն գիտենք, որ այդ դասի ֆունկցիաները բավարարում են

$$\|f'(w)\|^2 \leq 2\beta f(w) \quad (5.45)$$

անհավասարությունը: Ենթադրենք, որ $\beta \leq \lambda l/2$: Համաձայն $l(w, z_i)$ ֆունկցիայի β ողորկության՝ ունենք

$$l(A(S_i), z_i) - l(A(S), z_i) \leq (l'(A(S), z_i), A(S_i) - A(S)) + \frac{\beta}{2} \|A(S_i) - A(S)\|^2 :$$

Այստեղից, կիրառելով Կոշի-Բունյակովսկու անհավասարությունը և հաշվի առնելով (5.45)-ը, կստանանք՝

$$\begin{aligned} l(A(S_i), z_i) - l(A(S), z_i) &\leq \|l'(A(S), z_i)\| \|A(S_i) - A(S)\| + (\beta/2) \|A(S_i) - A(S)\|^2 \leq \\ &\sqrt{2\beta l(A(S), z_i)} \|A(S_i) - A(S)\| + (\beta/2) \|A(S_i) - A(S)\|^2 \end{aligned} \quad (5.46) :$$

Անալոգիայով կունենանք

$$\begin{aligned} l(A(S_i), z') - l(A(S), z') &\leq \|l'(A(S), z_i)\| \|A(S_i) - A(S)\| + (\beta/2) \|A(S_i) - A(S)\|^2 \leq \\ &\sqrt{2\beta l(A(S), z')} \|A(S_i) - A(S)\| + (\beta/2) \|A(S_i) - A(S)\|^2 : \end{aligned} \quad (5.47)$$

Տեղադրելով այս անհավասարությունները (5.43) անհավասարության մեջ՝ կստանանք

$$\|A(S) - A(S_i)\| \leq \frac{\sqrt{2\beta}}{\lambda l - \beta} \left(\sqrt{l(A(S), z_i)} + \sqrt{l(A(S_i), z')} \right):$$

Այստեղից՝ քանի որ $\beta \leq \lambda l/2$, ապա

$$\|A(S) - A(S_i)\| \leq \frac{\sqrt{8\beta}}{\lambda l} (\sqrt{l(A(S), z_i)} + \sqrt{l(A(S_i), z')}):$$

Այստեղից կստանանք՝

$$\begin{aligned} l(A(S_i), z_i) - l(A(S), z_i) &\leq \\ \sqrt{2\beta l(A(S), z_i)} \|A(S_i) - A(S)\| + \frac{\beta}{2} \|A(S_i) - A(S)\|^2 &\leq \\ \left(\frac{4\beta}{\lambda l} + \frac{8\beta^2}{(\lambda m)^2} \right) \left(\sqrt{l(A(S), z_i)} + \sqrt{l(A(S_i), z')} \right)^2 & \\ \leq \frac{8\beta}{\lambda l} \left(\sqrt{l(A(S), z_i)} + \sqrt{l(A(S_i), z')} \right)^2 & \\ \leq \frac{24\beta}{\lambda l} (l(A(S), z_i) + l(A(S_i), z')) : & \end{aligned} \quad (5.48)$$

Ունենք նաև

$$L_S(A(S)) \leq L_S(A(S)) + \lambda \|A(S)\|^2 \leq L_S(0) + \lambda \|0\|^2 \leq \max_z l(0, z) = C : \quad (5.49)$$

Գումարելով (5.48) անհավասարության երկու կողմերը $i = 1, 2, \dots, l$ ինդեքսների համար, հաշվի առնելով նաև (5.49) անհավասարությունը՝ միջինների համար կստանանք հետևյալ գնահատականը՝

$$E[l(A(S_i), z_i) - l(A(S), z_i)] \leq \frac{48\beta C}{\lambda l} : \quad (5.50)$$

Այսպիսով ապացուցվեց հետևյալ թեորեմը:

Թեորեմ 5.24. Դիցուք կորստի $l(w, z_i)$ ֆունկցիաները ուռուցիկ են, β -ողորկ u ոչ բացական են: Այդ դեպքում L_2 ռեգուլյարիզացիայով կառուցված $A(S)$ պրեդիկտորը ($\lambda \geq 2\beta/l$) բավարարում է (5.50) անհավասարությունը:

Այս թեորեմից հետևում է, որ A ալգորիթմը **կայուն** է:

Այժմ ցույց տանք, որ վերևում կառուցված $A(S)$ պրեդիկտորը նաև **օպտիմալ** է: Ակնհայտ է, որ

$$E_S[L_D(A(S))] = E_S[A(S)] + E_S[L_D(A(S)) - L_S(A(S))]:$$

Այս հավասարության երկրորդ գումարելին արդեն գնահատել ենք վերևում: Հետևաբար

$$E_S[L_D(A(S))] \leq E_S[A(S)] + \frac{2\rho^2}{\lambda l} :$$

Այժմ գնահատենք էմպիրիկ ռիսկը: Քանի որ $A(S) = \arg \min_w (L_S(w) + \lambda \|w\|^2)$, ապա

$$L_S(A(S)) \leq L_S(A(S)) + \lambda \|A(S)\|^2 \leq$$

$$L_S(w) + \lambda \|w\|^2 \quad \forall w:$$

Վերցնելով այս անհավասարության երկու մասի մաթեմատիկական սպասումը և հաշվի առնելով, որ $E_S[L_S(w)] = L_D(w)$ ՝ կստանանք՝

$$E_S[L_S(A(S))] \leq L_D(w) + \lambda \|w\|^2:$$

Արդյունքում կստանանք՝

$$E_S[L_D(AS)] \leq L_D(w) + \lambda \|w\|^2 + \frac{2\rho^2}{\lambda l}:$$

Դիցուք ենթադրենք, որ $\|w\| \leq B$ և

$$\lambda^* = \arg \min_{\lambda} (\lambda B^2 + \frac{2\rho^2}{\lambda l}):$$

Այդ դեպքում հեշտ է նկատել, որ

$$\lambda^* = \sqrt{\frac{2\rho^2}{B^2 l}}$$

և λ պարամետրի նման ընտրության դեպքում տեղի կունենա հետևյալ անհավասարությունը՝

$$E_S[L_D(AS)] \leq \min_w L_D(w) + \rho B \sqrt{\frac{8}{l}}:$$

Մասնավորապես, եթե $l \geq (8\rho^2 B^2)/\epsilon^2$, ապա կամայական D հավանականային բաշխման դեպքում կունենանք՝

$$E_S[L_D(A(S))] \leq \min_{w \in H} L_D(w) + \epsilon,$$

ինչը նշանակում է $A(S)$ հիպոթեզի օպտիմալություն:

5.18 Ստոխաստիկ գրադիենտային իջեցում (SGD with Momentum)

Դիցուք $\{x_t\}$ հաջորդականությունը կառուցվում է հետևյալ ռեկուրենտ առնչություններով՝

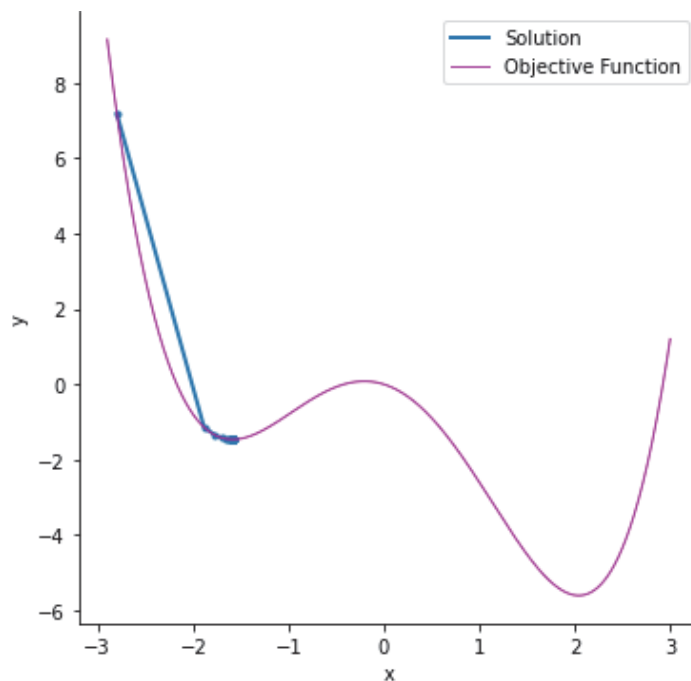
$$m_{t+1} = \mu m_t + f'(x_t),$$

$$x_{t+1} = x_t - \alpha m_{t+1}:$$

Գրականության մեջ սա կոչվում է մոմենտներով ստոխաստիկ գրադիենտային իջեցում: Այստեղ մենք չենք անդրադառնա գուգամիտության հարցերին, այլ քննարկենք թվային օրինակ: Դիցուք $y = 0.3x^4 - 0.1x^3 - 2x^2 - 0.8x$: Կատարենք սովորական գրադիենտային իջեցում $x_0 = -2.8$ սկզբնական կետից՝ քայլի երկարությունը վերցնելով հաստատուն,

$$x_{t+1} = x_t - 0.05f'(x_t):$$

Կատարելով ծրագրային իրականացում *Python*-ով՝ տեսնում ենք, որ պրոցեսը սկսում է կայունանալ $x = -1.6$ կետի շուրջը, որը մոտավոր ճշտությամբ ֆունկցիայի լոկալ մինիմումի կետն է: Այսինքն՝ պրոցեսը կանգնում է լոկալ մինիմումի կետում (տե՛ս նկ.5.6):



Նկ. 5.6: Գրադիենտային իջեցում

	x	y
1	-2.800	7.194
2	-1.885	-1.140
3	-1.766	-1.354
4	-1.702	-1.421
5	-1.663	-1.446
	...	...
99	-1.586	-1.464

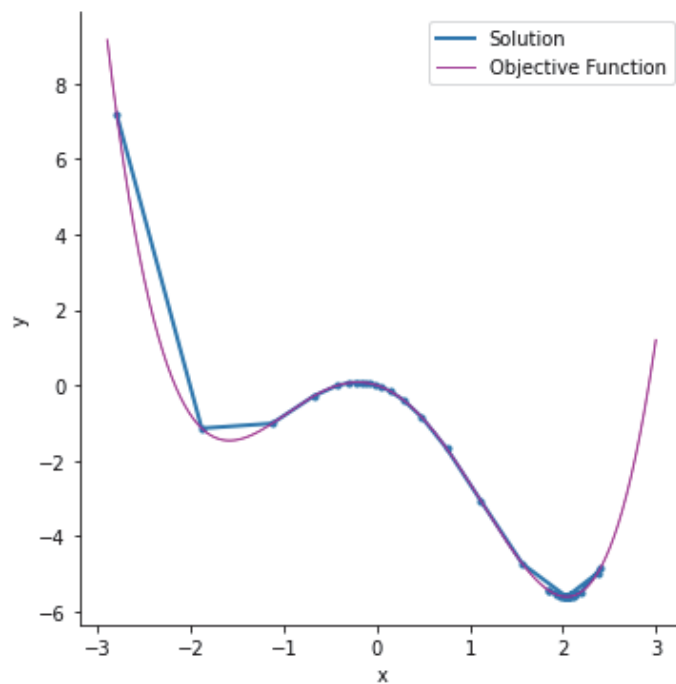
Նկ. 5.7: Տվյալների աղյուսակ

Տվյալները յուրաքանչյուր իտերացիայում տրված է աղյուսակ 5.7-ում:

Այժմ կիրառենք մոմենտներով ստոխաստիկ գրադիենտային իջեցման մեթոդը՝ $\mu = 0.7$ պարամետրով: Այդ դեպքում կստանանք տվյալների նոր աղյուսակ (աղյուսակ 5.8), և որ կարևորն է՝ պրոցեսը կայունանում է ֆունկցիայի գլոբալ մինիմումի կետում (տե՛ս նկ. 5.9):

	x	y
1	-2.800	7.194
2	-1.885	-1.140
3	-1.126	-1.011
4	-0.676	-0.279
	...	...
99	2.042	-5.608

Նկ. 5.8: Նոր տվյալների աղյուսակ



Նկ. 5.9: SGD with Momentum

Առաջադրանք: Դիտարկել 4.1 պարագրաֆում քննարկված գծային ռեգրեսիայի խնդիրը: Խնդիրը լուծելու համար կիրառել հետևյալ ալգորիթմները՝

1. 4.3 բանաձևը,
2. գրադիենտային մեթոդը հաստատուն քայլով,
3. SGD,
4. SGD with momentum,
5. AdaGrad,
6. Nesterov momentum method:

Գրել համապատասխան կոդերը *Python*-ով: Կատարել կանխատեսում և գնահատել յուրաքանչյուր մոդելի որակը: Որպես ալգորիթմի որակի գնահատական՝ վերցնել

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{w_i - \hat{w}_i}{w_i} \right|$$

չափանիշը (Mean Absolute Percentage Error), որտեղ w_i , $i = 1, 2, \dots, n$ թվերը իրական արժեքներն են, իսկ $\hat{w}_i$, $i = 1, 2, \dots, n$ ՝ կանխատեսումները:

5.19 Ֆունկցիայի մինիմիզացիայի Նյուտոն-Ռաֆսոնի մեթոդը

Այս պարագրաֆում քննարկելու ենք ֆունկցիայի մինիմիզացիայի երկրորդ կարգի դասական մեթոդ, որը գրականության մեջ կոչվում է Նյուտոն-Ռաֆսոնի մեթոդ: Այդ մեթոդում ֆունկցիայի նվազման ուղղության ընտրության պրոցեսում օգտագործվում է նաև նրա հեսիանը:

Դիցուք պետք է լուծել հետևյալ խնդիրը՝

$$f(x) \rightarrow \min_{x \in \mathbb{R}^n}:$$

Նախ պետք է գտնել այն կետերը, որոնք բավարարում են մինիմումի անհրաժեշտ պայմանը, այսինքն՝ $f'(x) = 0$ հավասարումը: Այս հավասարումը լուծելու համար կիրառում ենք հաջորդական մոտավորությունների մեթոդը՝ հենվելով գծայնացման մեթոդի վրա: Վերլուծում ենք $f'(x)$ վեկտոր ֆունկցիան թեյլորի բանաձևի x_k կետի շրջակայքում՝

$$f'(x) \approx f'(x_k) + f''(x_k)(x - x_k):$$

Լուծելով այս գծային հավասարումը՝ գտնում ենք հաջորդ մոտավորությունը

$$x_{k+1} = x_k - (f''(x_k))^{-1} f'(x_k):$$

Այս ռեկուրենտ առնչությունը կոչվում է Նյուտոնի մեթոդ: Սակայն պարզվում է, որ այն գուգամիտում է ոչ բոլոր սկզբնական կետերից: Ինչպես նկատելի է, այստեղ ֆունկցիայի նվազման ուղղությունը ընտրվել է ոչ թե հակադրադիենտը, այլ $p_k = -((f''(x_k))^{-1}f'(x_k))$ վեկտորը, իսկ քայլի երկարությունը՝ $\alpha_k = 1$ բոլոր պահերին: Այժմ դիտարկենք նման ռեկուրենտ առնչություն, սակայն քայլերը ընտրենք կիսման եղանակով՝

$$x_{k+1} = x_k + \alpha_k p_k,$$

որտեղ α_k -ն $1, 1/2, 1/4, \dots$ հաջորդականության այն առաջին անդամն է, որ բավարարում է

$$f(x_k + \alpha_k p_k) \leq f(x_k) + \epsilon \alpha_k (f'(x_k), p_k) \quad (5.51)$$

անհավասարությունը: Այստեղ $\epsilon \in (0, 1/2)$ -ը մեթոդի պարամետրն է: Այս եղանակը կոչվում է քայլի կարգավորմամբ Նյուտոն-Ռաֆսոնի մեթոդ: Այս մեթոդի գուգամիտության հարցը քննարկենք հետևյալ լրացուցիչ պայմանով: Դիցուք f ֆունկցիան բավարարում է

$$m||h||^2 \leq (f''(x)h, h) \leq M||h||^2 \quad \forall x, h, \in \mathbb{R}^n \quad (5.52)$$

անհավասարությունը:

Լեմմա 5.5. Եթե $f \in C^2(\mathbb{R}^n)$, և տեղի ունի (5.52) անհավասարությունը, ապա ցանկացած $x \in \mathbb{R}^n$ -ի համար գոյություն ունի $(f''(x))^{-1}$ հակադարձ մատրիցը, և տեղի ունի

$$\frac{1}{M}||h||^2 \leq ((f''(x))^{-1}h, h) \leq \frac{1}{m}||h||^2 \quad (5.53)$$

անհավասարությունը, ընդ որում՝

$$||f''(x)|| \leq M, \quad ||(f''(x))^{-1}|| \leq \frac{1}{m} \quad \forall x \in \mathbb{R}^n: \quad (5.54)$$

Ապացույց: Դիցուք $A = f''(x)$: Քանի որ այդ մատրիցը սիմետրիկ է, ապա գոյություն ունեն $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$ սեփական արժեքներին համապատասխան $h_1, h_2, \dots, h_n$ օրթոնորմավորված բազիս, ընդ որում՝

$$\lambda_1 = \min_{||h||=1} (Ah, h), \quad \lambda_n = \max_{||h||=1} (Ah, h):$$

Համաձայն (5.52) անհավասարության՝

$$0 < m \leq \lambda_1, \quad \lambda_n \leq M:$$

Ըստ սեփական արժեքի սահմանման՝ ունենք նաև

$$Ah_k = \lambda_k h_k, \quad k = 1, 2, \dots, n: \quad (5.55)$$

Դիցուք P -ն այնպիսի մատրից է, որի սյուները $h_1, \dots, h_n$ վեկտորներն են, իսկ Λ -ն

անկյունագծային մատրից է, որի գլխավոր անկյունագծում սեփական արժեքներն են: Այդ դեպքում (5.55) հավասարությունը կրնալուի հետևյալ տեսքը՝

$$AP = P\Lambda: \quad (5.56)$$

Քանի որ P մատրիցը օրթոնորմավորված է, ապա $P^T P = E$: Այստեղից՝

$$P = (P^T)^{-1}: \quad (5.57)$$

Հաշվի առնելով սա, և աջից բազմապատկելով (5.56)-ը հավասարության երկու կողմերը P^T -ով՝ կստանանք

$$A = P\Lambda P^T: \quad (5.58)$$

Այստեղից՝

$$A^{-1} = (P\Lambda P^T)^{-1} = P^{-1} \Lambda^{-1} P^T,$$

որտեղ Λ^{-1} -ն անկյունագծային է, որի գլխավոր անկյունագծում դասավորված են

$$\frac{1}{\lambda_1}, \dots, \frac{1}{\lambda_n}$$

թվերը: Այսինքն՝ A^{-1} մատրիցի սեփական արժեքներն են այդ թվերը: Հետևաբար՝

$$\frac{1}{\lambda_n} = \min_{\|h\|=1} (A^{-1}h, h), \quad \frac{1}{\lambda_1} = \max_{\|h\|=1} (A^{-1}h, h):$$

Մնաց ցույց տալու (5.54) անհավասարությունները: Ունենք

$$\|Ah\|^2 = (Ah, Ah) = (A^2h, h): \quad (5.59)$$

Նկատի ունենալով (5.57)-(5.58) հավասարությունները՝ կունենանք

$$A^2 = (P\Lambda)P^T(P\Lambda)P^T = P\Lambda^2P^T:$$

Ուստի A^2 մատրիցի սեփական արժեքներն են

$$\lambda_1^2 \leq \dots \leq \lambda_n^2$$

թվերը: Ուստի

$$\max_{\|h\|=1} (A^2h, h) = \lambda_n^2 \leq M^2: \quad (5.60)$$

Այստեղից, հաշվի առնելով նաև (5.59)-ը, կստանանք՝

$$\|Ah\|^2 \leq M^2\|h\|^2:$$

Այսինքն՝

$$\|A\| \leq M:$$

Նման ձևով հաշվի առնելով (5.53) անհավասարությունը՝ կստանանք՝

$$\|A^{-1}\| \leq \frac{1}{m}:$$

Թեորեմ 5.25. Դիցուք f ֆունկցիան բավարարում է լեմմա 5.5-ի պայմանները, իսկ $\{x_k\}$ հաջորդականությունը գեներացվում է Նյուտոն-Ռաֆսոնի մեթոդով: Այդ դեպքում այդ հաջորդականությունը գերզոնալիս արագությամբ ձգտում է այդ ֆունկցիայի միակ մինիմումի կետին: Այսինքն՝ գոյություն ունի այնպիսի $q_k \rightarrow 0$ հաջորդականություն, որ

$$\|x_{k+1} - x^*\| \leq q_k \|x_k - x^*\|,$$

որտեղ x^* -ը f -ի մինիմումի կետն է:

Ապացույց: Կատարենք նշանակումներ՝

$$f'_k = f'(x_k), \quad f''_k = f''(x_k):$$

Այդ դեպքում, հաշվի առնելով (5.53) անհավասարությունը, կստանանք՝

$$\begin{aligned} (f'_k, p_k) &= (f''_k (f''(x_k))^{-1} f'_k, p_k) = \\ &= -(f''_k p_k, p_k) \leq -m \|p_k\|^2: \end{aligned}$$

Այստեղից հետևում է, որ

$$\begin{aligned} \|p_k\|^2 &\leq -\frac{1}{m} (f'_k, p_k), \\ \|p_k\| &\leq \frac{1}{m} \|f'_k\|: \end{aligned} \tag{5.61}$$

Կիրառենք միջին արժեքի վերաբերյալ թեորեմը: Նկատի ունենալով նաև (5.54) և (5.61) անհավասարությունները՝ կստանանք

$$\begin{aligned} f(x_k + \alpha p_k) &= f(x_k) + \alpha (f'_k, p_k) + \frac{1}{2} \alpha^2 (f''(x_k + \theta_k \alpha p_k) p_k, p_k) \leq \\ &\leq f(x_k) + \alpha (f'_k, p_k) (1 - \frac{\alpha M}{2m}): \end{aligned} \tag{5.62}$$

Այստեղից, եթե

$$0 < \alpha \leq \frac{2m(1-\epsilon)}{M}, \tag{5.63}$$

ապա տեղի կունենա

$$f(x_k + \alpha p_k) \leq f(x_k) + \epsilon \alpha (f'_k, p_k) \tag{5.64}$$

անհավասարությունը: Այժմ ցույց տանք, որ

$$\alpha_k \geq \frac{m(1-\epsilon)}{M}: \tag{5.65}$$

Եթե $\frac{2m(1-\epsilon)}{M} \geq 1$, ապա (5.64) անհավասարությունը տեղի կունենա $\alpha = 1$ արժեքի դեպքում: Հետևաբար՝ այս դեպքում $\alpha_k = 1$:

Եթե $\frac{2m(1-\epsilon)}{M1} < 1$, ապա գոյություն կունենա այնպիսի բնական s թիվ, որ

$$\frac{1}{2^s} \leq \frac{2m(1-\epsilon)}{M} \leq \frac{1}{2^{s-1}}:$$

Այս պարագայում $\alpha = \frac{1}{2^s}$ արժեքի դեպքում տեղի կունենա (5.64) անհավասարությունը: Ըստ սահմանման $\alpha_k \geq \frac{1}{2^s}$, և քանի որ

$$\frac{1}{2^s} = \frac{1}{2^{2^{s-1}}} > \frac{m(1-\epsilon)}{M},$$

ապա ճշմարիտ է (5.65) անհավասարությունը: (5.64) անհավասարությունը գրենք հետևյալ տեսքով՝

$$\begin{aligned} f(x_k) - f(x_{k+1}) &\geq \epsilon \alpha_k (-(f'(x_k), p_k)) \geq \\ &\geq \epsilon \frac{m(1-\epsilon)}{M} (f_k'')^{-1} f_k', f_k' \geq \\ &\geq \frac{\epsilon(1-\epsilon)m}{M^2} \|f_k'\|^2: \end{aligned} \quad (5.66)$$

Քանի որ ըստ կառուցման $\{f(x_k)\}$ հաջորդականությունը մոնոտոն նվազող է և ներքևից սահմանափակ (քանի որ ֆունկցիան ունի միակ գլոբալ մինիմումի կետ (տե՛ս [50], թեորեմ 1.1.7)), ապա (5.66)-ից հետևում է, որ $\|f_k'\| \rightarrow 0$: Այստեղից, հաշվի առնելով $\|p_k\| \leq \frac{1}{m} \|f_k'\|$ անհավասարությունը, ստանում ենք, որ $p_k \rightarrow 0$: Պարագրաֆ 5.3-ում ապացուցվել է

$$\|x - x^*\| \leq \frac{1}{m} \|f'(x)\|$$

անհավասարությունը (անհավասարություն 5.8): Հետևաբար՝

$$\|x_k - x^*\| \leq \frac{1}{m} \|f_k'\| \rightarrow 0:$$

Այսպիսով ապացուցվեց, որ $x_k \rightarrow x^*$:

Մնաց ցույց տալու զուգամիտության արագության գնահատականը: Նախ ցույց տանք, որ ինչ-որ համարից սկսած՝ $\alpha_k = 1$: Ունենք

$$f(x_k + p_k) = f(x_k) + (f_k', p_k) + \frac{1}{2} (f_k'' p_k, p_k) + \frac{1}{2} ((f_{kc}'' - f_k'') p_k, p_k),$$

որտեղ

$$f_{kc}'' = f''(x_k + \theta_k p_k), \quad \theta_k \in (0, 1):$$

Քանի որ

$$p_k = -(f_k'')^{-1} f_k',$$

ապա

$$(f_{kc}'' p_k, p_k) = -(f_k', p_k): \quad (5.67)$$

Բացի դրանից՝ համաձայն (5.61)-ի ունենք

$$((f''_{kc} - f''_k)p_k, p_k) \leq \|f''_{kc} - f''_k\| \|p_k\|^2 \leq -\frac{1}{m} \|f''_{kc} - f''_k\| (f'_k, p_k): \quad (5.68)$$

Հաշվի առնելով (5.67)-(5.68) անհավասարությունները՝ կստանանք

$$f(x_k + p_k) \leq f(x_k) + (f'_k, p_k) \left(\frac{1}{2} - \frac{\|f''_{kc} - f''_k\|}{2m} \right): \quad (5.69)$$

Քանի որ

$$x_k \rightarrow x^*, \quad x_{kc} \rightarrow x^*,$$

ապա $\|f''_{kc} - f''_k\| \rightarrow 0$: Հետևաբար՝ ինչ-որ համարից սկսած՝ տեղի կունենա

$$\frac{1}{2} - \frac{\|f''_{kc} - f''_k\|}{2m} \leq \epsilon$$

անհավասարությունը, որովհետև ըստ ենթադրության, $\epsilon \in (0, 1/2$: Այսպիսով, այդ համարից սկսած, (5.69) անհավասարությունը կընդունի հետևյալ տեսքը՝

$$f(x_k + p_k) \leq f(x_k) + \epsilon (f'_k, p_k):$$

Այսինքն՝ գոյություն ունի այնպիսի N համար, որ $\alpha_k = 1$, $k \geq N$:

Դիցուք $k \geq N$: Այդ դեպքում

$$\begin{aligned} \|x_{k+1} - x^*\|^2 &= (x_k - x^* - f''(x_k)^{-1} f'_k, x_{k+1} - x^*) = \\ &= (x_k - x^* - (f''_k)^{-1} (f'_k - f'(x^*)), x_{k+1} - x^*) = \\ &= (f''(x_k)^{-1} (f''_k - f''_{k^*}) (x_k - x^*), x_{k+1} - x^*) \leq \\ &\leq \|(f''_k)^{-1}\| \|f''_k - f''_{k^*}\| \|x_{k+1} - x^*\| \|x_k - x^*\| \end{aligned}$$

որտեղ $f''_{k^*} = f''(x_* + \theta_k(x_k - x^*))$: Նշանակենք

$$q_k = \|(f''_k)^{-1}\| \cdot \|f''_k - f''_{k^*}\|:$$

Քանի որ $\|(f''(x_k))^{-1}\| \leq \frac{1}{m}$ (լեմմա 5.5 (5.64) անհավասարություն), ապա $q_k \rightarrow 0$: Թեորեմն ապացուցվեց:

Հետևանք: Եթե լրացուցիչ պահանջենք, որ

$$\|f''(x) - f''(y)\| \leq L \|x - y\|, \quad x, y \in \mathbb{R}^n,$$

ապա մեծ k համարների դեպքում տեղի կունենա

$$\|x_{k+1} - x^*\| \leq \frac{L}{M} \|x_k - x^*\|^2$$

անհավասարությունը: Այս դեպքում ասում են, որ գուգամիտությունը քառակուսային է: Այսպիսով, Նյուտոն-Ռաֆսոնի մեթոդն ավելի արագ է գուգամիտում,

քան սովորական գրադիենտային իջեցումը քայլի ընտրության նույն եղանակի դեպքում:

5.20 Մինիմիզացիայի երկրորդ կարգի մեթոդները Էքսպոնենցիալ (exp-concave) Ֆունկցիաների դասի համար

Նախորդ պարագրաֆում քննարկեցինք Նյուտոն-Ռաֆսոնի մեթոդը ուժեղ ուռուցիկ ֆունկցիաների համար: Այժմ դիտարկենք այն դեպքը, երբ ֆունկցիաները երկու անգամ դիֆերենցել են, բայց ուժեղ ուռուցիկ չեն:

Սահմանում: f ուռուցիկ ֆունկցիան կոչվում է α - Էքսպոնենցիալ գոգավոր $M \subset \mathbb{R}^n$ ուռուցիկ բազմության վրա, եթե

$$g(x) = e^{-\alpha f(x)}$$

ֆունկցիան գոգավոր է այդ բազմության վրա:

Տեղի ունի հետևյալ պնդումը:

Լեմմա 5.6. Որպեսզի f ֆունկցիան լինի α - գոգավոր M -ի վրա, անհրաժեշտ է և բավարար, որ

$$f''(x) - \alpha f'(x)(f'(x))^T$$

մատրիցը լինի ոչ բացասական որոշյալ ցանկացած $x \in M$ -ի դեպքում:

Ծշմարիտ է նաև հետևյալ պնդումը:

Լեմմա 5.7. Դիցուք f -ը α -Էքսպոնենցիալ գոգավոր ֆունկցիա է M ուռուցիկ կոմպակտ բազմության վրա: Այդ դեպքում տեղի ունի հետևյալ անհավասարությունը՝

$$f(x) \geq f(y) + (f'(y), x - y) + \frac{\gamma}{2}(f'(y)(f'(y))^T(x - y), x - y) \quad \forall x, y \in M,$$

որտեղ

$$\gamma \leq \frac{1}{2} \min(1/(4GD), \alpha), \quad D = \text{diam} M, \quad G \geq \|f'(x)\|, \quad x \in M:$$

Հայտնի է α -գոգավոր ֆունկցիաների մինիմիզացիայի հետևյալ ալգորիթմը.

- դիցուք $x_1 \in M$, իսկ $\gamma, \epsilon > 0$ թվերը պարամետրեր են, $A_0 = \epsilon I_n$,
- $k = 1, 2, \dots, T$ արժեքների համար կատարել
- $A_k = A_{k-1} + f'(x_k)(f'(x_k))^T$,
- $y_{k+1} = x_k - \frac{1}{\gamma} A_k^{-1} f'(x_k)$, $x_{k+1} = \Pi_M^{A_k}(y_{k+1})$:

Թեորեմ 5.26. Դիցուք $\{x_k\}$ հաջորդականությունը գեներացվում է վերը նշված ալգորիթմով

$$\gamma = \frac{1}{2} \min\left(\frac{1}{4GD}, \alpha\right) \epsilon = \frac{1}{\gamma^2 D^2}$$

սարամերերի դեպքում:

Այդ պարագայում տեղի ունի

$$f(\bar{x}_T) - \min_{x \in M} f(x) \leq 5\left(\frac{1}{\alpha} + DG\right)n \frac{\ln T}{T}, \quad T > 4$$

անհավասարությունը, որտեղ

$$\bar{x}_T = \frac{x_1 + x_2 + \dots + x_t}{T}:$$

Ապացույց: Դիցուք

$$x^* \in \arg \min_{x \in M} f(x):$$

Հանաձայն լեմմա 5.7-ի՝ ունենք

$$f(x_k) - f(x^*) \leq R_k,$$

որտեղ

$$R_k = (f'(x_k), x - x^*) - \frac{\gamma}{2} (f'(x_k) ((f'(x_k))^T (x^* - x_k), x^* - x_k)):$$

Համաձայն y_{k+1} -ի սահմանման՝ ունենք նաև

$$y_{k+1} - x^* = x_k - x^* - \frac{\gamma}{2} A_k^{-1} f'(x_k), \quad (5.70)$$

$$A_k(y_{k+1} - x^*) = A_k(x_k - x^*) - \frac{\gamma}{2} f'(x_k): \quad (5.71)$$

Բազմատկելով այս հավասարությունների աջ ու ձախ մասերը սկալյար արտադրյալի իմաստով՝ կստանանք

$$\begin{aligned} & (A_k(y_{k+1} - x^*), y_{k+1} - x^*) = \\ & = (A_k(x_k - x^*), x_k - x^*) - \frac{2}{\gamma} (f'(x_k), x_k - x^*) + \frac{1}{\gamma^2} (A_k^{-1} f'(x_k), f'(x_k)): \end{aligned} \quad (5.72)$$

Մյուս կողմից, ըստ պրոյեկցիայի սահմանման, ունենք նաև

$$\begin{aligned} & (A_k(y_{k+1} - x^*), y_{k+1} - x^*) = \|y_{k+1} - x^*\|_{A_k}^2 \geq \\ & \|x_{k+1} - x^*\|_{A_k}^2 = (A_k(x_{k+1} - x^*), x_{k+1} - x^*): \end{aligned}$$

Հաշվի առնելով այս անհավասարությունը, (5.72)-ից՝ կստանանք՝

$$(f'(x_k), x_k - x^*) \leq \frac{1}{2\gamma} (A_k^{-1} f'(x_k), f'(x_k)) + \frac{\gamma}{2} (A_k(x_k - x^*), x_k - x^*) -$$

$$-\frac{\gamma}{2}(A_k(x_{k+1} - x^*), x_{k+1} - x^*):$$

Գումարելով այս անհավասարությունները $k = 1, 2, \dots, T$ ինդեքսների համար՝ կստանանք

$$\begin{aligned} \sum_{k=1}^T (f'(x_k), x_k - x^*) &\leq \frac{1}{2\gamma} \sum_{k=1}^T (A_k^{-1} f'(x_k), f'(x_k)) + \frac{\gamma}{2} (A_1(x_1 - x^*), x_1 - x^*) + \\ &+ \frac{\gamma}{2} \sum_{k=2}^T (A_k - A_{k-1})(x_k - x^*), x_k - x^*) - \\ &- \frac{\gamma}{2} A_T(x_{T+1} - x^*), x_{T+1} - x^*) \leq \\ &\leq \frac{1}{2\gamma} \sum_{k=1}^T (A_k^{-1} f'(x_k), f'(x_k)) + \frac{\gamma}{2} \sum_{k=1}^T (f'(x_k)(f'(x_k))^T(x_k - x^*), x_k - x^*) + \\ &+ \frac{\gamma}{2} (A_1 - f'(x_1)f'(x_1)^T(x_1 - x^*), x_1 - x^*) \leq \\ &\frac{1}{2\gamma} \sum_{k=1}^T (A_k^{-1} f'(x_k), f'(x_k)) + \frac{\gamma}{2} D^2 \epsilon \leq \\ &\frac{1}{2\gamma} \sum_{k=1}^T (A_k^{-1} f'(x_k), f'(x_k)) + \frac{1}{2\gamma} \leq \\ &\frac{1}{2\gamma} \sum_{k=1}^T (A_k^{-1} f'(x_k), f'(x_k)) + 4\left(\frac{1}{\alpha} + GD\right): \end{aligned}$$

Այսպիսով ստացանք, որ

$$\sum_{k=1}^T (f(x_k) - f(x_*)) \leq \frac{1}{2\gamma} \sum_{k=1}^T (A_k^{-1} f'(x_k), f'(x_k)) + 4\left(\frac{1}{\alpha} + GD\right):$$

Այստեղից, կիրառելով Յենսենի անհավասարությունը ուռուցիկ f ֆունկցիայի համար, կստանանք

$$T(f(\bar{x}_T) - f(x^*)) \leq \sum_{k=1}^T (f(x_k) - f(x_*)) \leq \frac{1}{2\gamma} \sum_{k=1}^T (A_k^{-1} f'(x_k), f'(x_k)) + 4\left(\frac{1}{\alpha} + GD\right)$$

անհավասարությունը: Մնաց ցույց տալու, որ

$$\sum_{k=1}^T (A_k^{-1} f'(x_k), f'(x_k)) \leq n \ln T:$$

Ունենք

$$(A_k)^{-1} f'(x_k), f'(x_k)) = A_k^{-1} \bullet f'(x_k) f'(x_k)^T = A_k^{-1} \bullet (A_k - A_{k-1}),$$

որտեղ

$$A \dot{\bullet} B = \sum_{i=1}^n \sum_{j=1}^n a_i j b_i j = \text{Tr}(AB^T):$$

Հայտնի է ([6], լեմմա 4.5) նաև, որ

$$A^{-1} \dot{\bullet} (A - B) \leq \ln \frac{|A|}{|B|},$$

որտեղ $|A|$, $|B|$ թվերը մատրիցների դետերմինանտներն են: Այստեղից՝

$$\begin{aligned} \sum_{k=1}^T (A_k^{-1} f'(x_k), f'(x_k)) &= \sum_{k=1}^T A_k^{-1} \dot{\bullet} (A_k - A_{k-1}) = \\ &\leq \sum_{k=1}^T \ln \frac{|A_k|}{|A_{k-1}|} = \ln \frac{|A_T|}{|A_0|}: \end{aligned}$$

Քանի որ

$$A_T = \sum_{k=1}^T f'(x_k) f'(x_k)^T + \epsilon I_n,$$

ապա

$$|A_T| \leq (TG^2 + \epsilon)^n:$$

Հետևաբար՝

$$\sum_{k=1}^n (A_k^{-1} f'(x_k), f'(x_k)) \leq \ln \left(\frac{TG^2 + \epsilon}{\epsilon} \right)^n \leq n \ln(TG^2 \gamma^2 D^2 + 1) \leq n \ln T, \quad n \geq 1, \quad T > 4:$$

Թեորեմն ապացուցվեց:

Դիտողություն: Թեորեմ 5.16-ում ցույց է տրված, որ ուժեղ ուռուցիկ ֆունկցիաների համար գրադիենտային իջեցման մեթոդը ապահովում է

$$f(\bar{x}_T) - \min_{x \in M} f(x) = O\left(\frac{\ln T}{T}\right)$$

գնահատականը: Թեորեմ 5.24 ում գույքամիտության արագության նման գնահատական ապացուցվում է էքսպոնենցիալ գոգավոր ֆունկցիաների դասի համար:

ԽՆԴԻՐՆԵՐ

1) Ապացուցել լեմմաներ 5.6-5.7 պնդումները:

2) Ապացուցել, որ

$$|A_T| \leq (TG^2 + \epsilon)^2:$$

3) Ապացուցել, որ հետևյալ ֆունկցիաները էքսպոնենցիալ գոգավոր են՝

ա) $f(x) = ((x, w) - y)^2, x \in \mathbb{R}^n,$

բ) $f(x) = \ln(1 + e^{y(w, x)}), y \in \{-1, +1\},$

գ) $f(x) = -\ln(w, x):$

4) Ապացուցել, որ ուժեղ ուռուցիկ ֆունկցիան էքսպոնենցիալ գոգավոր է:

5.21 Նյուտոն-Փաուսի մեթոդը ոչ գծային ռեգրեսիայում

Դիցուք ունենք ուսուցանող $Z = \{(x_1, y_1), \dots, (x_l, y_l)\}$ հավաքածու, որտեղ $x_i \in \mathbb{R}^n, y_i \in R$: Դիտարկենք ռեգրեսիայի հետևյալ մոդելը՝

$$Q(Z, w) = \sum_{i=1}^l (f(x_i, w) - y_i)^2 \rightarrow \min_w,$$

որտեղ $f(x, w)$ -ն $w \in \mathbb{R}^p$ պարամետրից կախված ոչ գծային ֆունկցիա է: Այս կորստի ֆունկցիայի մինիմիզացիայի համար կիրառենք գրականության մեջ լավ հայտնի Նյուտոն-Ռաֆսոնի մեթոդը՝

$$w_{k+1} = w_k - \alpha_k (Q''(w_k))^{-1} Q'(w_k),$$

որտեղ $Q'(w_k)$ -ն Q -ի գրադիենտն է w_k կետում, իսկ $Q''(w_k)$ -ն հեսիանն է նշված կետում: Հայտնի է, որ այս մեթոդը ունի գույքամիտության քառակուսային արագություն: Սակայն մեծ չափերի դեպքում հաշվողական բարդություններ են առաջանում հեսիանի հակադարձ մատրիցը որոշելու պրոցեսում: Այս դժվարությունը կարելի է հաղթահարել հետևյալ կերպ: Ունենք

$$\frac{\partial Q(w)}{\partial w_j} = 2 \sum_{i=1}^l (f(x_i, w) - y_i) \frac{\partial f(x_i, w)}{\partial w_j}:$$

Այստեղից հեսիանի կոմպոնենտները կունենան հետևյալ տեսքը՝

$$\frac{\partial^2 Q(w)}{\partial w_j \partial w_k} = 2 \sum_{i=1}^l \frac{\partial f(x_i, w)}{\partial w_j} \frac{\partial f(x_i, w)}{\partial w_k} + 2 \sum_{i=1}^l (f(x_i, w) - y_i) \frac{\partial^2 f(x_i, w)}{\partial w_j \partial w_k}:$$

Եթե երկրորդ կարգի ածանցյալները մոդուլով բավականաչափ փոքր են, ապա դեն նետելով երկրորդ գումարելին՝ կստանանք հեսիանի կոմպոնենտների մոտավոր հետևյալ ներկայացումները՝

$$\frac{\partial^2 Q(w)}{\partial w_j \partial w_k} \cong 2 \sum_{i=1}^l \frac{\partial f(x_i, w)}{\partial w_j} \frac{\partial f(x_i, w)}{\partial w_k}:$$

Նշանակենք

$$F_k = \left(\frac{\partial f}{\partial w_j}(x_i, w_p) \right)_{l \times p} \quad f_k = (f(x_i, w_k))_{l \times 1} :$$

Հաշվի առնելով այս նշանակումները և տեղադրելով հեսիանի մոտավոր արժեքը Նյոտոն-Ռաֆսոնի բանաձևի մեջ՝ կստանանք հետևյալ ռեկուրենտ առնչությունը՝

$$w_{k+1} = w_k - \alpha_k (F_k^T F_k)^{-1} F_k^T (f_k - y) :$$

Այս առնչությունը կոչվում է ֆունկցիայի մինիմիզացիայի Նյոտոն-Գաուսի մեթոդ։ Հեշտ է նկատել, որ $\beta = (F_k^T F_k)^{-1} F_k^T (f_k - y)$ արտահայտությունը գծային ռեգրեսիայի հետևյալ խնդրի լուծումն է՝

$$\|F_k \beta - (f_k - y)\|^2 \rightarrow \min_{\beta} :$$

Այսպիսով ոչ գծային ռեգրեսիայի խնդիրը բերվում է գծային ռեգրեսիայի խնդրի։ Այս մեթոդի զուգամիտության արագությունը նույն է ինչ Նյոտոն-Ռաֆսոնինը։

5.22 Գրադիենտային իջեցման մեթոդում քայլի ընտրության Պոյսկի եղանակը

Դիցուք դիտարկում ենք

$$f(x) \rightarrow \min, \quad x \in R^n$$

խնդիրը։ Ենթադրենք հայտնի է այդ ֆունկցիայի մինիմալ արժեքը՝

$$f(x^*) = \min_{x \in R^n} f(x) \equiv f^* :$$

Օրինակ՝ ենթադրենք ուզում ենք լուծել գծային հավասարումների համատեղելի

$$(a_i, x) = b_i, \quad i = 1, 2, \dots, n$$

համակարգը։ Ակնհայտ է, որ այս խնդիրը համարժեք է

$$f(x) = \sum_{i=1}^n ((a_i, x) - b_i)^2 \rightarrow \min_x$$

խնդրին, որտեղ $f^* = 0$ ։

Պահանջվում է կառուցել $\{x_k\}$ հաջորդականություն, այնպիսին, որ $f(x_k) \rightarrow f(x^*)$ և գնահատել զուգամիտության արագությունը։ Դիցուք կատարում ենք գրադիենտային իջեցում հետևյալ եղանակով՝

$$x_{k+1} = x_k - \eta_k f'(x_k),$$

որտեղ

$$\eta_k = \begin{cases} \frac{f(x_k) - f^*}{\|f'(x_k)\|}, & \text{եթե; } f'(x_k) \neq 0, \\ 1, & \text{հակառակ դեպքում:} \end{cases}$$

Տեղի ունի հետևյալ թեորեմը:

Թեորեմ 5.27. Դիցուք f -ը ուռուցիկ դիֆերենցելի ֆունկցիա է, և նրա գրադիենտը սահմանափակ է՝ $\|f'(x)\| \leq G \quad \forall x \in R^n$: Դիցուք $\{x_k\}$ հաջորդականությունը գեներացվում է վերը նշված ռեկուրենտ առնչությամբ: Այդ դեպքում՝

$$f(\tilde{x}_k) - f^* = O\left(\frac{1}{\sqrt{k}}\right),$$

որտեղ

$$\tilde{x}_k = \arg \min_{i \in [1:k]} f(x_i):$$

Ապացույց: Ունենք

$$\begin{aligned} \|x_{j+1} - x^*\|^2 &\leq \|x_j - x^*\|^2 - \frac{(f(x_j) - f^*)^2}{\|f'(x_j)\|^2} \leq \\ &\|x_j - x^*\|^2 - \frac{(f(x_j) - f^*)^2}{G^2}: \end{aligned}$$

Այստեղից հետևում է, որ

$$\sum_{j=1}^k (f(x_j) - f^*)^2 \leq G^2(\|x_1 - x^*\|^2 - \|x_{k+1} - x^*\|^2):$$

Հետևաբար՝

$$\sum_{j=1}^k (f(x_j) - f^*)^2 \leq G^2\|x_1 - x^*\|^2:$$

Այստեղից

$$f(\tilde{x}_k) - f^* \leq \frac{G\|x_1 - x^*\|}{\sqrt{k}}:$$

Թեորեմն ապացուցվեց:

Երբ f ֆունկցիան α -ուժեղ ուռուցիկ է կամ β -ողորկ, ապա այս եղանակի զուգամիտության արագության տարբեր գնահատականների հետ կարելի է ծանոթանալ [7] աշխատանքում:

Առաջադրանք: Փորձարարական կենսաբանության մեջ, ուսումնասիրելով սուրստրատի կոնցենտրացիայի՝ (S)-ի և ռեակցիայի արագության փոխհարաբերությունները ֆերմենտային մոդուլյացիայի ռեակցիայում, ստացվել են հետևյալ տվյալները (տե՛ս 5.10): Ենթադրենք արագությունը արտահայտվում է հետևյալ բանաձևով՝

$$f(S, w) = \frac{w_1 S}{w_2 + S}:$$

Նյուտոն-Գաուսի մեթոդով գտնել $w = (w_1, w_2)$ պարամետրերի արժեքները և

հարթության վրա պատկերել տվյալները ու արագության կորը:

i	1	2	3	4	5	6	7
[S]	0.038	0.194	0.425	0.626	1.253	2.500	3.740
արագություն	0.050	0.127	0.094	0.2122	0.2729	0.2665	0.3317

Նկ. 5.10: Ոչ գծային ռեգրեսիա

6

Հենման վեկտորների մեթոդը (support vector machine- SVM)

Այս մեթոդը բինար դասակարգման հիմնական մեթոդներից է: Դրա օգնությամբ կատարվում է օբյեկտների դասակարգում երկու դասերից որևէ մեկում: Դիտարկենք պրակտիկ մի օրինակ: Ենթադրենք ունենք մի հարցաթերթիկ (անկետա) բաղկացած չորս հարցերից՝

- սիրելի ըմպելիքը՝ թեյ կամ սուրճ ,
- սիրելի կենդանին՝ շուն կամ կատու,
- լավատեսության աստիճանը,
- հոռետեսության աստիճանը:

Դիցուք կատարվել է հարցում «արտույտ» և «բու» մարդկանց շրջանում և ստացվել են տվյալներ, որոնք ներկայացված են աղյուսակ 6.1-ում:

Աղյուսակ 6.1, Հարցման թերթիկ				
Թեյ, Սուրճ	Կատու, Շուն	Լավատես	Հոռետես	Արտույտ
սուրճ	կատու	3	8	նչ (-1)
սուրճ	կատու	5	5	նչ (-1)
թեյ	կատու	6	4	այո (+1)
թեյ	շուն	8.	3	այո (+1)
սուրճ	շուն	7	3	այո(+1)
թեյ	շուն	6	3	այո (+1)
թեյ	շուն	4	4	նչ (-1)

Ենթադրենք՝ կատարվել է հարցում նոր մարդու, և նրա պատասխանը եղել է հետևյալը՝ (թեյ, շուն, 3, 8): Մենք ուզում ենք կանխատեսել, նա «արտույտ» է, թե՞ «բու»:

Դիցուք տրված է նմուշների $S = \{(x_i, y_i), i = 1, 2, \dots, l\}$ բազմություն, որտեղ $y_i \in \{1, -1\}, i = 1, 2, \dots, l$: Նշանակենք՝

$$S^+ = \{x/(x, 1) \in S\}, S^- = \{x/(x, -1) \in S\}$$

Գծային դասակարգման (կլասիֆիկացիայի) խնդիրն է կառուցել մի հիպերհարթություն, որը կանջատի S^+ և S^- բազմությունները իրարից: Սա համարժեք է հետևյալին. գոյություն ունեն $w_0, w_1, \dots, w_n$ թվեր, այնպիսին, որ

$$\forall((x^1, \dots, x^n), y) \in S, \text{sign}\left(\sum_{i=1}^n x^i w_i - w_0\right) = y:$$

Այսինքն՝

$$\forall((x^1, \dots, x^n), y) \in S \left(\sum_{i=1}^n x^i w_i - w_0\right)y \geq 0: \quad (6.1)$$

S^+ և S^- բազմությունները կոչվում են «խիստ» անջատող, եթե (6.1)-ում տեղի ունենա խիստ անհավասարություն:

Թեորեմ 6.1. Որպեսզի S^+ S^- բազմությունները լինեն «խիստ» անջատվող, անհրաժեշտ է և բավարար, որ

$$\text{conv}(S^+) \cap \text{conv}(S^-) = \emptyset:$$

Ապացույց: Դիցուք S նմուշների հավաքածուն գծորեն խիստ անջատելի է: Դա նշանակում է, որ գոյություն ունի հիպերհարթություն, որը անջատում է S^+ և S^- բազմությունները: Ենթատարածությունները նշանակենք Z^+ և Z^- , իսկ նրանց ներքին մասերը՝ համապատասխանաբար՝ $\text{int}Z^+, \text{int}Z^-$: S^+ և S^- բազմությունների «խիստ» անջատում նշանակում է՝

$$S^+ \subseteq \text{int}Z^+, S^- \subseteq \text{int}Z^-:$$

Այստեղից՝

$$\text{conv}(S^+) \subseteq \text{int}Z^+, \text{conv}(S^-) \subseteq \text{int}Z^-:$$

Հետևաբար՝

$$\text{conv}(S^+) \cap \text{conv}(S^-) = \emptyset: \quad (6.2)$$

Դիցուք տեղի ունի (6.2) պայմանը: Քանի որ $\text{conv}(S^+)$ և $\text{conv}(S^-)$ բազմությունները կոմպակտ են, ապա նրանց հեռավորությունը դրական թիվ է՝

$$\rho(\text{conv}(S^+), \text{conv}(S^-)) = \rho(x^+, x^-):$$

x^+ x^- կետերից տանենք ուղղահայաց հիպերհարթություններ $[x^+, x^-]$ հատվածին:

Այդ հիպերհարթությունները նշանակենք համապատասխանաբար P_{S^+} - ով և P_{S^-} -ով, իսկ կիսատարածությունները՝ Z^+ , Z^- : Հեշտ է նկատել, որ

$$Z_{S^+} = \{x/x = x^+ \text{ կամ } \angle xx^+x^- \geq \pi/2\}:$$

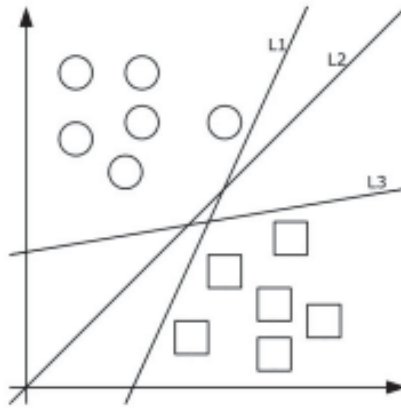
Ցույց տանք, որ $S^+ \subseteq Z_{S^+}$, այսինքն՝ $\forall x \in S^+ \angle xx^+x^- \geq \pi/2$: Իրոք, եթե գոյություն ունի այնպիսի $x \in S^+$ կետ, որ $\angle xx^+x^- < \pi/2$, ապա $[x, x^+]$ հատվածի վրա կգտնվի x' կետ, այնպիսին, որ $\rho(x', x^-) < \rho(x^+, x^-)$: Այդ դեպքում $[x', x^+] \subseteq \text{conv}(S^+)$:

Հետևաբար՝

$$\rho(x^+, x^+) < \rho(\text{conv}(S^+), \text{conv}(S^-)),$$

ինչը հակասություն է:

Այսպիսով, $[x^+, x^-]$ հատվածի մեջտեղով անցնող և նրան ուղղահայց հիպերհարթությունը «խիստ» կանջատի S^+ և S^- բազմությունները: Նկար 6.1-ում պատկերված են երկու բազմությունները անջատող հիպերհարթություններ:



Նկ. 6.1: Անջատող հիպերհարթություններ

6.1 Օպտիմալ անջատող հիպերհարթության կառուցման ալգորիթմը

Նախորդ պարագրաֆի քննարկումներից ակնհայտ դարձավ, որ խիստ անջատող հիպերհարթությունները կարող են միակը չլինեն (տես, օրինակ՝ նկ. 6.1): Հետևաբար հարց է առաջանում, թե այդ հիպերհարթություններից, որը համարել օպտիմալ, ինչը վերցնել օպտիմալության չափանիշ: Որպես օպտիմալության չափանիշ այստեղ վերցվում է հետևյալ թիվը՝

$$\rho(S^+ \cup S^-, P),$$

որտեղ P -ն անջատող հիպերհարթությունն է, իսկ

$$\rho(A, B) = \inf_{a \in A, b \in B} \|a - b\|:$$

Դիցուք P_{S^+} -ը այն հիպերհարթությունն է, որը ստացվել է P անջատող հիպերհարթությունը S^+ բազմության ուղղությամբ իր նորմալով՝ $\rho(P, S^+)$ չափով գուգահեռ տեղաշարժով:

Նման կերպ՝ P_{S^-} -ը այն հիպերհարթությունն է, որը ստացվել է P անջատող հիպերհարթությունը S^- բազմության ուղղությամբ իր նորմալով $\rho(P, S^-)$ չափով գուգահեռ տեղաշարժելով: $\rho(P_{S^+}, P_{S^-})$ թիվը կոչվում է $[P_{S^+}, P_{S^-}]$ շերտի հաստություն: Ակնհատ է, որ

$$\rho(S^+ \bigcup S^-, P)$$

արտահայտությունը կրնդունի մաքսիմալ արժեք այն դեպքում, երբ

$$\rho(P_{S^+}, P_{S^-}) = \rho(S^+, S^-)$$

և P հիպերհարթությունը անցնի այդ շերտի մեջտեղով: Այդպիսի հիպերհարթությունը անվանում ենք օպտիմալ:

Տեղի ունի հետևյալ պնդումը:

Թեորեմ 6.2. Դիցուք S^+ և S^- բազմությունները խիստ անջատվում են: Այդ դեպքում օպտիմալ անջատող հիպերհարթություն գոյություն ունի և միակն է:

Ապացույց: Դիցուք S հավաքածուն խիստ անջատելի է: Այսինքն՝ գոյություն ունի հիպերհարթություն $(\bar{w}, x) - c = 0$, այնպիսին, որ

$$(\bar{w}, x_i) - c > 0 \text{ } y_i = 1, \text{ } (\bar{w}, x_i) - c < 0, \text{ } y_i = -1:$$

Նշանակենք՝

$$c_1(\bar{w}) = \min_{i: y_i=1} (\bar{w}, x_i),$$

$$c_2(\bar{w}) = \max_{i: y_i=-1} (\bar{w}, x_i):$$

Ունենք

$$c_1(\bar{w}) > c > c_2(\bar{w}):$$

Նշանակենք

$$\begin{aligned} \rho(\bar{w}) &= \frac{c_1(\bar{w}) - c_2(\bar{w})}{2} = \\ &= \frac{1}{2}((c_1(\bar{w}) - c) + (c - c_2(\bar{w}))) : \end{aligned}$$

Դժվար չէ նկատել, որ այս թիվը հավասար է

$$\frac{1}{2}(\rho(P, S^+) + \rho(P, S^-)):$$

Քանի որ $\rho(\bar{w})$ ֆունկցիան անընդհատ է, ապա այն հասնում է իր մեծագույն արժեքին միավոր գնդի վրա: Նախ նկատենք, որ այդ մաքսիմումը հասանելի է գնդի եզրի վրա: Իրոք, եթե $\|\bar{w}_0\| < 1$, և այդ կետի վրա ρ ֆունկցիան հասնում է մաքսիմումի, ապա

$$\rho\left(\frac{\bar{w}_0}{\|\bar{w}_0\|}\right) = \frac{\rho(\bar{w}_0)}{\|\bar{w}_0\|} > \rho(\bar{w}_0),$$

որը հակասաությունն է:

Այժմ ցույց տանք, որ մաքսիմումը հասանելի է միայն մեկ կետում: Սկզբից համոզվենք, որ $\rho(\bar{w})$ ֆունկցիան գոգավոր է: Տեղի ունեն հետևյալ ակնհայտ անհավասարությունները՝

$$\begin{aligned}\min_{i \in I} (f(i) + g(i)) &\geq \min_{i \in I} f(i) + \min_{i \in I} g(i), \\ \max_{i \in I} (f(i) + g(i)) &\leq \max_{i \in I} f(i) + \max_{i \in I} g(i): \end{aligned}$$

Նշանակելով

$$f(i) = (\bar{w}, x_i), \quad g(i) = (\bar{u}, x_i)$$

կունենանք՝

$$\begin{aligned}\min_{y_i=1} (\lambda \bar{w} + (1 - \lambda) \bar{u}, x_i) &= \\ = \min_{y_i=1} (\lambda (\bar{w}, x_i) + (1 - \lambda) (\bar{u}, x_i)) &\geq \lambda \min_{y_i=1} (\bar{w}, x_i) + (1 - \lambda) \min_{y_i=1} (\bar{u}, x_i): \end{aligned}$$

Նման անհավասարություն տեղի ունի մաքսիմումի համար, և իրարից հանելով ստացված անհավասարությունները՝ կստանանք ρ ֆունկցիայի գոգավորությունը: Եթե ρ ֆունկցիան հասնի մաքսիմումի երկու տարբեր կետերում, ապա այդ կետերը միացնող հատվածի ներքին կետը կլինի այդ գոգավոր ֆունկցիայի մաքսիմումի կետ, որը հնարավոր չէ:

Այժմ նկարագրենք օպտիմալ հիպերհարթության կառուցման ալգորիթը: Ըստ կառուցման՝ P_{S^+} , P_{S^-} հիպերհարթությունները գուգահեռ են, և հետևաբար դրանք տարբերվում են միայն ազատ անդամով՝

$$(x, v) - a = 0, \quad (x, v) - b = 0:$$

Եթե այս հավասարումները բազմապատկենք $\lambda = \frac{2}{a-b}$ թվով, ապա նրանք կընդունեն հետևյալ տեսքը՝

$$(x, w) - w_0 = 1, \quad (x, w) - w_0 = -1,$$

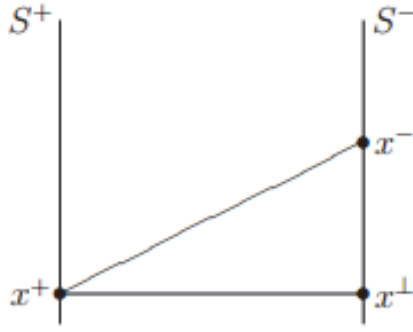
որտեղ

$$w = \lambda v, \quad w_0 = \frac{a+b}{a-b}:$$

Հետևաբար՝ S^+ և S^- բազմությունները գտնվում են համապատասխանաբար

$$H_{S^+} \equiv \{w : (x, w) - w_0 \geq 1\}, \quad H_{S^-} \equiv \{x : (x, w) - w_0 \leq -1\}$$

կիսատարածություններում: P_{S^+} և P_{S^-} հարթությունների վրա վերցնենք համապատասխանաբար x^+ և x^- կետեր: Դիցուք $x^\perp$ -ը x^+ -ի պրոյեկցիան է P_{S^-} հարթության վրա: Հաշվենք P_{S^+}, P_{S^-} շերտի ρ հաստությունը: Այդ մեծությունը հավասար է $x^+, x^-, x^\perp$ գագաթներով ուղղանկյուն եռանկյան $[x^+, x^\perp]$ էջին:



Նկ. 6.2: Օպտիմալ հիպերհարթության կառուցումը

Նկար 6.2-ից երևում է, որ

$$\cos \varphi = \frac{(x^+ - x^-, x^+ - x^\perp)}{\|x^+ - x^-\| \|x^+ - x^\perp\|} :$$

Այստեղից կստանանք

$$\rho = \frac{(x^+ - x^-, x^+ - x^\perp)}{\|x^+ - x^\perp\|} :$$

Քանի որ $x^+ - x^\perp$ վեկտորը գուգահեռ է w վեկտորին, ապա $x^+ - x^\perp = \eta w$: Հետևաբար՝

$$\rho = \frac{(x^+ - x^-, \eta w)}{\|\eta w\|} = \sigma \frac{(x^+ - x^-, w)}{\|w\|},$$

որտեղ $\sigma = +1$, կամ $\sigma = -1$: Քանի որ $x^+ \in P_{S^+}$, $x^- \in P_{S^-}$, ապա

$$(x^+, w) - w_0 = 1, \quad (x^-, w) - w_0 = -1 :$$

Այստեղից կստանանք՝

$$(x^+ - x^-, w) = 2 :$$

Ուստի

$$\rho = \frac{2}{\|w\|} :$$

Վերջնականում օպտիմալ հիպերհարթություն գտնելու խնդիրը կարելի է ձևակերպել հետևյալ կերպ՝

$$\frac{\|w\|^2}{2} \rightarrow \min, \quad y_x((x, w) - w_0) - 1 \geq 0, \quad x \in X_S = \{x \in \mathbb{R}^n : (x, y_x) \in S\} :$$

Այս խնդրի՝ Լագրանժի ֆունկցիան ունի հետևյալ տեսքը՝

$$L = \frac{\|w\|^2}{2} - \sum_{x \in X_S} \lambda_x (y_x((w, x) - w_0) - 1) :$$

Գրենք Կուն-Թակերի օպտիմալության պայմանները և պասիվ-ակտիվ սահմանափակումների պայմանները: Կունենանք՝

- $L'_w = \hat{w} - \sum_{x \in X_S} \hat{\lambda}_x y_x x = 0,$
- $L'_{w_0} = \sum_{x \in X_S} \hat{\lambda}_x y_x = 0,$
- $\hat{\lambda}_x (y_x((x, \hat{w}) - \hat{w}_0) - 1) = 0, \quad x \in X_S,$
- $y_x((x, \hat{w}) - \hat{w}_0) - 1 \geq 0:$

Այժմ ցույց տանք, որ $\hat{\lambda}_x$ պարամետրերի որոշման համար պետք է լուծել քառակուսային ծրագրավորման խնդիր, որի լուծման վերջավոր ալգորիթմը տրված է ([43], էջ 35) աշխատանքում:

Դրա համար

$$\begin{aligned} \hat{w} &= \sum_{x \in X_S} \hat{\lambda}_x y_x x, \\ \sum_{x \in X_S} \hat{\lambda}_x y_x &= 0 \end{aligned}$$

քանաձևերը տեղադրելով L Լագրանժյանի մեջ կստանանք՝

$$\begin{aligned} L &= \frac{1}{2} \|\hat{w}\|^2 - \sum_{x \in X_S} \hat{\lambda}_x (\hat{y}_x((x, \hat{w}) - \hat{w}_0) - 1) = \\ &= -\frac{1}{2} \left(\sum_{x \in X_S} \hat{\lambda}_x y_x x, \hat{w} \right) + \left(\sum_{x \in X_S} \hat{\lambda}_x y_x \right) \hat{w}_0 + \sum_{x \in X_S} \hat{\lambda}_x = \\ &= -\frac{1}{2} \sum_{x, x' \in X_S} \hat{\lambda}_x \hat{\lambda}_{x'} y_x y_{x'}(x, x') + \sum_{x \in X_S} \hat{\lambda}_x : \end{aligned}$$

Այսպիսով, համաձայն ուռուցիկ ծրագրավորման երկակիության տեսության, կարելի է ձևակերպել հետևյալ թեորեմը:

Թեորեմ 6.3. Լագրանժի $\hat{\lambda}_x, \quad x \in X_S$ անորոշ գործակիցները հետևյալ խնդրի լուծումներն են՝

$$\begin{aligned} &= -\frac{1}{2} \sum_{x, x' \in X_S} \lambda_x \lambda_{x'} y_x y_{x'}(x, x') + \sum_{x \in X_S} \lambda_x \rightarrow \max, \\ &\lambda_x \geq 0, \quad \sum_{x \in X_S} \lambda_x y_x = 0, \quad x \in X_S: \end{aligned}$$

Լուծելով այս քառակուսային ծրագրավորման խնդիրը՝ գտնում ենք $\hat{\lambda}_x$ պարամետրերը: Այնուհետև որոնվում է այնպիսի x օբյեկտ, որ $\hat{\lambda}_x > 0$: Պասիվ-ակտիվ $\hat{\lambda}_x (y_x((x, \hat{w}) - \hat{w}_0) - 1) = 0$ համապատասխան պայմանից ստանում ենք ազատ $\hat{w}_0$ հաստատունի արժեքը՝

$$\hat{w}_0 = (x, \hat{w}) - y_x:$$

Օպտիմալ անջատող հիպերհարթության հավասարումը կունենա հետևյալ վերջնական տեսքը՝

$$(x, \hat{w}) - \hat{w}_0 = 0:$$

Այժմ ցույց տանք, որ գոյություն ունի այնպիսի x օբյեկտ, որ $\hat{\lambda}_x > 0$: Եթե բոլոր $\hat{\lambda}_x = 0$, ապա $\hat{w} = 0$, և հետևաբար՝

$$y_x((x, \hat{w}) - \hat{w}_0) - 1 \geq 0, \forall x:$$

Այստեղից կստանանք

$$-y_x \hat{w}_0 \geq 1:$$

Քանի որ ենթադրվում է, որ ունենք ոչ տրիվիալ դեպք, այսինքն՝ կա առնվազն երկու օբյեկտ $+1$ և -1 պատասխաններով, կստանանք

$$\hat{w}_0 \geq 1, \hat{w}_0 \leq -1,$$

որը հնարավոր չէ:

Այժմ դիտարկենք անջատող օպտիմալ հիպերհարթության կառուցման մի օրինակ: Դիցուք ունենք հետևյալ հավաքածուն՝

$$x_1 = (1, 1), y_1 = 1, x_2 = (1, 2), y_2 = -1, x_3 = (2, 3), y_3 = -1:$$

Կազմենք

$$\begin{aligned} F(\lambda) &= \sum_{i=1}^3 \lambda_i - \frac{1}{2} \sum_{i=1, j=1}^3 \lambda_i \lambda_j y_i y_j (x_i, x_j) = \\ &= \lambda_1 + \lambda_2 + \lambda_3 - \frac{1}{2} (2\lambda_1^2 + 5\lambda_2^2 + \\ &+ 13\lambda_3^2 - 6\lambda_1 \lambda_2 - 10\lambda_1 \lambda_3 + 16\lambda_2 \lambda_3) \end{aligned}$$

Ֆունկցիան և այն մաքսիմիզացնենք

$$\lambda_1 - \lambda_2 - \lambda_3 = 0, \lambda_1 \geq 0, \lambda_2 \geq 0, \lambda_3 \geq 0$$

պայմաններով: $\lambda_3 = \lambda_1 - \lambda_2$ արժեք տեղադրելով ֆունկցիոնալի մեջ՝ կստանանք

$$F(\lambda_1, \lambda_2) = 2\lambda_1 - \frac{5}{2}\lambda_1^2 - \lambda_2^2 + 3\lambda_1 \lambda_2:$$

Հաշվելով այս ֆունկցիայի մասնական ածանցյալները և հավասարեցնելով դրանք զրոյի՝ կստանանք

$$\lambda_1 = 4, \lambda_2 = 6,$$

և հետևաբար՝ $\lambda_3 = -2$: Սա նշանակում է, որ F ֆունկցիան մաքսիմալ արժեք ընդունում է տիրույթի եզրի վրա: Ուրեմն հետազոտենք այն

$$\lambda_i \geq 0, i = 1, 2, 3, \lambda_3 = \lambda_1 - \lambda_2$$

բազմության վրա:

- Եթե $\lambda_1 = 0$, ապա $\lambda_3 = -\lambda_2$: Ուստի $\lambda_i = 0, i = 1, 2, 3, F(\lambda) = 0$:

- եթե $\lambda_2 = 0$, ապա $\lambda_1 = \lambda_3 = \lambda$: Այստեղից կստանանք

$$F(\lambda) = 2\lambda - \frac{5}{2}\lambda^2:$$

Ածանցելով այն և հավասարեցնելով զրոյի՝ կստանանք $\lambda = 2/5$, $F(\lambda) = 2/5$:

- Եթե $\lambda_3 = 0$, ապա

$$\lambda_1 = \lambda_2 = \lambda, F(\lambda) = 2\lambda - \frac{1}{2}\lambda^2:$$

Ուրեմն, $\lambda = 2$ -ը F ֆունկցիայի մաքսիմումի կետն է, իսկ նրա մաքսիմալ արժեքը հավասար է 2-ի:

Այսպիսով, ունենք $(2, 2, 0)$ վեկտորը: Ուստի՝

$$\hat{w} = \sum_{i=1}^3 \lambda_i y_i x_i = 2x_1 - 2x_2 = (0, -2),$$

$$\hat{w}_0 = y_1 - (\hat{w}, x_1) = 3:$$

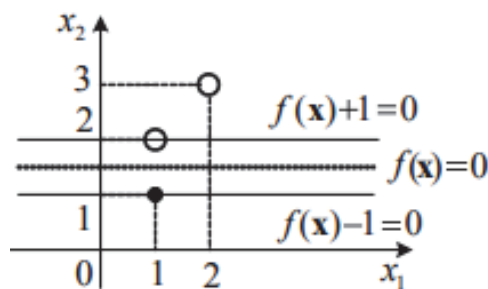
Օպտիմալ անջատող հիպերհարթության հավասարումը կլինի՝

$$f(x_1, x_2) = -2x_2 + 3 = 0,$$

իսկ շերտի հաստությունը՝

$$h = \frac{2}{\|w\|} = 1 \quad (\text{տե՛ս նկ. 6.3}):$$

Այս նկարում $f(x) + 1 = 0$, $f(x) - 1 = 0$ ուղիղները մաքսիմալ շերտի եզրագծերն են: Այժմ ցույց տանք, որ դիտարկած քառակուսային ծրագրավորման խնդիրը կարելի



Նկ. 6.3: Օպտիմալ հիպերհարթության կառուցման օրինակ

է բերել գծային խնդրի: Ունենք հետևյալ խնդիրը՝

$$f(\lambda) = \frac{1}{2}(A\lambda, \lambda) - (b, \lambda) \rightarrow \min,$$

$$\sum_{i=1}^l y_i \lambda_i = (y, \lambda) = 0, \quad \lambda \geq 0:$$

Գրենք Կուն-Թակերի պայմանները այս դեպքի համար.

- $\frac{\partial L(\lambda, z)}{\partial \lambda} \geq 0 \equiv \lambda - b + zy \geq 0,$
- $\lambda_i(A\lambda - b + zy)_i = 0, \quad i = 1, 2, \dots, l,$
- $\lambda_i \geq 0, \quad i = 1, 2, \dots, l:$

Այս պայմանները կարող ենք ներկայացնել նաև հետևյալ տեսքով`

- $A\lambda + zy - v = b,$
- $(b, \lambda) = 0,$
- $\lambda \geq 0, \quad v \geq 0,$
- $\lambda_i v_i = 0, \quad i = 1, 2, \dots, l:$

Փաստորեն խնդիրն այն է, որ պետք է գտնել λ, v, z պարամետրերը այնպես, որ բավարարվեն վերը նշված պայմանները: Հայտնի է Ֆրանկ-Վուլֆի ալգորիթմը, հիմնված` սիմպլեքս ալգորիթմի վրա, որը վերջավոր քայլերում լուծում է այս խնդիրը:

Սակայն նշենք, որ այս ձեռնարկում շարադրված գծայնացման ալգորիթմը պարզագույն մոտավոր ալգորիթմ է օպտիմալ հիպերհարթության կառուցման համար:

6.2 Ռոգենբլատի ալգորիթմը

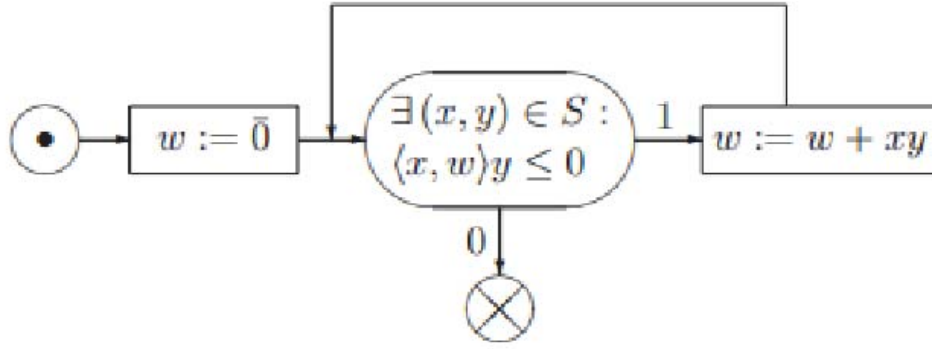
Արդեն գիտենք, որ խիստ անջատման պայմանը հետևյալն է` գոյություն ունի այնպիսի $\mathbf{w}^* = (w_1^*, \dots, w_n^*, w_0^*)$ վեկտոր, որ

$$(\mathbf{w}^*, \mathbf{x})y > 0, \quad \mathbf{x} = (x, -1), \quad (x, y) \in S, \quad y \in \{-1, +1\}:$$

Նկարագրենք այդպիսի $\mathbf{w}$ վեկտորի կառուցման վերջավոր ալգորիթմ, որը կրում է Ռոգենբլատի անունը:

- $\mathbf{w} = 0:$
- Եթե գոյություն ունի այնպիսի $(x, y) \in S$, որ $(\mathbf{w}, \mathbf{x})y \leq 0$, ապա անցնել երրորդ քայլին, հակառակ դեպքում անցնել չորրորդ քայլին:
- $\mathbf{w} = \mathbf{w} + \mathbf{x}y$, և վերադառնալ երկրորդ քայլին:
- Վերջ:

Նկար 6.4-ում պատկերված է Ռոգենբլատի ալգորիթմի սխեման:



Նկ. 6.4: Ռոգենբլատի ալգորիթմը

Ցույց տանք, որ այս ալգորիթմը վերջավոր քայլերի ընթացքում կառուցում է անհրաժեշտ $\mathbf{w}$ վեկտորը: Ունենք $\mathbf{w}^k = \mathbf{w}^{k-1} + \mathbf{x}y$: Հետևաբար՝

$$(\mathbf{w}^k, \mathbf{w}^*) = (\mathbf{w}^{k-1} + \mathbf{x}y, \mathbf{w}^*) = (\mathbf{w}^{k-1}, \mathbf{w}^*) + (\mathbf{x}, \mathbf{w}^*)y \geq (\mathbf{w}^{k-1}, \mathbf{w}^*) + \|\mathbf{w}^*\|\rho,$$

որտեղ

$$\rho = \frac{1}{\|\mathbf{w}^*\|} \min_{(x,y) \in S} (\mathbf{x}, \mathbf{w}^*)y: \quad (6.3)$$

Կիրառելով k անգամ այս անհավասարությունը և հաշվի առնելով, որ $\mathbf{w}^0 = \mathbf{0}$ ՝ կստանանք՝

$$(\mathbf{w}^k, \mathbf{w}^*) \geq k\|\mathbf{w}^*\|\rho: \quad (6.4)$$

Ունենք նաև Կոշի-Բունյակովսկու անհավասարությունը՝

$$(\mathbf{w}^k, \mathbf{w}^*) \leq \|\mathbf{w}^k\| \|\mathbf{w}^*\|: \quad (6.5)$$

Այստեղից, հաշվի առնելով նաև (6.4) անհավասարությունը, կստանանք

$$k^2 \rho^2 \leq \|\mathbf{w}^k\|^2: \quad (6.6)$$

Մյուս կողմից

$$\begin{aligned} \|\mathbf{w}^k\|^2 &= \|\mathbf{w}^{k-1}\|^2 + 2(\mathbf{x}, \mathbf{w}^{k-1})y + \|\mathbf{x}\|^2 \leq \\ &\leq \|\mathbf{w}^{k-1}\|^2 + \sigma^2, \end{aligned}$$

որտեղ $\sigma = \max_{(x,y) \in S} \|\mathbf{x}\|$: Կիրառելով այս անհավասարությունը k անգամ՝ կստանանք

$$\|\mathbf{w}^k\|^2 \leq k\sigma^2:$$

Այստեղից և (6.6) անհավասարությունից հետևում է, որ

$$k \leq \frac{\sigma^2}{\rho^2}:$$

Մեկնաբանենք այս ալգորիթմը օրինակի վրա: Դիցուք

$$x_1 = (1, 1), y_1 = 1; x_2 = (1, -1), y_2 = 1, x_3 = (-1, -1), y_3 = -1, x_4 = ((-1, 1), y_4 = -1:$$

- $\mathbf{w}^0 = 0$,
- $\mathbf{w}^1 = \mathbf{w}^0 + \mathbf{x}_1 y_1 = \mathbf{x}_1$,
- $\mathbf{w}^2 = \mathbf{w}^1 + \mathbf{x}_4 y_4 = \mathbf{x}_1 - \mathbf{x}_4 = (2, 0, 0)$,
- քանի որ $(\mathbf{w}^2, \mathbf{x}_i) y_i > 0$, $i = 1, 2, 3, 4$, ապա $2x_1 + 0x_2 - 0 = 0$ գիծը խիստ անջատում է օբյեկտները իրարից:

6.3 Հենման վեկտորների մեթոդը և ստոխաստիկ գրադիենտներ

Հենման վեկտորների մեթոդը կարելի է ձևակերպել որպես մինիմիզացիայի խնդիր.

$$\frac{1}{l} \sum_{i=1}^l L_i(\mathbf{w}) = \frac{1}{l} \sum_{i=1}^l [m_i < 0] \equiv \frac{1}{l} \sum_{i=1}^l [y_i((w, x_i) - w_0) < 0] \rightarrow \min_{\mathbf{w}},$$

որտեղ $\mathbf{w} \equiv (w, w_0)$, իսկ

$$[m_i < 0] = \begin{cases} 1 & \text{երբ } m_i < 0 \\ 0 & \text{երբ } m_i \geq 0 : \end{cases}$$

Այստեղ մինիմիզացվող ֆունկցիայի արժեքը տրված w , w_0 պարամետրերի դեպքում հավասար է սխալ դասակարգված օբյեկտների քանակի միջին արժեքին: Դետք է գտնել այդ պարամետրերի այնպիսի արժեքներ, որ այդ մեծությունը լինի մինիմալ:

Սակայն $[m_i < 0]$ ֆունկցիաները դիֆերենցելի չեն, և հետևաբար գրադիենտային մեթոդները այստեղ կիրառելի չեն: Սովորաբար այս դեպքերում $[m_i < 0]$ ֆունկցիաները վերևից մոտարկվում են դիֆերենցելի կամ կտոր առ կտոր դիֆերենցելի ֆունկցիաներով: Օրինակ՝ հետևյալ կերպ՝

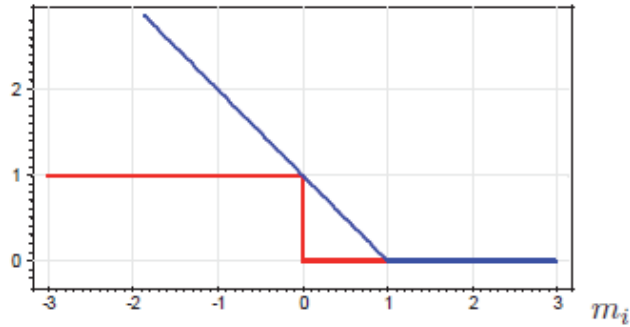
$$[m_i < 0] \leq \tilde{L}_i = (1 - m_i)_+,$$

ինչպես ցույց է տրված նկ.6.5-ում:

Երբ նմուշների քանակը՝ l -ը մեծ է, ապա մինիմիզացիայի համար կիրառվում է ստոխաստիկ գրադիենտային մեթոդը՝

$$\mathbf{w}^{k+1} = \mathbf{w}^k - \alpha_k \mathbf{g}_i^k,$$

որտեղ i -ն որևէ պատահական ինդեքս է, α_k -ն քայլի երկարությունն է, իսկ $\mathbf{g}_i^k \in \partial \tilde{L}_i(\mathbf{w}^k)$:



Նկ. 6.5: $[m_i < 0]$ ֆունկցիայի մոտարկում վերևից $\tilde{L}_i = (1 - m_i)_+$ ֆունկցիայով

6.4 Օպտիմալ բաժանող հիպերհարթության կառուցումը «աղմկոտ» նմուշներից բաղկացած հավաքածուի համար

Ենթադրենք՝ մեր հավաքածուն գծորեն բաժանելի չէ, և SVM ալգորիթմը սխալվում է որոշ օբյեկտների վրա: Բայց աժմ պետք է այնպես անել, որ այդ սխալանքները լինեն մինիմալ: Հիշենք, որ օպտիմալ հիպերհարթության կառուցման համար պետք է լուծել քառակուսային ծրագրավորման հետևյալ խնդիրը՝

$$\frac{1}{2}(w, w) \rightarrow \min, \quad y_x((w, x) - w_0) \geq 1, \quad x \in X_S:$$

Յուրաքանչյուր x օբյեկտի հետ կապենք մի ξ_x փոփոխական, որն ալգորիթմի սխալանքի չափն է այդ օբյեկտի վրա: Այդ սխալանքը գնահատվում է $1 - m_x$ տարբերությամբ, որտեղ $m_x = y_x((x, w) - w_0)$:

Իսկ օպտիմալ բաժանող հիպերհարթության կառուցման խնդիրը այս դեպքում ձևակերպվում է հետևյալ կերպ՝

$$\begin{cases} \frac{1}{2}||w||^2 + C \sum_{i=1}^l \xi_i \rightarrow \min_{w, w_0, \xi}, \\ y_i((w, x_i) - w_0) \geq 1 - \xi_i, \quad i = 1, 2, \dots, l, \\ \xi_i \geq 0, \quad i = 1, 2, \dots, l: \end{cases}$$

Գրականության մեջ այս խնդիրը կոչվում է *Soft - SVM* ոչ համասեռ կիսատարածությունների դեպքում ($w_0 \neq 0$):

Նպատակային ֆունկցիան արտահայտում է հետևյալ պահանջները՝

- անջատող շերտի հաստությունը պետք է լինի մաքսիմալ.

$$\frac{2}{||w||^2} \rightarrow \max,$$

այսինքն՝

$$\frac{1}{2}||w||^2 \rightarrow \min:$$

- Գումարային սխալները պետք լինի մինիմալ.

$$\sum_{i=1}^l \xi_i \rightarrow \min:$$

C հաստատունը կարգավում է բալանսը այս երկու պահանջների մեջ: Դա կատարվում է հետևյալ կերպ: Սկզբից խնդիրը լուծվում է C -ի որևէ արժեքի դեպքում: Հավաքածուից հեռացվում են բոլոր այն օբյեկտները, որոնց դեպքում ալգորիթմը շատ է սխալվել: Իսկ պարամետրը փոխում են հետևյալ սկզբունքով.

- եթե շերտի հաստությունը փոքր է, ապա այդ պարամետրը փոքրացնում են:
- Եթե բավականին շատ օբյեկտներ են գտնվում հեռու իրենց դասի եզրից, ապա C պարամետրը մեծացնում են:
- Խնդիրը լուծում են նոր հավաքածուի և նոր C պարամետրի դեպքում: Այսպես կատարվում է մի քանի իտերացիա:

Այսինքն՝ որոնվում է, այսպես ասած, հավասարակշռության կետ: Լագրանժի ֆունկցիան տվյալ խնդրի համար ունի հետևյալ տեսքը՝

$$L(w, w_0, \xi, \lambda, \eta) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^l \lambda_i (y_i((w, x_i) - w_0) - 1) - \sum_{i=1}^l \xi_i (\lambda_i + \eta_i - C):$$

Համաձայն Կուն-Թակերի թեորեմի՝ օպտիմալության պայմանները կունենան հետևյալ տեսքերը՝

$$\begin{cases} \frac{\partial L}{\partial w} = w - \sum_{i=1}^l \lambda_i y_i x_i = 0 \Rightarrow w = \sum_{i=1}^l \lambda_i y_i x_i, \\ \frac{\partial L}{\partial w_0} = - \sum_{i=1}^l \lambda_i y_i = 0, \\ \frac{\partial L}{\partial \xi_i} = -\lambda_i - \eta_i + C = 0, \Rightarrow \eta_i + \lambda_i = C, i = 1, 2, \dots, l, \\ \lambda_i \geq 0, \eta_i \geq 0, \xi_i \geq 0, i = 1, 2, \dots, l, \\ \lambda_i ((w, x_i) - w_0) = 1 - \xi_i, i = 1, 2, \dots, l, \\ \eta_i \xi_i = 0, i = 1, 2, \dots, l: \end{cases}$$

Այստեղից՝ $\lambda_i, \xi_i, \eta_i, m_i$ պարամետրերի միջև տեղի ունեն հետևյալ հարաբերակցությունները՝

- $\lambda_i = 0, \eta_i = C, \xi_i = 0, m_i > 0 \Rightarrow x_i$ օբյեկտը դասվել է ճիշտ և հեռու է բաժանող շերտից: Այդպիսի օբյեկտներ կոչվում են «ծայրամասային»:
- $0 < \lambda_i < C, 0 < \eta_i < C, \xi_i = 0, m_i = 1, \Rightarrow x_i$ օբյեկտը ճիշտ է դասվել և բաժանող շերտի եզրին է: Դրանք կոչվում են հենքային օբյեկտներ: ,
- $\lambda_i = C, \eta_i = 0, \xi_i > 0, m_i < 1, \Rightarrow x_i$ օբյեկտը կոչվում է «խախտող»:

Իսկ λ_i պարամետրերի որոշման համար լուծվում է քառակուսային ծրագրավորման հետևյալ երկակի խնդիրը՝

$$\begin{cases} L(\lambda) = -\sum_{i=1}^l \lambda_i + \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l \lambda_i \lambda_j y_i y_j (x_i, x_j) \rightarrow \min_{\lambda}, \\ 0 \leq \lambda_i \leq C, \quad i = 1, 2, \dots, l, \\ \sum_{i=1}^l \lambda_i y_i = 0 : \end{cases}$$

Հիշենք, որ այս խնդրում x_i օբյեկտի լուսանցք (margin) կոչվում է հետևյալ մեծությունը՝

$$m_i = y_i((w, x_i) - w_0) :$$

Ալգորիթմը սխալվում է x_i օբյեկտի վրա, եթե $m_i < 0$: օբյեկտը գտնվում է բաժանող շերտի մեջ, եթե $m_i \in (-1, 1)$: Եթե $m_i > 1$, ապա օբյեկտը դասակարգված է ճիշտ: Այս դեպքում ալգորիթմի սխալների քանակը հավասար կլինի

$$\sum_{i=1}^l [m_i < 1] :$$

Այժմ նկարագրենք այդ սխալների մինիմիզացման սուբգրադիենտային մեթոդները: Ծեմային կորստի ֆունկցիան փոխարինվում է կտոր առ կտոր գծային ֆունկցիայով, որը վերևից մոտարկում է նրան՝ ինչպես ցույց է տրված նկար 6.5-ում: Այս փոխարինման իմաստը նրանում է, որ կորստի ֆունկցիան սխալի մեծության նկատմամբ ավելի զգայուն է դառնում և ալգորիթմը տուգանում է օբյեկտին դասի սահմանին մոտենալու համար: Հեշտ է նկատել, որ վերևի խնդիրը համարժեք է

$$Q(w, w_0) = \sum_{i=1}^l (1 - m_i)_+ + \frac{1}{2C} \|w\|_2^2 \rightarrow \min_{w, w_0}$$

խնդիրին:

Դիտարկենք այս խնդիրը հետևյալ տեսքով՝

$$f(w, w_0) = \frac{1}{l} \sum_{i=1}^l (1 - y_i((w, x_i) - w_0)_+ + \frac{\lambda}{2} \|w\|_2^2 \rightarrow \min :$$

Այժմ ցույց տանք, որ եթե դիտարկենք միայն համասեռ ենթատարածությունների դեպքը ($w_0 = 0$), ապա խնդիրը կարելի է լուծել ստոխաստիկ գրադիենտի պրոյեկտման մեթոդով: Այդ դեպքում $f(w)$ ֆունկցիան λ -ուժեղ ուռուցիկ է, և հետևաբար այն ունի միակ մինիմումի կետ՝ $w = w^*$:

Քանի որ $f(0) = 1$, ապա օպտիմալ w^* լուծումը բավարարում է $\|w^*\|_2^2 \leq 2/\lambda$: Հետևաբար խնդիրը համարժեք է հետևյալին՝

$$f(w) \rightarrow \min, \quad \|w\|_2^2 \leq \frac{2}{\lambda} :$$

Նշանակենք $g_i(w) = \max\{0, 1 - y_i(w, x_i)\}$ և հաշվենք f ֆունկցիայի սուբգրադիեն-

րենցիալը՝ օգտվելով գումարի և մաքսիմումի սուբգրադիենցիալի վերաբերյալ կանոններից.

$$\partial f(w) = \frac{1}{l} \sum_{i=1}^l \partial g_i(w) + \lambda w:$$

Ունենք

$$-\frac{1}{l} \sum_{i=1}^l g_i + \lambda w \in \partial f(w),$$

որտեղ

$$v_i = \begin{cases} 0, & 1 \leq y_i(w, x_i), \\ -y_i x_i, & 1 > y_i(w, x_i) \end{cases}$$

վեկտորը պատկանում է $\partial g_i(w)$ սուբգրադիենցիալին:

Սուբգրադիենտային իջեցման մեթոդը w պարամետրի նկատմամբ կրնդունի հետևյալ տեսքը՝

$$\begin{aligned} w^{t+1} &= \Pi(w^t - \eta_t(-\frac{1}{l} \sum_{i=1}^l v_i^t + \lambda w^t)) = \\ &= \Pi((1 - \lambda \eta_t)w_t + \frac{1}{l} \eta_t \sum_{i=1}^l v_i^t): \end{aligned}$$

Այստեղ

$$v_i^t = \begin{cases} 0, & 1 \leq y_i(w^t, x_i), \\ -y_i x_i, & 1 > y_i(w^t, x_i) \end{cases},$$

իսկ

$$\Pi(z) = \frac{z}{\max\{1, \|z\|_2/r\}}, \quad r = \sqrt{2/\lambda}$$

պրոյեկտման օպերատորն է գրո կենտրոնով և r շառավղով գնդի վրա:

Այժմ նկարագրենք ստոխաստիկ գրադիենտային իջեցման մեթոդը: Ինչպես նշվեց վերևում, խնդիրը համարժեք է

$$f(w) \rightarrow \min_{w \in A}, \quad A = \{w : \|w\|^2 \leq \frac{2}{\lambda}\}$$

խնդրին:

Դիցուք ξ_t անկախ պատահական մեծություններ են, միևնույն ձևով բաշխված, որոնք արժեքներ են ընդունում $\{1, 2, \dots, l\}$ բազմությունից: Ստոխաստիկ գրադիենտի մեթոդը կնկարագրվի հետևյալ ռեկուրենտ առնչությամբ՝

$$w^{t+1} = \Pi_A(w^t - \frac{1}{\lambda} \eta_t v(x_t, \xi_t)), \quad v(x_t, \xi_t) = \begin{cases} \lambda w^t, & 1 \leq y_{\xi_t}(w^t, x_{\xi_t}) \\ \lambda w^t - y_{\xi_t} x_{\xi_t}, & 1 > y_{\xi_t}(w^t, x_{\xi_t}) \end{cases}:$$

Այստեղ մենք կիրառեցինք $(1 - y(w, x))_+$ ֆունկցիայի սուբգրադիենտը հաշվելու բանաձևը, որը տրված է ուռուցիկ ֆունկցիայի սուբգրադիենցիալի բաժնում:

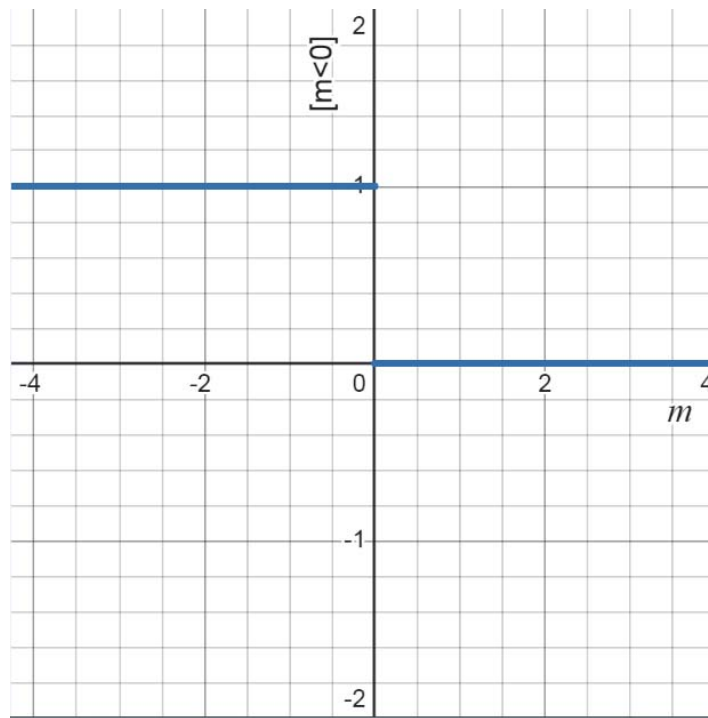
Հիշենք նաև, որ ստոխաստիկ սուբգրադիենտային մեթոդում, ինչպես հայտնի է, որպես օպտիմալ լուծում T իտերացիայից հետո կարող ենք վերցնել այդ իտերեցիաների ընթացքում կառուցված վեկտորների միջին արժեքը:

6.5 Կորստի Ֆունկցիաները բինար դասակարգման խնդիրներում

Բինար դասակարգման ժամանակ Էմպիրիկ ոիսկի Ֆունկցիոնալը ունի հետևյալ տեսքը՝

$$L = \frac{1}{l} \sum_{i=1}^l [m_i < 0],$$

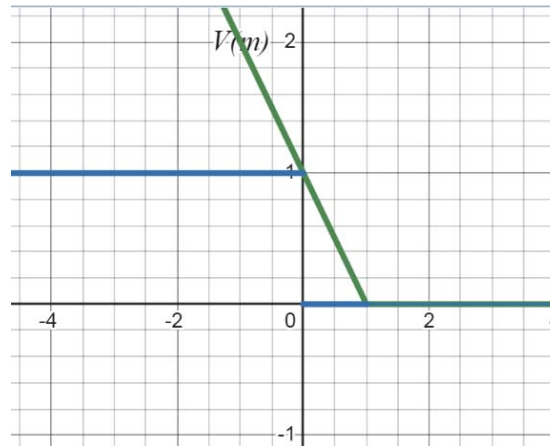
որտեղ $m_i = y_i((w, x_i) - w_0)$: Դիցուք $\mathbf{w} = (w_0, w_1, \dots, w_n)$, $\mathbf{x}_i = (-1, x_{1i}, \dots, x_{ni})$: Այդ դեպքում $[m = y(\mathbf{w}, \mathbf{x}) < 0]$ Ֆունկցիան խզվող է և կոչվում է հենքային: Նկար 6.6-ում պատկերված է դրա գրաֆիկը: Այս Ֆունկցիան մոտարկվում է ողորկ Ֆունկցիաներով: Այդպիսի մոտարկող Ֆունկցիաներ կարելի է նշել բավականին շատ, որոնցից մի քանիսը ներկայացված են նկ. 6.7- 6.9-ում:



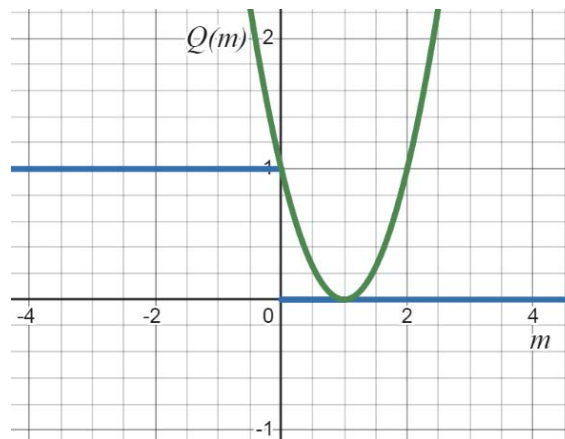
Նկ. 6.6: $[m < 0]$ հենքային կորստի Ֆունկցիայի գրաֆիկը

Օրինակ՝ ընտրելով $Q(m)$ Ֆունկցիան՝ կստանանք գծային ռեգրեսիայի խնդիրը: Այսինքն՝ կորստի Ֆունկցիան այս դեպքում հավասար կլինի $(1 - y(\mathbf{w}, \mathbf{x}))^2$ արտահայտությանը: Արդյունքում կունենանք հետևյալ խնդիրը՝

$$\tilde{L}(\mathbf{w}) = \frac{1}{l} \sum_{i=1}^l (1 - (\mathbf{w}, \mathbf{x}_i) y_i)^2 \rightarrow \min:$$



Նկ. 6.7: $[m < 0]$ հենքային և կորստի $V(m) = \max(0, 1 - m)$ ֆունկցիաների գրաֆիկները (SVM)



Նկ. 6.8: $[m < 0]$ հենքային և կորստի $Q(m) = (1 - m)^2$ ֆունկցիաների գրաֆիկները (FLD-քառակուսային)

Հաշվելով այս ֆունկցիայի գրադիենտը և հավասարեցնելով զրոյի՝ կստանանք

$$\tilde{L}'(\mathbf{w}) = -\frac{2}{l} \sum_{i=1}^l (1 - (\mathbf{w}, \mathbf{x}_i) y_i) \mathbf{x}_i^T y_i = 0:$$

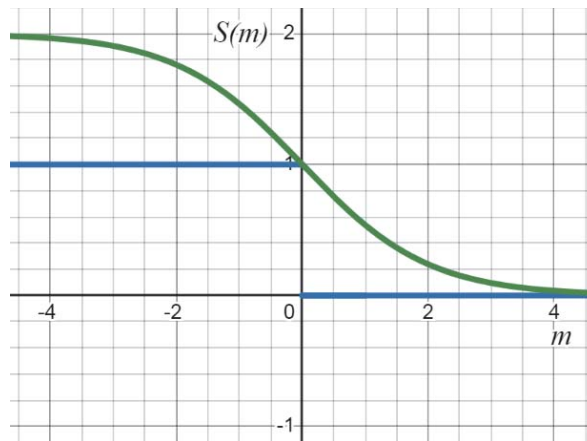
Այստեղից կստանանք

$$\sum_{i=1}^l \mathbf{x}_i^T y_i - \mathbf{w}^T \sum_{i=1}^l \mathbf{x}_i \mathbf{x}_i^T = 0:$$

Այսինքն՝

$$\mathbf{w}^T = \sum_{i=1}^l \mathbf{x}_i^T y_i \left(\sum_{i=1}^l \mathbf{x}_i \mathbf{x}_i^T \right)^{-1}:$$

Վերջում նշենք, որ վերցնելով վերը նշված կորստի ողորկ ֆունկցիաներից որևէ մեկը և կատարելով միջին ռիսկի ֆունկցիայի մինիմիզացիա՝ կառուցում ենք դասերը անջատող հիպերհարթություններ:



Նկ. 6.9: $[m < 0]$ հենքային և կորստի $S(m) = 2(1 + \exp(m))^{-1}$ ֆունկցիաների գրաֆիկները (ANN-սիզմոյիդ)



Նկ. 6.10: Գծորեն չբաժանվող հավաքածուի օրինակ

6.6 Միջուկների մեթոդը մեքենայական ուսուցման մեջ

Ոչ բոլոր դեպքերում է S ուսուցանող հավաքածուն գծորեն բաժանելի: Օրինակ՝ դիտարկենք S հավաքածուն, որը տրված է նկ. 6.10-ում, որում աստղանիշներով նշված են S^- բազմության տարրերը, իսկ մնացած կետերը պատկանում են S^+ դասին: Այս դասերը գծորեն անջատելի չեն:

Այս դեպքում որոշիչ ֆունկցիան կարելի է ընտրել հետևյալ բանաձևով՝

$$a_S(x) = \text{sign}(\|x\| - w_0):$$

Եթե դասերը գծորեն չեն անջատվում, ապա կառուցում են ոչ գծային $\phi(x)$ ֆունկցիոնալ, այնպիսին, որ

$$\phi(x) = \begin{cases} 1, & \text{եթե } x \in S^+, \\ -1, & \text{եթե } x \in S^-: \end{cases}$$

Նման ոչ գծային ֆունկցիոնալի կառուցումը կատարվում է հետևյալ ընթացակարգով.

- կառուցվում է ոչ գծային

$$\varphi: \mathbb{R}^n \rightarrow \mathbb{R}^N$$

ձևափոխություն, այնպիսին, որ

$$S_\varphi = \{(\varphi(x), y_x) \mid x \in X_S\}$$

հավաքածուն լինի գծորեն բաժանելի ($\mathbb{R}^N$ տարածությունը կոչվում է ուղղիչ տարածություն):

- Հենման վեկտորների մեթոդով S_φ հավաքածուի համար կառուցվում է որոշիչ ֆունկցիա՝ հետևյալ բանաձևով՝

$$a_{S_\varphi}(x) = \text{sign}((\varphi(x), w) - w_0):$$

Վերը նշված օրինակում որպես ուղղիչ տարածություն կարելի է վերցնել $\mathbb{R}^2$ -ը, իսկ $\varphi(x) \equiv (x_1^2, x_2^2)$: Կարելի է ցույց տալ, որ ամեն մի S հավաքածու ունի ուղղիչ տարածություն: Իրոք, եթե S հավաքածուի օբյեկտների քանակը հավասար է N -ի, ապա նրա համար ուղղիչ տարածություն կլինի $\mathbb{R}^N$ -ը, և այս պարագայում որպես $\varphi: \mathbb{R}^n \rightarrow \mathbb{R}^N$ կարելի է վերցնել այնպիսի ինյեկտիվ արտապատկերում, որը S հավաքածուի յուրաքանչյուր տարրին համապատասխանության մեջ դնի $\mathbb{R}^N$ տարածության օբյեկտնորմալ բազիսի որևէ էլեմենտ: Այդ դեպքում S_φ հավաքածուն $\mathbb{R}^N$ -ում կլինի գծորեն անկախ:

Տեղի ունի հետևյալ ընդհանուր թեորեմը:

Թեորեմ 6.4. *Գծորեն անկախ վեկտորների կամայական հավաքածու $\mathbb{R}^n$ -ում կարելի է գծորեն բաժանել երկու դասի:*

Ապացույց: Դիցուք $x_1, x_2, \dots, x_m$ վեկտորները գծորեն անկախ են $\mathbb{R}^n$ -ում: Այդ դեպքում $m \leq n$: Դիտարկենք այն դեպքը, երբ $m = n$: Այդ դեպքում

$$V \equiv [x_1, \dots, x_n]^T$$

մատրիցը կլինի չվերասերվող: Հետևաբար կամայական y վեկտորի համար

$$Vw = y$$

համակարգը ունի լուծում: Դիցուք ենթադրենք, որ

$$x_1, \dots, x_p \in S^+, \quad x_{p+1}, \dots, x_n \in S^-:$$

Վերցնենք

$$y_1 > 0, \dots, y_p > 0, \quad y_{p+1} < 0, \dots, y_n < 0:$$

Քանի որ w վեկտորը $Vw = y$ հավասարման լուծումն է, ուստի

$$(w, x_i) > 0, \quad i = 1, \dots, p, \quad (w, x_i) < 0, \quad i = p+1, \dots, n,$$

այսինքն՝ S^+ և S^- դասերը գծորեն անջատվում են: Եթե $m < n$, ապա վեկտորների $x_1, x_2, \dots, x_m$ համակարգը կարելի է լրացնել մինչև գծորեն անկախ լրիվ համակարգի $\mathbb{R}^n$ -ում և կրկնել նախորդ դատողությունները:

Եթե S_ϕ հավաքածուն գծորեն բաժանելի է, ապա նրա որոշիչ ֆունկցիան կունենա հետևյալ տեսքը՝

$$\begin{aligned} a_{S_\phi}(x) &= \text{sign}((\phi(x), \tilde{w}) - \tilde{w}_0) = \\ &= \text{sign}(\phi(x), \sum_{x' \in X_S} \tilde{\lambda}_{x'} y_{x'} \phi(x')) - w_0) = \\ &= \text{sign}(\sum_{x' \in X_S} \tilde{\lambda}_{x'} y_{x'} (\phi(x), \phi(x')) - \tilde{w}_0), \end{aligned}$$

որտեղ $\tilde{\lambda}$ -ն

$$\begin{aligned} \sum_{x, x' \in X_S} \lambda_x \lambda_{x'} \phi(x), \phi(x')) - \sum_{x \in X_S} \lambda_x &\rightarrow \min, \\ x \in X_S, \lambda_x &\geq 0, \sum_{x \in X_S} \lambda_x y_x = 0 \end{aligned}$$

խնդրի լուծումն է: Այստեղից հետևում է, որ որոշիչ a_S ֆունկցիայի կառուցման համար պետք չէ ճշտել ուղղիչ տարածության չափողականությունը և $\phi(x)$, $x \in X_S$ վեկտորների տեսքը, այլ բավական է իմանալ $(\phi(x), \phi(x'))$ սկալյար արտադրյալը:

Մեքենայական ուսուցման միջուկների մեթոդի էությունն այն է, որ որոշիչ ֆունկցիայի կառուցման պրոցեսում $(\phi(x), \phi(x'))$ սկալյար արտադրյալը փոխարինվում է

$$K : \mathbb{R}^n \times \mathbb{R}^n \rightarrow R$$

ֆունկցիայով, որը կոչվում միջուկ: Այն պետք է բավարարի հետևյալ պայմանները:

- սիմետրիկություն՝ $K(x, x') = K(x', x) \forall x, x' \in \mathbb{R}^n$,
- դրական կիսաորոշյալություն՝

$$\sum_{i,j=1}^m \alpha_i \alpha_j K(x_i, x_j) \geq 0, \forall x_1, \dots, x_m \in \mathbb{R}^n, \alpha_1, \alpha_2, \dots, \alpha_m \in R:$$

Ջեյմս Մերսերի կողմից ապացուցվեց նաև հակադարձ պնդումը, որ եթե K ֆունկցիոնալը բավարարում է վերը նշված երկու պայմանները, ապա գոյություն ունի $\phi : \mathbb{R}^n \rightarrow H$ արտապատկերում, այնպիսին, որ

$$K(x, x') = \langle \phi(x), \phi(x') \rangle:$$

Այստեղ H -ը $\langle \bullet, \bullet \rangle$ սկալյար արտադրյալով ինչ-որ տարածություն է:

Այժմ դիտարկենք ուղղիչ տարածության կառուցման մի օրինակ քառակու-

սային միջուկի միջոցով.

$$X = \mathbb{R}^2, K(u, v) = (u, v)^2, u = (u_1, u_2), v = (v_1, v_2):$$

Գտնենք H տարածությունը և արտապատկերում $\varphi: X \rightarrow H$, այնպիսին, որ

$$K(x, x') = \langle \varphi(x), \varphi(x') \rangle:$$

Ունենք

$$\begin{aligned} K(u, v) &= \langle u, v \rangle^2 = \langle (u_1, u_2), (v_1, v_2) \rangle^2 = (u_1 v_1 + u_2 v_2)^2 = \\ &= u_1^2 v_1^2 + u_2^2 v_2^2 + 2u_1 v_1 u_2 v_2 = ((u_1^2, u_2^2, \sqrt{2}u_1 u_2), (v_1^2, v_2^2, \sqrt{2}v_1 v_2)): \end{aligned}$$

Այսպիսով՝

$$H = \mathbb{R}^3, \varphi: (u_1, u_2) \rightarrow (u_1^2, u_2^2, \sqrt{2}u_1 u_2):$$

Այսինքն՝ X տարածության էլիպսային մակերևույթը արտապատկերվում է H տարածության հարթությանը:

Այժմ նկարագրենք միջուկների մեթոդով որոշիչ ֆունկցիայի կառուցման ալգորիթմը:

- Գտնում ենք

$$\sum_{x \in X_S} \lambda_x - 0.5 \sum_{x, x' \in X_S} \lambda_x \lambda_{x'} y_x y_{x'} K(x, x')$$

ֆունկցիայի մաքսիմալ արժեքը

$$\sum_{x \in X_S} \lambda_x y_x = 0, \lambda_x \geq 0, x \in X_S$$

բազմության վրա:

- Որոշիչ ֆունկցիան կլինի՝

$$a_S(x) = \text{sign}\left(\sum_{x' \in X_S} \hat{\lambda}_{x'} y_{x'} K(x, x') - w_0\right),$$

որտեղ

$$w_0 = \sum_{x' \in X_S} \hat{\lambda}_{x'} y_{x'} K(x'', x') - y_{x''},$$

ընդ որում $x'' \in X_S, \hat{\lambda}_{x''} \neq 0$:

Թեորեմ 6.5. Դիցուք հարթության վրա տրված են $m + 1$ հատ կետեր, որոնք բաժանված են երկու խմբի՝ S^+ և S^- : Այդ դեպքում կարելի է կառուցել m աստիճանի այնպիսի բազմանդամ, որի գրաֆիկով անջատվում են այդ խմբերը:

Ապացույց: Ենթադրենք՝ բոլոր այդ $(x_1, y_1), (x_2, y_2), \dots, (x_{m+1}, y_{m+1})$ կետերի արգիսները իրարից տարբեր են: Այդ դեպքում համաձայն ինտերպոլիացիայի

թեորեմի, այդ կետերով անցնում է $y = f(x) = a_0x^m + a_1x^{m-1} + \dots + a_m$ բազմանդամ:
Այսինքն՝

$$y_i = f(x_i), \quad i = 1, 2, \dots, m+1:$$

Վերցնենք նոր կետեր՝

$$(x_i, y_i^\epsilon), \quad i = 1, 2, \dots, m+1,$$

որտեղ $\epsilon > 0$, և

$$y_i^\epsilon = \begin{cases} y_i + \epsilon, & \text{եթե } (x_i, y_i) \in S^+, \\ y_i - \epsilon, & \text{եթե } (x_i, y_i) \in S^-: \end{cases}$$

Տանենք հանրահաշվական կոր այս նոր կետերով: Ակնհայտ է, որ այս կորը կրաժանի S^+ և S^- դասերը իրարից:

Եթե որոշ կետերի արգիսներ համընկնում են, ապա կոորդինատային համակարգը միշտ կարելի է պտտել ինչ-որ անկյան տակ այնպես, որ բոլոր կետերի արգիսները այդ նոր համակարգում լինեն իրարից տարբեր, և այնտեղ կառուցել անջատող հանրահաշվական կորը:

Որոշիչ ֆունկցիայի միջուկով կառուցման ալգորիթմը մեկնաբանենք օրինակի վրա:

Դիցուք

$$x_1 = (0, 0), y_1 = 1, \quad x_2 = (2, 0), y_2 = 1, \quad x_3 = (1, 0), y_3 = -1:$$

Պարզ է, որ այս կետերը հնարավոր չէ գծորեն անջատել: Քանի որ կետերի քանակը հավասար է երեքի, ապա կարելի է վերցնել հետևյալ միջուկը՝

$$K(x, y) = ((xy) + 1)^2, \quad (m = 2):$$

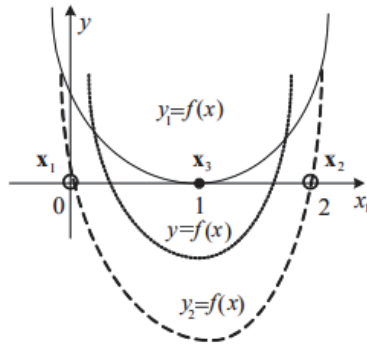
Կազմենք հետևյալ ֆունկցիան՝

$$\begin{aligned} F(\lambda) &= \lambda_1 + \lambda_2 + \lambda_3 - \frac{1}{2}(\lambda_1^2 + 25\lambda_2^2 + 4\lambda_3^2 + 2\lambda_1\lambda_2 - \\ &\quad - 2\lambda_1\lambda_3 - 18\lambda_2\lambda_3) = 2\lambda_1 + 2\lambda_2 - \\ &\quad - \frac{1}{2}(3\lambda_1^2 + 11\lambda_2^2 - 10\lambda_1\lambda_2): \end{aligned}$$

Այստեղ հաշվի է առված $\lambda_2 = \lambda_1 + \lambda_2$ պայմանը: Ածանցելով այս ֆունկցիան և հավասարեցնելով այն զրոյի՝ կստանանք՝ $\lambda = (4, 2, 6)$: Հետևաբար՝

$$f(x) = 4K(x_1, x) + 2K(x_2, x) - 6K(x_3, x) + 1 - (4K(x_1, x_1) + 2K(x_2, x_1) - 6K(x_3, x_1)) = 2x_1^2 - 4x_1 + 1,$$

$$f_1(x) = f(x) + 1, \quad f_2(x) = f(x) - 1 \quad (\text{տե՛ս նկ.6.11}):$$



Նկ. 6.11: $K(x, y) = ((x, y) + 1)^m$ միջուկով կլաստերավորում

ԽՆԴԻՐՆԵՐ

1) Յույց տալ, որ տրված ֆունկցիաները միջուկներ են:

- $K(x, x') = (c(x, x') + d)^m$, որտեղ $c \geq 0$, $d \geq 0$, իսկ m -ը՝ ոչ բացասական ամբողջ թիվ (պոլինոմիալ միջուկ):
- $K(x, x') = \sigma(c(x, x') + d)$, որտեղ

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

(սիգմոյիդալ միջուկ):

2) Սույց տալ, որ եթե $K_1(x, x')$ և $K_2(x, x')$ ֆունկցիաները միջուկներ են, ապա միջուկներ են նաև հետևյալ ֆունկցիաները.

- $K_1(x, x')K_2(x, x')$:
- $\alpha_1 K_1(x, x') + \alpha_2 K_2(x, x')$, $\alpha_1 \geq 0$, $\alpha_2 \geq 0$:

3) Յույց տալ, որ $K(x, y) = e^{(x, y)}$ ֆունկցիան միջուկ է:

Լուծում: Ունենք

$$ep(x, y) = \sum_{k=0}^{\infty} \frac{(x, y)^k}{k!}:$$

Գումարի ամեն մի անդամ բազմանդամ է դրական գործակիցներով և ուրեմն միջուկ է: Ուստի ամբողջ գումարը նույնպես միջուկ է:

4) Գտնել $\varphi(x)$ ֆունկցիայի նորմը՝ իմանալով $K(x, y)$ միջուկը:

Լուծում:

$$\|\varphi(x)\| = \sqrt{(\varphi(x), \varphi(x))} = \sqrt{K(x, x)}:$$

5) Գտնել $\varphi(x)$, $\varphi(z)$ վեկտորների միջև եղած հեռավորությունը $K(x, z)$ միջուկի միջոցով:

Լուծում:

$$\begin{aligned} \rho^2(\varphi(x), \varphi(z)) &= \|\varphi(x) - \varphi(z)\|^2 = \\ &= (\varphi(x), \varphi(x)) + (\varphi(z), \varphi(z)) - 2(\varphi(x), \varphi(z)) = \end{aligned}$$

$$= K(x, x) + K(z, z) - 2K(x, z):$$

6) Դիցուք $K(x, x')$ ֆունկցիան միջուկ է, ապացուցել, որ տեղի ունի հետևյալ անհավասարությունը՝

$$K(x, x') \leq \sqrt{K(x, x)}\sqrt{K(x', x')}, \quad \forall x, x' \in \mathbb{R}^n,$$

որը Կոչի-Բունյակովսկու անհավասարությունն է:

6.7 Բազմությունների ռադեմախերյան բարդություն

Այս պարագրաֆում ներմուծվում է բազմության այսպես ասած ռադեմախերյան բարդության գաղափարը, որը հետագայում կօգտագործվի SVM ալգորիթմի բարդության գնահատման հարցերում: Կքննարկվի նաև օպտիմալ պրեդիկտորների գոյության հարցը հենման վեկտորների մեթոդում:

Սահմանում: Դիցուք $A \subset \mathbb{R}^m$: A բազմության բարդություն է կոչվում հետևյալ մեծությունը՝

$$R(A) = \frac{1}{m} E_{\sigma} \left[\sup_{a \in A} \sum_{i=1}^m \sigma_i a_i \right],$$

որտեղ $P(\sigma_i = 1) = P(\sigma_i = -1) = \frac{1}{2}$:

Բերենք բազմության բարդությունը հաշվելու երկու օրինակ:

Օրինակ 1: Դիցուք $A = \{(a_1, a_2) \in \mathbb{R}^2\}$: Այդ դեպքում

$$R(A) = \frac{1}{2} \left(\frac{1}{4} ((a_1 + a_2) + (a_1 - a_2) + (-a_1 + a_2) + (-a_1 - a_2)) \right) = 0:$$

Օրինակ 2: Դիցուք $A = \{(1, 1), (1, 2)\}$: Այդ դեպքում կունենանք

$$\begin{aligned} R(A) &= \frac{1}{2} \left(\frac{1}{4} \max\{1 + 1, 1 + 2\} + \frac{1}{4} \max\{1 - 1, 1 - 2\} + \right. \\ &\quad \left. + \frac{1}{4} \max\{-1 - 1, -1 - 2\} + \frac{1}{4} \max\{-1 + 1, -1 + 2\} \right) = \frac{1}{4}: \end{aligned}$$

Այս սահմանումից անմիջապես բխում է հետևյալ պնդումը:

Լեմմա 6.1. Կամայական $c \in \mathbb{R}$ սկալյարի և $a_0 \in \mathbb{R}^m$ վեկտորի համար տեղի ունի

$$R(\{ca + a_0, a \in A\}) \leq |c|R(A):$$

Հաջորդ պնդումը հաստատում է, որ բազմության բարդությունը համընկնում է նրա ուռուցիկ թաղանթի բարդության հետ:

Լեմմա 6.2. Դիցուք $A \subset \mathbb{R}^m$: Այդ դեպքում տեղի ունի $R(A) = R(\text{conv}(A))$ անհավասարությունը:

Ապացույց: Ակնհայտ է, որ կամայական v վեկտորի համար տեղի ունի հետևյալ հավասարությունը՝

$$\sup_{\alpha \geq 0, |\alpha|_1=1} \sum_{j=1}^N \alpha_j v_j = \max_j v_j:$$

Հետևաբար՝

$$\begin{aligned} mR(\text{conv}(A)) &= E_\sigma \left[\sup_{\alpha \geq 0: |\alpha|_1=1} \sup_{a^{(1)}, a^{(2)}, \dots, a^{(N)}} \sum_{i=1}^m \sigma_j \sum_{j=1}^N \alpha_j a_i^{(j)} \right] \\ &= E_\sigma \left[\sup_{\alpha \geq 0, |\alpha|=1} \sum_{j=1}^N \alpha_j \sup_{a^{(j)}} \sum_{i=1}^m \sigma_i a_i^{(j)} \right] = \\ &= E_\sigma \left[\sup_{a \in A} \sum_{i=1}^m \sigma_i a_i \right] = mR(A): \end{aligned}$$

Լեմման ապացուցվեց:

Հաջորդ լեմման ցույց է տալիս, որ վերջավոր բազմության բարդությունը լոգարիթմական կախվածություն ունի բազմության հզորությունից:

Լեմմա 6.3. (Մասարտ) Դիցուք $A = \{a^1, \dots, a^N\}$ -ն վեկտորների ենթաբազմություն է $\mathbb{R}^m$ -ից:

Այդ դեպքում

$$R(A) \leq \max_{a \in A} \|a - \bar{a}\| \frac{\sqrt{2 \ln N}}{m},$$

որտեղ

$$\bar{a} = \frac{1}{N} \sum_{i=1}^N a^{(i)}:$$

Ապացույց: Ընդհանրությունը չխախտելով՝ կարող ենք ենթադրել, որ $\bar{a} = 0$: Դիցուք

$$A' = \{\lambda a^{(1)}, \dots, \lambda a^{(N)}\}, \quad \lambda > 0:$$

Ունենք

$$\begin{aligned} mR(A') &= E_\sigma [\max_{a \in A'} (\sigma, a)] = \\ &= E_\sigma [\ln(\max_{a \in A'} e^{(\sigma, a)})] \leq \\ &\leq E_\sigma [\ln(\sum_{a \in A'} e^{(\sigma, a)})] \leq \\ &\leq \ln(E_\sigma [\sum_{a \in A'} e^{(\sigma, a)}]) = \\ &\leq \ln(\sum_{a \in A'} \prod_{i=1}^m E_{\sigma_i} [e^{\sigma_i a_i}]): \end{aligned}$$

Այնուհետև օգտվենք հետևյալ անհավասարությունից՝

$$\frac{e^a + e^{-a}}{2} \leq e^{a^2/2}.$$

Նախ ապացուցենք այս անհավասարությունը: Ունենք

$$e^a = \sum_{n=0}^{\infty} \frac{a^n}{n!} :$$

Հետևաբար՝

$$\frac{e^a + e^{-a}}{2} = \sum_{n=0}^{\infty} \frac{a^{2n}}{(2n)!}, \quad e^{a^2/2} = \sum_{n=0}^{\infty} \frac{a^{2n}}{2^n n!} :$$

Քանի որ $(2n)! \geq 2^n n!$, $n \geq 0$, ապա նշված անհավասարության ապացույցը ավարտված է: Օգտվելով այս անհավասարությունից՝ ստանում ենք

$$E_{\sigma_i}[e^{\sigma_i a_i}] = \frac{e^{a_i} + e^{-a_i}}{2} \leq e^{a_i^2/2} :$$

Հետևաբար՝

$$\begin{aligned} mR(A') &\leq \ln \left(\sum_{a \in A'} \prod_{i=1}^m e^{a_i^2/2} \right) = \ln \left(\sum_{a \in A'} e^{\|a\|^2/2} \right) \leq \\ &\leq \ln(|A'| \max_{a \in A'} e^{\|a\|^2/2}) = \ln(|A'|) \max_{a \in A'} (\|a\|^2/2) : \end{aligned}$$

Քանի որ $R(A) = \frac{1}{\lambda} R(A')$, ապա

$$R(A) \leq \frac{\ln(|A'|) + \lambda^2 \max_{a \in A} (\|a\|^2/2)}{\lambda m} :$$

Տեղադրելով $\lambda = \sqrt{2 \ln(|A|) / \max_{a \in A} \|a\|^2}$ կստանանք անհրաժեշտ անհավասարությունը:

Լեմմա 6.4. Դիցուք

$$H_2 = \{x \rightarrow (w, x) : \|w\|_2 \leq 1\},$$

իսկ

$$S = \{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$$

վեկտորների հավաքածու է հիլբերտյան տարածությունում:
Նշանակենք՝

$$H_2 \circ S = \{(w, x^{(1)}), \dots, (w, x^{(m)}), \|w\|_2 \leq 1\} :$$

Այդ դեպքում

$$R(H_2 \circ S) \leq \frac{\max_i \|x^{(i)}\|_2}{\sqrt{m}} :$$

Ապացույց: Ունենք

$$mR(H_2 \circ S) = E_{\sigma} \left[\sup_{a \in H_2 \circ S} \sum_{i=1}^m \sigma_i a_i \right] =$$

$$\begin{aligned}
&= E_\sigma \left[\sup_{w: \|w\|_2 \leq 1} \sum_{i=1}^m \sigma_i(w, x^{(i)}) \right] = \\
&= E_\sigma \left[\sup_{w: \|w\|_2 \leq 1} \left(w, \sum_{i=1}^m \sigma_i x^{(i)} \right) \right] \leq \\
&\leq E_\sigma \left[\left\| \sum_{i=1}^m \sigma_i x^{(i)} \right\|_2 \right]:
\end{aligned}$$

Կիրառենք այստեղ Յենսենի անհավասարությունը՝

$$\begin{aligned}
E_\sigma \left[\left\| \sum_{i=1}^m \sigma_i x^{(i)} \right\|_2 \right] &= E_\sigma \left[\left(\left\| \sum_{i=1}^m \sigma_i x^{(i)} \right\|_2^2 \right)^{1/2} \right] \leq \\
&\leq \left(E_\sigma \left[\sum_{i=1}^m \sigma_i^2 \left\| x^{(i)} \right\|_2^2 \right] \right)^{1/2}:
\end{aligned}$$

Քանի որ $\sigma_1, \sigma_2, \dots, \sigma_m$ անկախ պատահական մեծություններ են, ապա

$$\begin{aligned}
E_\sigma \left[\left\| \sum_{i=1}^m \sigma_i x^{(i)} \right\|_2^2 \right] &= \\
&= E_\sigma \left[\sum_{i,j=1}^m \sigma_i \sigma_j (x^{(i)}, x^{(j)}) \right] = \\
&= \sum_{i \neq j} (x^{(i)}, x^{(j)}) E[\sigma_i \sigma_j] + \sum_{i=1}^m ((x^{(i)}, x^{(i)}) E[\sigma_i^2]) = \\
&= \sum_{i=1}^m \|x^{(i)}\|^2 \leq m \max_i \|x^{(i)}\|^2:
\end{aligned}$$

Լեմման ապացուցվեց:

Լեմմա 6.5. Դիցուք $S = \{x^{(1)}, \dots, x^{(m)}\}$ կետերի հավաքածու է $\mathbb{R}^n$ տարածությունից: Այդ դեպքում տեղի ունի

$$R(H_1 \circ S) \leq \max_i \|x^{(i)}\|_\infty \sqrt{\frac{2 \ln(2n)}{m}}$$

անհավասարությունը, որտեղ

$$H_1 \circ S = \{(w, x^{(1)}), \dots, (w, x^{(m)}) : \|w\|_1 \leq 1\}:$$

Ապացույց: Կամայական երկու w, v վեկտորների համար տեղի ունի Գյուդերի անհավասարությունը՝

$$(w, x) \leq \|w\|_1 \|v\|_\infty,$$

որտեղ

$$\|w\|_1 = \sum_{i=1}^n |w_i|, \quad \|v\|_\infty = \max_i |v_i|:$$

Օգտագործելով այս անհավասարությունը՝ կստանանք

$$\begin{aligned} mR(H_1 \circ S) &= E_\sigma \left[\sup_{a \in H_1 \circ S} \sum_{i=1}^m \sigma_i a_i \right] = \\ &= E_\sigma \left[\sup_{w: \|w\|_1 \leq 1} \sum_{i=1}^m \sigma_i (w, x^{(i)}) \right] = \\ &= E_\sigma \left[\sup_{w: \|w\|_1 \leq 1} \left(w, \sum_{i=1}^m \sigma_i x^{(i)} \right) \right] \leq \\ &\leq E_\sigma \left[\left\| \sum_{i=1}^m \sigma_i x^{(i)} \right\|_\infty \right]: \end{aligned}$$

Կամայական $j \in [1 : n]$ ինդեքսի համար նշանակենք

$$u^{(j)} = (x_{1,j}, x_{2,j}, \dots, x_{m,j}):$$

Նկատենք, որ

$$\|u^j\|_2 \leq \sqrt{m} \max_i \|x^{(i)}\|_\infty:$$

Նշանակենք

$$U = \{u^{(1)}, \dots, u^{(n)}, -u^{(1)}, \dots, -u^{(n)}\}:$$

Վերևի անհավասարության աջ մասը հավասար է $mR(U)$: Իսկ, ըստ Մասարտի լեմմայի, կունենանք

$$R(U) \leq \max_i \|x^{(i)}\|_\infty \sqrt{2 \ln(2n)/m},$$

ինչը որ պահանջվում էր ապացուցվեց:

Լեմմա 6.6. Դիցուք $\varphi_i : R \rightarrow R$, $i = 1, 2, \dots, m$ ρ -լիպշիցյան ֆունկցիաներ են՝

$$|\varphi(\alpha) - \varphi(\beta)| \leq \rho |\alpha - \beta|:$$

Դիցուք $a \in \mathbb{R}^m$ և

$$\varphi(a) = (\varphi_1(a_1), \varphi_2(a_2), \dots, \varphi_m(a_m)), \quad \varphi \circ A = \{\varphi(a), a \in A\}:$$

Այդ դեպքում՝

$$R(\varphi \circ A) \leq \rho R(A):$$

Ապացույց: Ընդհանրությունը չիսխտելով՝ ենթադրենք $\rho = 1$: Նշանակենք

$$A_i = \{(a_1, a_2, \dots, a_{i-1}, \varphi_i(a_i), a_{i+1}, \dots, a_m) : a \in A\}:$$

Լեմման ապացուցելու համար բավական է ցույց տալ, որ

$$R(A_i) \leq R(A) \quad \forall i:$$

Ահավասարությունը ապացուցենք $i = 1$ -ի դեպքում: Ունենք

$$\begin{aligned} mR(A_1) &= E_\sigma \left[\sup_{a \in A_1} \sum_{a \in A_1} \sigma_i a_i \right] = \\ &= E_\sigma \left[\sup_{a \in A} (\sigma_1 \varphi_1(a_1) + \sum_{i=2}^m \sigma_i a_i) \right] \\ &= \frac{1}{2} E_{\sigma_2, \dots, \sigma_m} \left[\left(\sup_{a \in A} (\varphi_1(a_1) + \sum_{i=2}^m \sigma_i a_i) + \sup_{a \in A} (-\varphi_1(a_1) + \sum_{i=2}^m \sigma_i a_i) \right) \right] = \\ &= \frac{1}{2} E_{\sigma_2, \dots, \sigma_m} \left[\sup_{a, a' \in A} \left(\varphi_1(a_1) - \varphi_1(a'_1) + \sum_{i=2}^m \sigma_i a_i + \sum_{i=2}^m \sigma_i a'_i \right) \right] \leq \\ &\leq \frac{1}{2} E_{\sigma_1, \dots, \sigma_m} \left[\sup_{a, a' \in A} \left(|a_1 - a'_1| + \sum_{i=2}^m \sigma_i a_i + \sum_{i=2}^m \sigma_i a'_i \right) \right]: \end{aligned}$$

Այստեղ $|a_1 - a'_1|$ արտահայտության մոդուլի նշանը կարելի է բաց թողնել, քանի որ a, a' տարրերը ընտրվում են միևնույն A բազմությունից, և վերին եզրի նշանի տակ գտնվող գումարելիները չեն փոխվում, երբ a -ն փոխարինում ենք a' -ով: Հետևաբար՝

$$mR(A_1) \leq \frac{1}{2} E_{\sigma_1, \dots, \sigma_m} \left[\sup_{a, a' \in A} \left(a_1 - a'_1 + \sum_{i=2}^m \sigma_i a_i + \sum_{i=2}^m \sigma_i a'_i \right) \right]:$$

Հեշտ է նկատել, որ անհավասարության աջ մասը հավասար է $mR(A)$ -ի: Լեմման ապացուցվեց:

6.8 Օպտիմալ արեդիկտորները գծային դասերում և հենման վեկտորների մեթոդում

Դիտարկենք հիպոթեզների $H = \{w : \|w\|_2 \leq B\}$ և նմուշների $Z = X \times Y$ բազմությունները: Ենթադրենք կորստի $l : H \times Z \rightarrow R$ ֆունկցիան ունի հետևյալ տեսքը՝

$$l(w, (x, y)) = \varphi((w, x), y),$$

որտեղ $\varphi : R \times Y \rightarrow R$ -ն ρ -լիպշիցյան ֆունկցիա է յուրաքանչյուր ֆիքսած $y \in Y$ արժեքի դեպքում: Նման ֆունկցիայի օրինակ է արդեն հայտնի հենքային

$$l(w, (x, y)) = \max\{0, 1 - y(w, x)\}$$

ֆունկցիան, որը 1-լիպշիցյան է $Y = \{+1, -1\}$ դեպքում:

Նորից դիտարկենք հենման վեկտորների մեթոդի հիմնական խնդիրը հետևյալ տեսքով՝

$$\arg \min_w ||w||^2, \quad y_i(w, x_i) \geq 1, \quad \forall i:$$

Գրականության մեջ այս խնդիրը կոչվում է *Hard-SVM* համասեռ կիսատարածությունների դեպքում ($w_0 = 0$):

Տեղի ունի հետևյալ թեորեմը:

Թեորեմ 6.6. Դիցուք D հավանականային բաշխում է $X \times \{+1, -1\}$ բազմության վրա և գոյություն ունի այնպիսի w^* վեկտոր, որ

$$P_{(x,y) \sim D}[y(w^*, x) \geq 1] = 1, \quad ||x||_2 \leq R:$$

Դիցուք w_S -ը վերևի խնդրի լուծումն է: Այդ դեպքում նմուշների պատահական $S \sim D^m$ հավաքածուի համար առնվազն $1 - \delta$ հավանականությամբ տեղի ունի հետևյալ գնահատականը՝

$$P_{(x,y) \sim D^m}[y \neq \text{sign}((w_S, x))] \leq \frac{4R||w_S||}{\sqrt{m}} + \sqrt{\frac{\ln(4 \log_2(||w_S||)/(\delta))}{m}}:$$

Նախքան այս թեորեմի ապացույցը մեզ անհրաժեշտ են որոշ պնդումներ:

Դիցուք ունենք հիպոթեզների H բազմությունը և կորստի l ֆունկցիան: Նշանակենք

$$F \equiv l \circ H = \{z \rightarrow l(h, z), \quad h \in H\}:$$

Կամայական $f \in F$ ֆունկցիայի համար սահմանենք

$$L_D(f) = E_{z \sim D}[f(z)], \quad L_S(f) = \frac{1}{m} \sum_{i=1}^m f(z_i):$$

Այժմ գնահատենք իրական և էմպիրիկ ռիսկերի տարբերությունը՝

$$\text{Rep}_D(F, S) \equiv \sup_{f \in F}(L_D(f) - L_S(f)):$$

Նշանակենք

$$F \circ S = \{(f(z_1), f(z_2), \dots, f(z_m)), \quad f \in F\}:$$

Այս բազմության ռադեմախերյան բարդությունը հավասար կլինի

$$R(F \circ S) = \frac{1}{m} E_{\sigma \sim \{+1, -1\}} \left[\sup_{f \in F} \sum_{i=1}^m \sigma_i f(z_i) \right]:$$

Պնդում: Ճշմարիտ է հետևյալ անհավասարությունը՝

$$E_{S \sim D^m} [\text{Rep}_D(F, S)] \leq 2 E_{S \sim D^m} R(F \circ S):$$

Ապացույց: Դիցուք $S' = \{z'_1, z'_2, \dots, z'_m\}$ օբյեկտների մի ուրիշ հավաքածու է: Ակնհայտ է, որ $L_D(f) = E_{S'}(L_{S'}(f))$:
Հետևաբար՝

$$L_D(f) - L_S(f) = E_{S'}(L_{S'}(f)) - L_S(f) = E_{S'}(L_{S'}(f) - L_S(f)):$$

Այստեղից կստանանք

$$\sup_{f \in F} (L_D(f) - L_S(f)) = \sup_{f \in F} E_{S'}[L_{S'}(f) - L_S(f)] \leq E_{S'}\left[\sup_{f \in F} (L_{S'}(f) - L_S(f))\right]:$$

Այնուհետև վերցնելով այս անհավասարության երկու մասերի մաթեմատիկական սպասումը ըստ S -ի,՝ կստանանք

$$E_S\left[\sup_{f \in F} (L_D(f) - L_S(f))\right] \leq E_{S,S'}\left[\sup_{f \in F} (L_{S'}(f) - L_S(f))\right] =$$

$$\frac{1}{m} E_{S,S'}\left[\sup_{f \in F} \sum_{i=1}^m (f(z_i) - f(z'_i))\right]:$$

Քանի որ z_j, z'_j պատահական մեծությունները անկախ են և միևնույն բաշխումից, ապա մեկի փոխարինումը մյուսով մաթեմատիկական սպասումը չի փոխում:
Հետևաբար՝

$$\begin{aligned} E_{S,S'}\left[\sup_{f \in F} \left((f(z'_j) - f(z_j)) + \sum_{i \neq j} (f(z'_i) - f(z_i))\right)\right] &= \\ &= E_{S,S'}\left[\sup_{f \in F} \left((f(z_j) - f(z'_j)) + \sum_{i \neq j} (f(z'_i) - f(z_i))\right)\right]: \end{aligned}$$

Դիցուք σ_j այնպիսի պատահական մեծություն է, որ

$$P(\sigma_j = 1) = \frac{1}{2}, \quad P(\sigma_j = -1) = \frac{1}{2}:$$

Այդ դեպքում

$$\begin{aligned} E_{S,S',\sigma_j}\left[\sup_{f \in F} \left(\sigma_j((f(z'_j) - f(z_j)) + \sum_{i \neq j} (f(z'_i) - f(z_i)))\right)\right] &= \\ \frac{1}{2} \left(E_{S,S'}\left[\sup_{f \in F} \left((f(z'_j) - f(z_j)) + \sum_{i \neq j} (f(z'_i) - f(z_i))\right)\right]\right) + \\ + \frac{1}{2} \left(E_{S,S'}\left[\sup_{f \in F} \left((f(z_j) - f(z'_j)) + \sum_{i \neq j} (f(z'_i) - f(z_i))\right)\right]\right): \end{aligned}$$

Կիրառելով այս պրոցեսը ամեն մի j ինդեքսի համար՝ կստանանք

$$\begin{aligned} E_{S,S'}\left[\sup_{f \in F} \sum_{i=1}^m (f(z'_i) - f(z_i))\right] &= E_{S,S',\sigma}\left[\sup_{f \in F} \sum_{i=1}^m \sigma_i (f(z'_i) - f(z_i))\right] \leq \\ &\leq E_{S,S',\sigma}\left[\sup_{f \in F} \sum_{i=1}^m \sigma_i f(z'_i) + \sup_{f \in F} \sum_{i=1}^m -\sigma_i f(z_i)\right] = \end{aligned}$$

$$\begin{aligned}
&= \leq E_{S, S', \sigma} \left[\sup_{f \in F} \sum_{i=1}^m \sigma_i f(z'_i) + \sup_{f \in F} \sum_{i=1}^m \sigma_i f(z_i) \right] = \\
&= m E_{S'} \left[R(F \circ S') \right] + m E_S \left[R(F \circ S) \right] = 2m E_S \left[R(F \circ S) \right]:
\end{aligned}$$

Լեմման ապացուցվեց:

Լեմմա 6.7. *Տեղի ունեն հետևյալ պնդումները՝*

- $E_{S \sim D^m} \left[L_D(ERM_H(S)) - L_S(ERM_H(S)) \right] \leq 2E_{S \sim D^m} R(l \circ H \circ S):$
- $E_{S \sim D^m} \left[L_D(ERM_H(S)) - L_S(h) \right] \leq 2E_{S \sim D^m} R(l \circ H \circ S) \forall h \in H:$
- **Դիցուք** $h^* = \arg \min_{h \in H} L_D(h):$

Այդ դեպքում կամայական $\delta \in (0, 1)$ թվի համար առնվազն $(1 - \delta)$ հավանականությամբ տեղի ունի

$$L_D(ERM_H(S)) - L_S(h^*) \leq \frac{2E_{S \sim D^m} R(l \circ H \circ S)}{\delta}$$

անհավասարությունը:

Այստեղ $ERM_H(S) \in \arg \min_{h \in H} L_S(h):$

Ապացույց: Առաջին պնդումը անմիջապես բխում է նախորդ լեմմայից: Երկրորդը բխում է հետևյալ անհավասարությունից՝

$$L_D(h) = E_S \left[L_S(h) \right] \geq E_S \left[L_S(ERM_H(S)) \right]:$$

Երրորդ անհավասարությունը բխում է երկրորդից՝ կիրառելով Չեբիշևի անհավասարությունը (նկատենք, որ $L_D(ERM_H(S)) - L_D(h^*)$ պատահական մեծությունը ոչ բացասական է):

Լեմմա 6.8. *(Մակդարմիդի անհավասարություն):* Դիցուք ունենք m փոփոխականի $f(x_1, x_2, \dots, x_m) : V^m \rightarrow R$ ֆունկցիա, որը V բազմության վրա բավարարում է հետևյալ պայմանը. գոյություն ունի այնպիսի $c > 0$ թիվ, որ

$$|f(x_1, x_2, \dots, x_m) - f(x_1, x_2, \dots, x_{i-1}, x'_i, x_{i+1}, \dots, x_m)| \leq c, \quad x_1, x_2, \dots, x_m, x'_i \in V:$$

Ենթադրենք նաև, որ $X_1, X_2, \dots, X_m$ անկախ պատահական մեծությունները արժեքներ են ընդունում V բազմությունից: Այդ դեպքում առնվազն $1 - \delta$ հավանականությամբ տեղի ունի

$$|f(X_1, X_2, \dots, X_m) - E[f(X_1, X_2, \dots, X_m)]| \leq c \sqrt{\ln(2/\delta)m/2}$$

անհավասարությունը:

Այստեղից հետևում է հետևյալ թեորեմը, որով գնահատվում է ռեալ և էմպիրիկ ռիսկերի տարբերությունը:

Թեորեմ 6.7. Դիցուք գոյություն ունի այնպիսի $c > 0$, որ $|l(h, z)| \leq c$, $z \in Z, h \in H$: Այդ դեպքում առնվազն $1 - \delta$ հավանականությամբ տեղի ունեն հետևյալ անհավասարությունները՝

- $L_D(h) - L_S(h) \leq 2E_{S' \sim D^m} R(l \circ H \circ S') + c\sqrt{\ln(2/\delta)/m}$, $\forall h \in H$:
- $L_D(h) - L_S(h) \leq 2R(l \circ H \circ S) + c\sqrt{2\ln(4/\delta)/m}$, $\forall h \in H$:
- $L_D(ERM_H(S)) - L_D(h) \leq 2R(l \circ H \circ S) + c\sqrt{2\ln(8/\delta)/m}$:

Ապացույց: Դիտարկենք

$$Rep_D(F, S) = \sup_{f \in F} (L_D(h) - L_S(h))$$

պատահական մեծությունը: Այն բավարարում է նախորդ լեմմայի պայմանները $2c/m$ հաստատունով: Իրոք, ենթադրենք $S'_i = \{z_1, \dots, z_{i-1}, z'_i, z_{i+1}, \dots, z_m\}$ նոր հավաքածու է: Այդ դեպքում

$$\begin{aligned} |Rep_D(F, S) - Rep_D(F, S'_i)| &= \\ |\sup_{f \in F} (L_D(h) - L_S(h)) - \sup_{f \in F} (L_D(h) - L_{S'_i}(h))| &\leq \sup_{h \in F} |L_S(h) - L_{S'_i}(h)| = \\ &= \sup_{h \in H} \left| \frac{1}{m} (l(h, z_1, \dots, z_m) - l(h, z_1, \dots, z_{i-1}, z'_i, z_{i+1}, \dots, z_m)) \right| \leq \frac{2c}{m}: \end{aligned}$$

Այստեղից օգտվելով Մակդարմիդի անհավասարությունից և վերևում ապացուցված պնդման եզրակացությունից, $1 - \delta$ հավանականությամբ տեղի ունի

$$\begin{aligned} Rep_D(F, S) &\leq ERep_D(F, S) + c\sqrt{2\ln(2/\delta)/m} \leq \\ &\leq 2E_{S'} R(l \circ H \circ S') + c\sqrt{2\ln(2/\delta)/m} \end{aligned}$$

անհավասարությունը:

Այստեղից էլ անմիջապես հետևում է առաջին անհավասարությունը: Կիրառելով Մակդարմիդի անհավասարությունը $(l \circ H \circ S)$ պատահական մեծության նկատմամբ և հաշվի առնելով առաջին անհավասարությունը՝ կստանանք երկրորդը: Վերջապես, երրորդ անհավասարության ապացույցի համար նկատենք, որ $h_S = ERM_H(S)$ հիպոթեզի և կամայական $h, h^* \in H$ հիպոթեզների համար տեղի ունի հետևյալ անհավասարությունը՝

$$\begin{aligned} L_D(h_S) - L_D(h^*) &= \\ L_D(h_S) - L_S(h_S) + L_S(h_S) - L_S(h^*) + L_S(h^*) - L_D(h^*) &= \\ &\leq (L_D(h_S) - L_S(h_S)) + (L_S(h^*) - L_D(h^*)): \end{aligned}$$

Այս անհավասարության աջ մասի առաջին գումարելին գնահատենք թեորեմի երկրորդ անհավասարությամբ: Իսկ երկրորդ գումարելին, համաձայն Հոֆմանի

անհավասարության՝ առնվազն $1 - \delta/2$ հավանականությամբ տեղի ունի

$$L_S(h^*) - L_D(h^*) \leq c\sqrt{2\ln(4/\delta)/m}$$

անհավասարությունը: Միավորելով այս երկուսը՝ կստանանք թեորեմի երրորդ անհավասարությունը: Նշենք, որ թեորեմի առաջին երկու անհավասարությունները տեղի ունեն $h = EMR_H(S)$ հիպոթեզի համար:

Թեորեմն ապացուցվեց:

Թեորեմ 6.8. Դիցուք ունենք հավանականային D բաշխում $X \times Y$ արտադրյալի վրա և 1-հավանականությամբ տեղի ունի $\|x\|_2 \leq R$ անհավասարությունը: Դիցուք

$$H = \{w : \|w\|_2 \leq B\}$$

և կորստի $l : H \times Z \rightarrow R$ ֆունկցիան ունի $l(h, (x, y)) = \varphi((w, x), y)$ տեսքը, որտեղ ամեն մի ֆիքսված $y \in Y$ դեպքում $a \rightarrow \varphi(a, y)$ ֆունկցիան ρ -լիպշիցյան է և $\max_{a \in [-BR, BR]} |\varphi(a, y)| \leq c$: Այդ դեպքում կամայական $\delta \in (0, 1)$ թվի համար առնվազն $1 - \delta$ հավանականությամբ տեղի ունեն հետևյալ անհավասարությունները՝

- $L_D(w) \leq L_S(w) + \frac{2\rho BR}{\sqrt{m}} + c\sqrt{2\ln(2/\delta)/m}$:
- $L_D(EMR_H(h_S)) \leq \inf_{h \in H} L_D(h) + \frac{2\rho BR}{\sqrt{m}} + c\sqrt{2\ln(8/\delta)/m}$:

Ապացույց: Դիցուք

$$F = \{(x, y) \rightarrow \varphi((w, x), y), w \in H\}:$$

Այդ դեպքում

$$F \circ S = \{\varphi((w, x_1), y_1) \dots \varphi((w, x_m), y_m)\}, w \in H:$$

Համաձայն լեմմաներ 6.4-6.6, 1-հավանականությամբ կունենանք

$$R(F \circ S) \leq \frac{\rho BR}{\sqrt{m}}:$$

Այնուհետև կիրառելով նախորդ թեորեմ 6.7-ի երկրորդ և երրորդ անհավասարությունները՝ կստանանք թեորեմի անհրաժեշտ պնդումները:

Այժմ գնահատենք, թե ինչ երկարությամբ հավաքածու վերցնենք, որ անհավասարությունների աջ մասերը դառնան նախօրոք վերցրած $\epsilon > 0$ թվից փոքր:

Ճշմարիտ է

$$\begin{aligned} \frac{2\rho BR}{\sqrt{m}} + c\sqrt{2\ln(8/\delta)/m} &\leq \\ &\leq \sqrt{\frac{4(\rho BR)^2 + 2c^2 \ln(8/\delta)}{m}} \end{aligned}$$

անհավասարությունը: Իրոք, եթե $a, b > 0$, ապա

$$\sqrt{a} + \sqrt{b} \leq 2\sqrt{a+b} \iff \sqrt{ab} \leq 3(a+b)$$

$$\iff 9a^2 - 17ab + 9b^2 \geq 0:$$

Վերջին անհավասարությունը ակնհայտ է: Այսպիսով

$$2\sqrt{\frac{4(\rho BR)^2 + 2c^2 \ln(8/\delta)}{m}} \leq \epsilon$$

անհավասարությունից կարող ենք որոշել պահանջվող հավաքածուի երկարությունը: Այստեղից կստանանք

$$m \geq \frac{4(\rho BR)^2 + 2c^2 \ln(8/\delta)}{\epsilon^2}:$$

Թեորեմն ապացուցվեց:

Դիցուք h_S -ը հետևյալ խնդրի լուծումն է՝

$$\arg \min_w \|w\|^2, \quad y_i(w, x_i) \geq 1, \quad i = 1, 2, \dots, m:$$

Թեորեմ 6.9. Դիցուք D հավանականային բաշխումը $Z = X \times \{+1, -1\}$ բազմության վրա այնպիսին է, որ գոյություն ունի այնպիսի w^* վեկտոր, որ

$$P_{(x,y) \sim D}[y(w^*, x) \geq 1] = 1$$

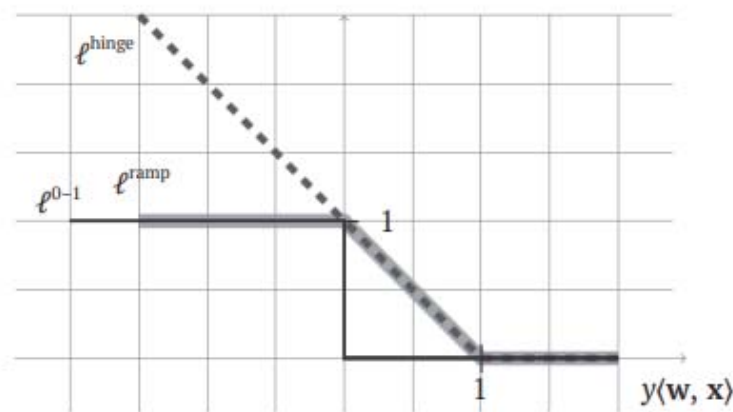
և 1-հավանականությամբ տեղի ունի $\|x\|_2 \leq R$ անհավասարությունը: Այդ դեպքում ցանկացած $\delta \in (0, 1)$ թվի համար առնվազն $1 - \delta$ հավանականությամբ տեղի ունի հետևյալ անհավասարությունը՝

$$P_{(x,y) \sim D}[y \neq \text{sign}(w_S, x)] \leq \frac{2R\|w^*\|}{\sqrt{m}} + \sqrt{\frac{2 \ln(2/\delta)}{m}}:$$

Ապացույց: Ապացույցի համար որպես կորստի ֆունկցիա վերցնենք հետևյալ ֆունկցիան՝

$$l^{\text{ramp}}(w, (x, y)) \equiv \min\{1, \max\{0, 1 - y(w, x)\}\}:$$

Նրա գրաֆիկը պատկերված է 6.12 նկարում: Այս ֆունկցիայի արժեքների



Նկ. 6.12: l^{ramp} ֆունկցիայի գրաֆիկը

բազմությունը $[0, 1]$ հատվածն է և 1-լիպշիցյան է: Եվ քանի որ այն վերևից մոտարկում է հենքային ֆունկցիային, ապա

$$P_{(x,y) \sim D}[y \neq \text{sign}(w_S, x)] \leq L_D(w_S):$$

Դիցուք $B = \|w^*\|_2$: դիտարկենք հիպոթեզների $H = \{w : \|w\|_2 \leq B\}$ բազմությունը: Անհայտ է, որ 1- հավանականությամբ $w_S \in H$ և $L_S(w_S) = 0$: Հետևաբար, ըստ նախորդ թեորեմի,

$$L_D(w_S) \leq L_S(w_S) + \frac{2\rho BR}{\sqrt{m}} + \sqrt{2 \ln(2/\delta)/m}:$$

Թեորեմն ապացուցվեց::

Այժմ գնահատենք ալգորիթմի բարդությունը: Պահանջենք, որ

$$\frac{2\rho BR}{\sqrt{m}} + \sqrt{2 \ln(2/\delta)/m} < \epsilon:$$

Այստեղից հետևում է, որ ալգորիթմի $m(\delta, \epsilon)$ բարդությունը մեծ է $\mathbb{R}^2 \|w^*\|^2 / (\epsilon)^2$: Ինչպես տեսնում ենք, այդ բարդությունը կախված է w^* վեկտորից, որը հայտնի չէ: Այդ իսկ պատճառով անհրաժեշտ է ստանալ բարդության նոր գնահատական: Դրա հնարավորությունը տալիս է հետևյալ թեորեմը:

Թեորեմ 6.10. Դիցուք տեղի ունեն թեորեմ 6.9-ի պայմանները: Այդ դեպքում առնվազն $1 - \delta$ հավանականությամբ կամայական $S \sim D^m$ հավաքածուի համար տեղի ունի հետևյալ անհավասարությունը՝

$$P_{(x,y) \sim D}[y \neq \text{sign}(w_S, x)] \leq \frac{4R\|w_S\|}{\sqrt{m}} + \sqrt{\frac{\ln(4 \log_2(\|w_S\|/\delta))}{m}}:$$

Ապացույց: Կամայական ամբողջ i -ի համար նշանակենք

$$B_i = 2^i, \quad H_i = \{w : \|w\|_2 \leq B_i\}, \quad \delta_i = \frac{\delta}{4i^2}:$$

Յուրաքանչյուր ֆիքսված i համար կիրառենք թեորեմ 6.8-ի անհավասարությունը, կստանանք՝

$$L_D(w) \leq L_S(w) + \frac{2\rho B_i R}{\sqrt{m}} + \sqrt{2 \ln(2/\delta)_i/m}:$$

Քանի որ $\sum_{i=1}^{\infty} \delta_i \leq \delta$, ապա վերևի անհավասարությունը տեղի ունի առնվազն $1 - \delta$ հավանականությամբ: Ամեն մի ֆիքսած w վեկտորի համար վերցնենք $i = \lceil \log_2 \|w\| \rceil$: Այդ դեպքում կունենանք

$$w \in H_i, \quad B_i \leq 2\|w\|, \quad 2/\delta_i = \frac{4i^2}{\delta} \leq (4 \log_2(\|w\|))^2 / \delta:$$

Հետևաբար՝

$$\begin{aligned} L_D(w) &\leq L_S(w) + \frac{2B_i R}{\sqrt{m}} + \sqrt{\frac{2 \ln(2/\delta_i)}{m}} \leq \\ &\leq L_S(w) + \frac{4||w||R}{\sqrt{m}} + \sqrt{\frac{4(\ln(4 \log_2(||w||) + \ln(1/\delta)))}{m}}. \end{aligned}$$

Մասնավորապես, տեղադրելով այստեղ w_S վեկտորը՝ կստանանք թեորեմի պնդումը:

Վերջում նշենք, որ այս ալգորիթմի բարդությունը բացահայտորեն կախված չէ w_S վեկտորի չափողականությունից: Թեորեմն ապացուցվեց:

7

Նեյրոնային ցանցերի ուսուցում

Նեյրոնային ցանցերը կատարում են ինֆորմացիայի ձևափոխություն: Նեյրոնները լինում են երկու տիպի՝

- Դենդրիտներ (dendrit), որոնց վրա գալիս է մուտքային ինֆորմացիա: Դրանք համարակալենք բնական թվերով՝ $1, 2, \dots, n$:
- Ակսոններ (aksion), դրանցով անցնում են ելքային ազդանշաններ:

Նեյրոնների աշխատանքի սկզբունքը հետևյալն է: Յուրաքանչյուր դենդրիտի հետ կապված է մի թիվ, որը կոչվում է նրա կշիռ: Նշանակենք դրանք հետևյալ կերպ՝ $w_1, w_2, \dots, w_n$: Կա ևս մի թիվ, որը կոչվում է գրգռման շեմ՝ w_0 : Դիցուք x_i ինտենսիվության ազդանշանները գալիս են i -րդ դենդրիտի վրա, նեյրոնի կենտրոնական մասում առաջացնում են $\sum_{i=1}^n w_i x_i$ ինտենսիվությամբ նոր ազդանշան:

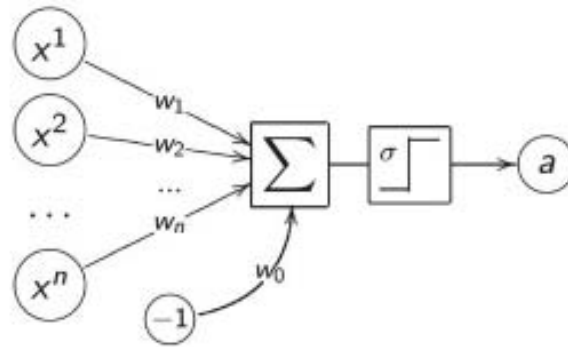
- Եթե այդ ինտենսիվությունը գերազանցում է նեյրոնի գրգռման շեմը, ապա նեյրոնը ակսոնի միջոցով բաց է թողնում էլեկտրական ազդանշան որոշակի ինտենսիվությամբ,
- հակառակ դեպքում այդ պրոցեսը տեղի չի ունենում:

Այսպիսով նեյրոնը կարող է դիտարկվել որպես թվային ինֆորմացիայի մշակման օբյեկտ, որի մուտքին տրվում է $(x^1, \dots, x^n)$ վեկտոր, իսկ ելքում ստացվում է $\sigma(\sum_{i=1}^n w_i x_i - w_0)$ հզորությամբ ազդանշան: Այստեղ σ -ն կոչվում է ակտիվության ֆունկցիա: Այն կարող է լինել տարբեր: Օրինակ՝

- $\sigma(x) = \frac{1}{1+e^{-x}}$,
- $\sigma(x) = 0, x < 0, \sigma(x) = x, x \geq 0$ (RELU):

Նկար 7.1-ում պատկերված է պարզագույն նեյրոնային ցանց: Այժմ ցույց տանք, որ նեյրոնային ցանցերի միջոցով կարելի է ռեալիզացնել բուլյան ֆունկցիաները: Նկար 7.2-ում տրված է գումարի, բազմապատկման բուլյան գործողությունների իրականացումը մեկ նեյրոնի միջոցով:

Իսկ նկար 7.3-ում՝ բուլյան հանրահաշվական գումարի իրականացումը երկ-շերտ նեյրոնային ցանցի միջոցով:

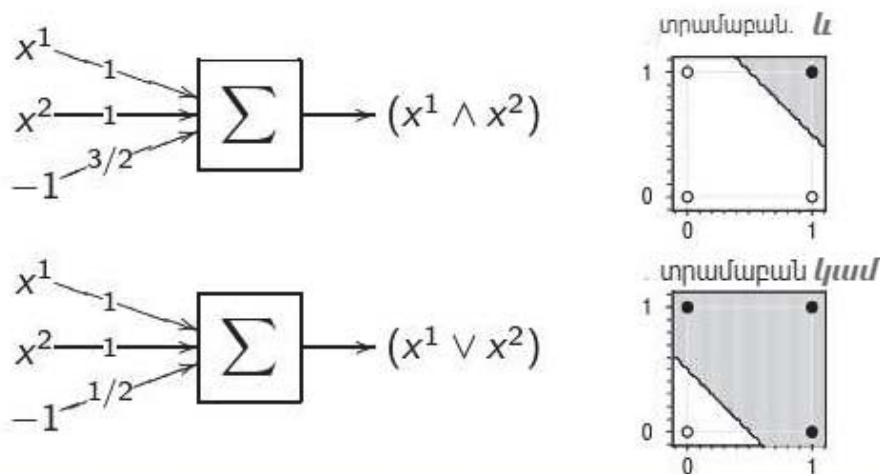


Նկ. 7.1: Պարզագույն նեյրոնային ցանց

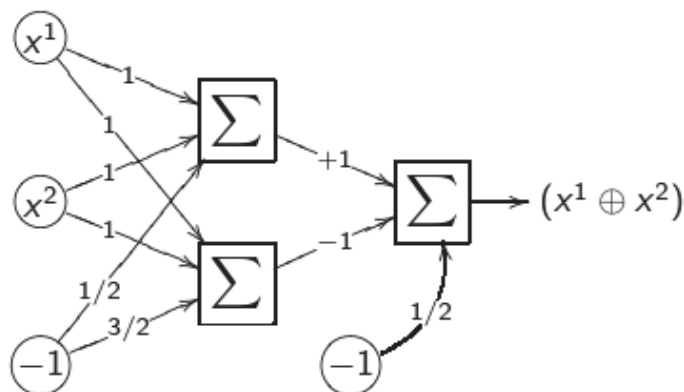
$$x^1 \wedge x^2 = [x^1 + x^2 - \frac{3}{2} > 0]$$

$$x^1 \vee x^2 = [x^1 + x^2 - \frac{1}{2} > 0]$$

$$\neg x^1 = [-x^1 + \frac{1}{2} > 0]$$



Նկ. 7.2: Բուլյան ֆունկցիաների իրականացումը նեյրոնային ցանցերով



Նկ. 7.3: $x_1 \oplus x_2 = [(x_1 \vee x_2) - (x_1 \wedge x_2) - 1/2 > 0]$

7.1 Սխալանքի ետ տարածման մեթոդը (Error back propagation method)

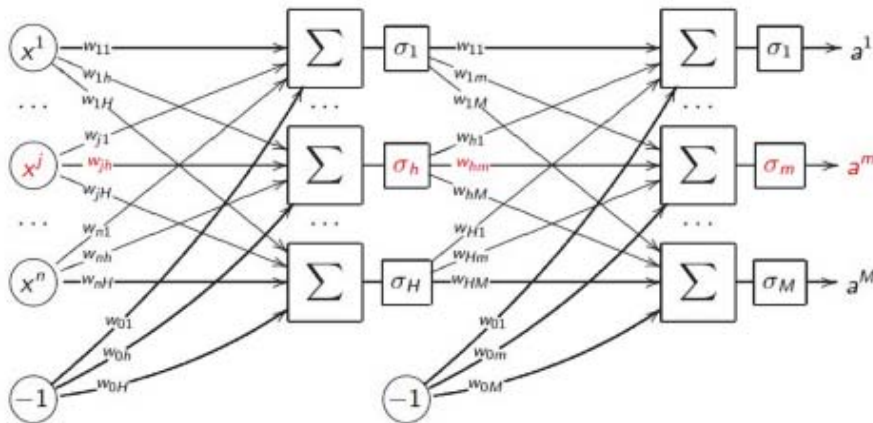
Սխալանքի ետ տարածման մեթոդը բազմաշերտ նեյրոնային ցանցերի ուսուցման լայն տարածված եղանակներից է: Այն գրադիենտային իջեցման մի մոդիֆիկացիա է: Հիշեցնենք, որ բազմաշերտ նեյրոնային շերտի կոմպոնենտները նեյրոններ են:

- Նեյրոնների մուտքին տրվում է $(x_1, x_2, \dots, x_n) \in \mathbb{R}^n$ վեկտորը,
- նեյրոնը ելքում տալիս է

$$a = \sigma((x, w) - w_0)$$

թիվը:

Նկար 7.4-ում պատկերված է երկշերտ նեյրոնային ցանց:



Նկ. 7.4: Երկշերտ նեյրոնային ցանց

Այս ցանցը կշռային գործակիցների w_{ij} ֆիքսված արժեքների դեպքում որոշում է a_w ֆունկցիա, որը մուտքի $x \in \mathbb{R}^n$ վեկտորին համապատասխանության մեջ է դնում ելքի $a_w(x) \in \mathbb{R}^M$ վեկտորը: Եթե տրված է $S \subset \mathbb{R}^n \times \mathbb{R}^M$ ուսուցանող հավաքածուն, ապա նեյրոնային ցանցի սխալանքը $(x, y) \in S$ գույգի վրա հետևյալ թիվն է՝

$$Q(x, y, w) = \frac{1}{2} \|a_w(x) - y\|^2:$$

Սխալանքի ետ տարածման ալգորիթմը գտնում է w_{ij} կշռային գործակիցների այնպիսի արժեքներ, որ ցանցի վերը նշված սխալանքը լինի հնարավորինս քիչ:

Ալգորիթմը դա անում է մի քանի իտերացիաների ընթացքում: Յուրաքանչյուր իտերացիա բաղկացած է հետևյալ քայլերից՝

- ընտրվում է (x, y) գույգը հավաքածուից,
- կատարվում է ցանցի սխալանքի հաշվարկ ձախից աջ w_{ij} կշիռների ընթացիկ արժեքների դեպքում,

- այն հուշեցնում է կշռային արժեքների վերահաշվարկ հսկանական ուղղությամբ՝ աջից ձախ (սկզբից կատարվում է վերջին շերտի կշիռների վերահաշվարկ, այնուհետև՝ նախավերջին շերտի և այսպես շարունակ):

Այժմ ավելի մանրամասնորեն քննարկենք այդ վերահաշվարկի կանոնները նկար 7.5 և 7.6-ում պատկերված նեյրոնային ցանցերի օրինակներով: Այստեղ որպես ակտիվացման ֆունկցիա ընտրված է սիգմոյիդը՝

$$\sigma(x) = \frac{1}{1 + e^{-x}}:$$

Ալգորիթմում օգտագործվում է հետևյալ ակնհայտ հավասարությունը՝

$$\sigma'(x) = \sigma(x)(1 - \sigma(x)):$$

Միավանքի ետ տարածման ընթացակարգը հետևյալն է:

- Կատարվում է կշիռների ինիցիալիզացիա, այսինքն՝ դրանց տրվում է պատահական արժեքներ:
- Պատահաանորեն վերցնում ենք $(x, y) \in S$:
- **Ուղիղ քայլ:** Հաշվում ենք բոլոր նեյրոնների ելքերը, և

$$Q(x, y, w) = \frac{1}{2} \sum_{m=1}^M (a^m - y^m)^2,$$

$$\frac{\partial Q}{\partial w_{hm}} = a^m - y^m \equiv \xi^m, \quad m = 1, 2, \dots, M:$$

- **Հետադարձ քայլ (կշիռների վերահաշվարկ):**

$$w'_{hm} = w_{hm} - \eta \frac{\partial Q}{\partial w_{hm}} \eta,$$

$$w'_{jh} = w_{jh} - \frac{\partial Q}{\partial w_{jh}} \eta,$$

որտեղ $\eta \in (0, 1)$ պարամետր է, իսկ $\frac{\partial Q}{\partial w_{hm}}$ և $\frac{\partial Q}{\partial w_{jh}}$ ածանցյալները հաշվվում են հետևյալ բանաձևերով: Դիցուք $u^1, u^2, \dots, u^H$ -ը առաջին շերտի ելքերն են: Այդ դեպքում

$$\forall m = 1, 2, \dots, M, \quad h = 1, 2, \dots, H, \quad j = 1, 2, \dots, n$$

$$a^m = \sigma\left(\sum_{h=1}^H w_{hm} u^h - w_{0m}\right),$$

$$\begin{aligned} \frac{\partial Q}{\partial w_{hm}} &= \frac{\partial Q}{\partial a^m} \frac{\partial a^m}{\partial w_{hm}} = \xi^m \sigma'\left(\sum_{h=1}^H w_{hm} u^h - w_{0m}\right) u^h = \\ &= \xi^m a^m (1 - a^m) u^h, \end{aligned}$$

$$\begin{aligned}\frac{\partial Q}{\partial w_{0m}} &= \frac{\partial Q}{\partial a^m} \frac{\partial a^m}{\partial w_{0m}} = \xi^m \sigma' \left(\sum_{h=1}^H w_{hm} u^h - w_{0m} \right) (-1) = \\ &= -\xi^m a^m (1 - a^m),\end{aligned}$$

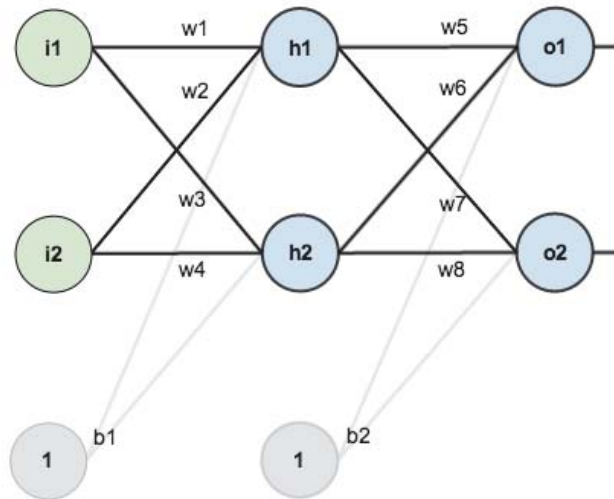
$$u^h = \sigma \left(\sum_{j=1}^n w_{jh} x^j - w_{0h} \right),$$

$$\begin{aligned}\frac{\partial Q}{\partial u^h} &= \sum_{m=1}^M \frac{\partial Q}{\partial a^m} \frac{\partial a^m}{\partial u^h} = \sum_{m=1}^M \xi^m \sigma' \left(\sum_{h=1}^H w_{hm} u^h - w_{0m} \right) w_{hm} = \\ &= \sum_{m=1}^M \xi^m a^m (1 - a^m) w_{hm} \equiv \zeta^h,\end{aligned}$$

$$\begin{aligned}\frac{\partial Q}{\partial w_{jh}} &= \frac{\partial Q}{\partial u^h} \frac{\partial u^h}{\partial w_{jh}} = \zeta^h \sigma' \left(\sum_{j=1}^n w_{jh} x^j - w_{0h} \right) x^j = \\ &= \zeta^h u^h (1 - u^h) x^j,\end{aligned}$$

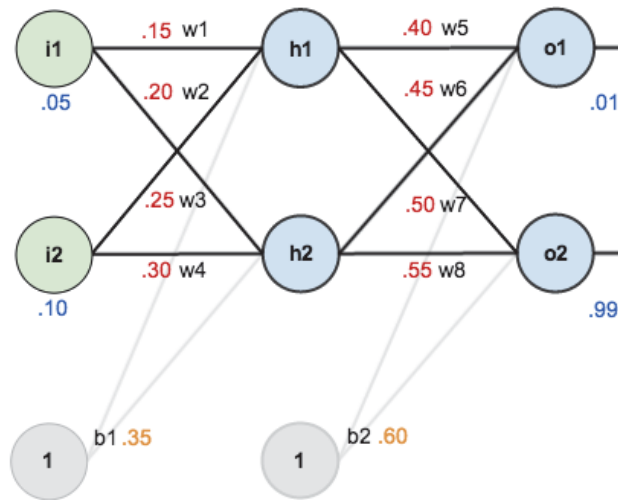
$$\begin{aligned}\frac{\partial Q}{\partial w_{0h}} &= \frac{\partial Q}{\partial u^h} \frac{\partial u^h}{\partial w_{0h}} = \zeta^h \sigma' \left(\sum_{j=1}^n w_{jh} x^j - w_{0h} \right) (-1) = \\ &= -\zeta^h u^h (1 - u^h): \end{aligned}$$

Դիտարկենք նեյրոնային ցանց, որի սխեման պատկերված է նկար 7.5-ում:



Նկ. 7.5: Նեյրոնային ցանց

Այն բաղկացած է երկու (I_1, I_2) մուտքերից, (h_1, h_2) թաքնված նեյրոններից և (O_1, O_2) ելքերից: Այս մեթոդի հիմնական նպատակը այն է, որ կշիռների օպտիմիզացիայի միջոցով ցանցը կամայական մուտքային տվյալների վրա ստանա ելքում տված տվյալները: Ելքային տվյալներն են՝ (0.05, 0.10). պահանջվում է, որ նեյրոնային ցանի ելքում լինի (0.01, 0.99) վեկտորը (տե՛ս նկ.7.6):



Նկ. 7.6: Նախնական տվյալներ

Ազդանշանների ուղիղ տարածում: Ներդրնային ցանցի աշխատանքը սկսվում է ներդրնային ցանցի շերտերով մուտքային տվյալների ուղիղ տարածումից: Այն կատարվում է հետևյալ ընթացակարգով.

- հաշվարկվում է h_1 թաքնված ներդրնի մուտքին տրված ազդանշանի մեծությունը՝

$$net(h_1) = w_1 i_1 + w_2 i_2 + b_1 = 0.15 \times 0.05 + 0.2 \times 0.1 + 0.35 = 0.3775,$$

- h_1 ներդրնի էլքի ազդանշանը ստանալու համար կիրառվում է ակտիվացման ֆունկցիան՝

$$Out(h_1) = \frac{1}{1 + e^{-net(h_1)}} = 0.593269992,$$

- կատարելով նման գործողություններ h_2 ներդրնի համար՝ կստանանք

$$out(h_2) = 0.596884378:$$

Հաջորդ փուլում պետք է կատարել ազդանշանի տարածում թաքնված շերտից դեպքի էլք:

- o_1 ներդրնի նկատմամբ կստանանք հետևյալ արդյունքները՝

$$net(o_1) = w_5 out(h_1) + w_6 out(h_2) + b_2 = 1.105905967,$$

$$out(o_1) = \frac{1}{e^{-net(o_1)}} = 0.75136507:$$

- Կատարելով նույն գործողությունները o_2 ներդրնի նկատմամբ՝ կստանանք $out(o_2) = 0.772928465$:

Ընդհանուր սխալանքի հաշվում:

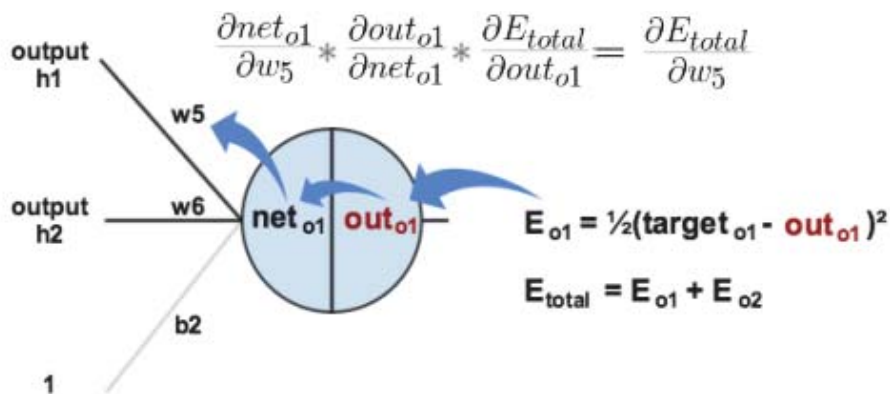
$$E(\text{total}) = E(o_1) + E(o_2) = \frac{1}{2}((\text{target}(o_1) - \text{out}(o_1))^2 + (\text{target}(o_2) - \text{out}(o_2))^2 =$$

$$= \frac{1}{2}(0.01 - 0.75136507)^2 + (0.99 - 0.772928465)^2 = 0.298371109:$$

Սխալանքի ետ տարածում: Ներդրային ցանցի ելքային մասում w_5, w_6 կշիռները կապված են o_1 ներդրմի հետ, իսկ w_7, w_8 կշիռները՝ o_2 ներդրմի հետ: Բացի դրանից՝ այդ երկու ներդրումների վրա ազդում է նաև b_2 շեղումը: Կշիռների օպտիմիզացիայի պրոցեսը սկսենք w_5 կշռից: Որպեսզի իմանանք թե ինչպես է ազդում w_5 կշռի փոփոխությունը ընդհանուր սխալանքի վրա, պետք է հաշվենք $E(\text{total})$ -ի մասնակի ածանցյալը ըստ w_5 փոփոխականի: Կիրառելով բարդ ֆունկցիայի ածանցման կանոնը՝ կստանանք:

$$\frac{\partial E(\text{total})}{\partial w_5} = \frac{\partial E(\text{total})}{\partial (\text{out}(o_1))} \frac{\partial (\text{out}(o_1))}{\partial (\text{net}(o_1))} \frac{\partial (\text{net}(o_1))}{\partial w_5}:$$

Սա կարելի է ներկայացնել նկար 7.7-ի սխեմայով: Հաշվենք բանաձևում առկա բոլոր մասնակի ածանցյալները:



Նկ. 7.7: Բարդ ֆունկցիայի ածանցման կանոնը

•

$$\frac{\partial E(\text{total})}{\partial (\text{out}(o_1))} = -2 \cdot \frac{1}{2}(\text{target}(o_1) - \text{out}(o_1)) = 0.74136507:$$

• Քանի որ

$$\text{out}(o_1) = \frac{1}{1 + e^{-\text{net}(o_1)}},$$

ապա

$$\frac{\partial \text{out}(o_1)}{\partial \text{net}(o_1)} = \text{out}(o_1)(1 - \text{out}(o_1)) = 0.186815602:$$

• Եվ վերջապես՝ քանի որ $\text{net}(o_1) = w_5 \text{out}(h_1) + w_6 \text{out}(h_2) + b_2$, ապա

$$\frac{\partial \text{net}(o_1)}{\partial w_5} = \text{out}(h_1) = 0.593269992:$$

- Տեղադրելով այս արժեքները ընդհանուր բանաձևի մեջ՝ կստանանք՝

$$\frac{\partial E(total)}{\partial w_5} = 0.082167041:$$

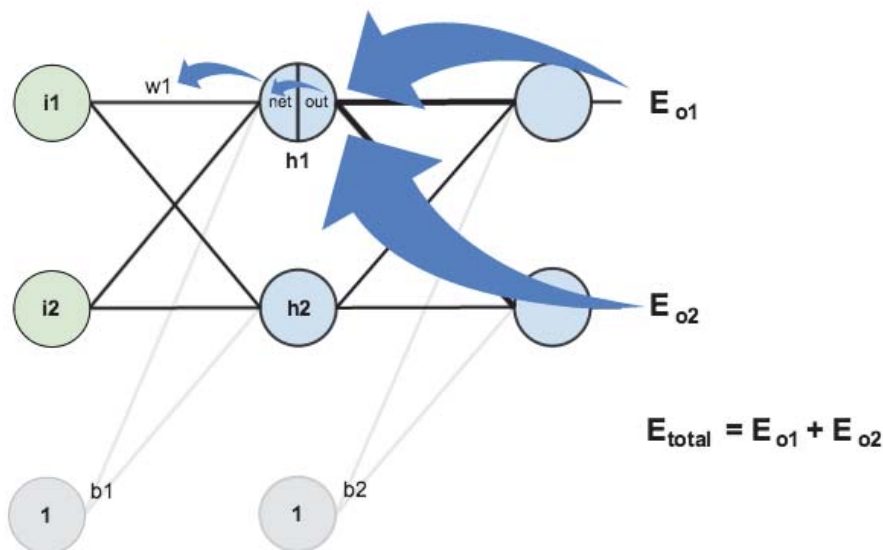
Այսպիսով կիրառելով գրադիենտային իջեցումը w_5 կշռի նկատմամբ՝ կստանանք w_5 -ի նոր արժեք՝

$$w_5^+ = w_5 - \eta \frac{\partial E(total)}{\partial w_5} = 0.35891648, \eta = 1/2:$$

Անալոգիայով կհաշվենք մնացած կշիռների նոր արժեքները՝

- $w_6^+ = 0.408666186,$
- $w_7^+ = 0.511301207,$
- $w_8^+ = 0.561370121:$

Թաքնված շերտի կշռայի գործակիցների թարմացում (նկ. 7.8):



Նկ. 7.8: Թաքնված շերտի կշիռների թարմացում

- Քանի որ $out(h_1)$ -ը ազդում է $out(o_1)$ -ի և $out(o_2)$ -ի վրա, ապա

$$\frac{\partial E(total)}{\partial out(h_1)} = \frac{\partial E(o_1)}{\partial out(h_1)} + \frac{\partial E(o_2)}{\partial out(h_1)}:$$

- Հաշվենք այդ գումարելիները առանձին-առանձին՝

$$\begin{aligned} \frac{\partial E(o_1)}{\partial out(h_1)} &= \frac{\partial E(o_1)}{\partial net(o_1)} \frac{\partial net(o_1)}{\partial out(h_1)} = \\ &= \frac{\partial E(o_1)}{\partial out(o_1)} \frac{\partial out(o_1)}{\partial net(o_1)} \frac{\partial net(o_1)}{\partial out(h_1)} = 0.055399425: \end{aligned}$$

- Կատարելով նման հաշվարկներ $\frac{\partial E(o_2)}{\partial out(h_1)}$ գումարելու նկատմամբ՝ կստանանք՝

$$\frac{\partial E(o_2)}{\partial out(h_1)} = -0.019049119:$$

Արդյունքում կստանանք՝

$$\frac{\partial E(total)}{\partial w_1} = 0.000438568:$$

Այստեղից w_1 -ի թարմացված նոր արժեքը կլինի

$$w_1^+ = w_1 - \eta \frac{\partial E(total)}{\partial w_1} = 0.149780716:$$

Անալոգիայով կստանանք՝

- $w_2^+ = 0.19956143,$
- $w_3^+ = 0.24975114,$
- $w_4 = 0.29950229:$

Ինչպես տեսնում ենք, մեկ իտերացիայի ժամանակ սխալանքը 0.298371109-ից նվազեց՝ դառնալով 0.291027924: Ծրագրային իրականացման միջոցով կարելի է համոզվել, որ 10000 իտերացիայից հետո սխալանքը կնվազի մինչև 0.0000351085:

Վերջում բերենք որոշ ընդհանուր թեորեմներ:

- **Երկշերտ նեյրոնային ցանցով կարելի է իրականացնել կամայական բույյան ֆունկցիա:**
- **Ցիբենկոյի թեորեմ 1989: Դիցուք $f : [0, 1]^n \rightarrow R$ -ը անընդհատ ֆունկցիա է: Այդ դեպքում կարելի է ընտրել H , α_H , w_H պարամետրերի այնպիսի արժեքներ, որ երկշերտ նեյրոնային շերտով**

$$a(x) = \sum_{h=1}^H \alpha_{\sigma}((w, h, x) - w_0)$$

ֆունկցիան հավասարաչափ մոտարկի $f(x)$ -ը ցանկացած $\epsilon > 0$ ճշտությամբ՝

$$|f(x) - a(x)| < \epsilon, \quad \forall x \in [0, 1]^n,$$

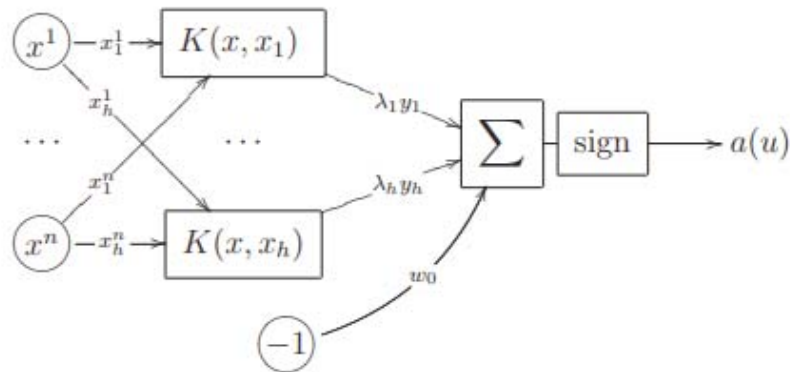
որտեղ σ -ն սիգմոյիդի ֆունկցիան է: (նկ. 7.9-ը երկշերտ նեյրոնային ցանց է:)

7.2 SVD ալգորիթմը և երկշերտ նեյրոնային ցանցեր

Հենման վեկտորների մեթոդում որոշիչ ֆունկցիան ունի հետևյալ տեսքը՝

$$a_S(x) = \text{sign}\left(\sum_{i=1}^h \lambda_i y_i K(x_i, x)\right):$$

Այստեղ ենթադրվում է, որ եթե x_i օբյեկտը ոչ հենքային է, ապա $\lambda_i = 0$: $a_S(x)$ ալգորիթմը կարելի է համարել որպես երկշերտ նեյրոնային ցանց, որն ունի n մուտքային նեյրոն և h նեյրոն թաքնված շերտում (տե՛ս նկ. 7.9):



Նկ. 7.9: SVD-ն որպես երկշերտ նեյրոնային ցանց

SVM մեթոդով կազմաձևված ցանցն ունի մի քանի ուշագրավ առանձնահատկություններ:

- Թաքնված շերտում նեյրոնների թիվը որոշվում է ավտոմատ կերպով:
- Թաքնված շերտի նեյրոնների քաշային վեկտորները համընկնում են հենման օբյեկտների հատկանիշների հետ:
- λ_i , $i = 1, 2, \dots, l$ երկակի փոփոխականները արտահայտում են $K(x_i, x)$, $i = 1, \dots, l$ միջուկների կարևորության աստիճանը:

Առաջադրանք: Գրել վերևում նկարագրված սխալանքի ետ տարածման ալգորիթմի կոդը *Python*-ով և ուսուցանել նկ. 7.6-ի նեյրոնային ցանցը:

8

Մեքենայական ուսուցման հավանականային մոդելներ

8.1 Պայմանական հավանականություններ, Բայեսի քանաձևը

Ինչպես հայտնի է A պատահարի պայմանական հավանականություն B պայմանով, կոչվում է հետևյալ մեծությունը՝

$$P(A|B) = \frac{P(A \cap B)}{P(B)}:$$

Այս սահմանումից հետևում է, որ եթե պատահարները անկախ են, ապա

$$P(A|B) = P(A):$$

Դիտարկենք պայմանական հավանականությունների հաշվման մի քանի օրինակներ:

Օրինակ 1: Ունենք 12 գնդիկներ, որոնցից 5-ը սև, 7-ը սպիտակ: Դրանք համարակալված են հերթով: Պատահական ձևով հանվում է մի գնդիկ: Գտնել հավանականությունը, որ հանված գնդիկը սև է՝ իմանալով, որ այն ունի զույգ համար: Դիցուք A -ն այն պատահարն է, որ հանված է սև գնդիկ, իսկ B -ն՝ հանված է զույգ համարով գնդիկ: Այդ դեպքում կունենանք

$$P(A \cap B) = \frac{2}{12} = \frac{1}{6}, \quad P(B) = \frac{1}{2}:$$

Հետևաբար՝

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1}{3}:$$

Այժմ քննարկենք լրիվ հավանականությունների քանաձև: Դիցուք ունենք իրարից անկախ $B_1, \dots, B_n$ պատահարների լրիվ համակարգ՝

$$B_1 \cup B_2 \dots \cup B_n = \Omega:$$

Այդ դեպքում տեղի ունի հետևյալ բանաձևը՝

$$P(A) = \sum_{i=1}^n P(A|B_i)P(B_i):$$

Ունենք՝

$$A = A \cap \Omega = A \cap \left(\bigcup_{i=1}^n B_i\right) = \bigcup_{i=1}^n (A \cap B_i):$$

Այստեղից, քանի որ $A \cap B_i$, $i = 1, 2, \dots, n$ պատահարները նույնպես անկախ են, կստանանք

$$P(A) = \sum_{i=1}^n P(A \cap B_i) = \sum_{i=1}^n P(A|B_i)P(B_i):$$

Օրինակ 2: Դիցուք ունենք երեք արկղ, որոնցում կան համապատասխանաբար առաջինում՝ երեք սպիտակ և չորս սև գնդիկ, երկրորդում երկու սպիտակ և հինգ սև գնդիկ, երրորդում՝ տասը սև գնդիկ: Պատահականորեն ընտրվում է արկղ և հանվում է գնդիկ:

Գտնել հավանականությունը, որ գնդիկը լինի սպիտակ: Դիցուք B_i -ն այն պատահարն է, որ ընտրված է i -րդ արկղը, իսկ A -ն՝ հանված է սպիտակ գնդիկ: Ըստ տվյալների կունենանք՝

$$P(B_i) = \frac{1}{3}; i = 1, 2, 3, P(A|B_1) = \frac{3}{7}, P(A|B_2) = \frac{2}{7}, P(A|B_3) = 0:$$

Այստեղից և լրիվ հավանականությունների բանաձևից կստանանք

$$P(A) = \frac{1}{3} \cdot \frac{3}{7} + \frac{1}{3} \cdot \frac{2}{7} = 0.238:$$

Լրիվ հավանականությունների բանաձևից ստանում ենք հետևյալ հավասարությունը, որը կոչվում է **Քայեսի բանաձև**՝

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{i=1}^n P(A|B_i) \cdot P(B_i)}:$$

Օրինակ 3: Դիտարկենք նախորդ օրինակը և հաշվենք $P(B_1|A)$ հավանականությունը, կստանանք

$$P(B_1|A) = \frac{P(A|B_1)P(B_1)}{P(A)} = \frac{1}{7} : 0.238 = 0.6:$$

Այժմ բերենք պայմանական հավանականության հաշվման կիրառական բնույթի հետևյալ հայտնի խնդիրը:

Օրինակ 4: Ենթադրենք կա երեք փակ դուռ: Երկուսի հետևում կա ուլ, իսկ մեկում՝ մեքենա: Իհարկե, դուք գիտեք՝ որտեղ է մեքենան: Դուք առաջարկում եք ընտրել մի դուռ: Դրանից հետո դուք բացում եք մյուս դուռը, որի հետևում ուլն է, և դուք հարցնում եք՝ ցանկանում եք փոխել ձեր որոշումը: Այսինքն՝ կմեծանա ձեր շանսը մեքենա շահելու, եթե որոշումը փոխեք՝ ընտրելով մյուս դուռը: Ի՞նչ կանեիք դուք (տես նկ.8.1):



Նկ. 8.1: Բայեսի բանաձևի կիրառություն

Ներմուծենք երկու պատահական մեծություններ՝

- A (մեքենան գտնվում է առաջին, երկրորդ կամ երրորդ դռան հետևում),
- B (հաղորդավարը բացում է առաջին, երկրորդ կամ երրորդ դուռը):

Հաշվենք

$$P(A = 1|B = 2)$$

հավանականությունը: Այսինքն՝ փոխում ենք մեր որոշումը. ընտրում ենք առաջին դուռը, երբ հաղորդավարը բացել է երկրորդ դուռը: Հաշվելով այն Բայեսի բանաձևով՝ կստանանք՝

$$P(A = 1|B = 2) = \frac{P(A = 1)P(B = 2|A = 1)}{P(B = 2)}:$$

Ունենք $P(A = 1) = 1/3$, իսկ $P(B = 2|A = 1) = 1$, քանի որ սա հավանականությունն է այն բանի, որ հաղորդավարը կբացի երկրորդ դուռը՝ իմանալով, որ մեքենան առաջին դռան հետևում է: Այն հավասար է մեկի, քանի որ Դուք նշել եք երրորդ դուռը: Իսկ $P(B = 2)$ հավանականության հաշվարկը կատարենք լրիվ հավանականությունների բանաձևով: Կստանանք՝

$$\begin{aligned} P(B = 2) &= \sum_{i=1}^3 P(B = 2|A = i) = P(B = 2|A = 1)P(A = 1) + P(B = 2|A = 2)P(A = 2) + \\ &+ P(B = 2|A = 3)P(A = 3) = \frac{1}{3} + \frac{1}{6} = \frac{2}{3}: \end{aligned}$$

Այստեղ $P(B = 2|A = 2) = 0$, որովհետև հավանականությունն այն բանի, որ հաղորդավարը կբացի երկրորդ դուռը՝ իմանալով, որ այնտեղ մեքենան է, հավասար է զրոյի: Իսկ $P(B = 2|A = 3) = 1/2$, քանի որ սա հավանականությունն է այն բանի, որ հաղորդավարը կբացի երկրորդ դուռը, իմանալով, որ մեքենան երրորդ դռան հետևում է հավասար է $1/2$, որովհետև հաղորդավարը ունի երկու հնարավորություն՝ բացելու կա՛մ առաջին դուռը, կա՛մ երկրորդը:

Այսպիսով, փոխելով որոշումը՝ մեքենա շահելու շանսը մեծանում է և հավասարվում է $2/3$ -ի: Չփոխելով որոշումը՝ մեքենա շահելու հավանականությունը հավասարվում է $1/2$ -ի:

8.2 Օպտիմալ բայեսյան կանոններ

Դիտարկենք բինար կլասիֆիկացիայի պարզագույն խնդիրը: Դիտարկենք պատահական մեծությունների զույգ (X, Y) , որոնք ընդունում են արժեքներ համապատասխանաբար A և $B = \{1, -1\}$ բազմություններից: Ենթադրենք՝ հայտնի է առաջի և երկրորդ դասի օբյեկտների հավանականային բաշխման ֆունկցիաները՝ $P(X = x|Y = 1)$, $P(X = x|Y = -1)$: Այդ դեպքում առաջին և երկրորդ դասերի հանդես գալու հավանականությունները հավասար կլինեն համապատասխանաբար՝

$$P(Y = 1) = \int_A P(X = x|Y = 1)dx, \quad P(Y = -1) = \int_A P(X = x|Y = -1)dx:$$

Այստեղից երևում է, որ բոլոր այս հավանականությունները կարելի է հաշվել, եթե հայտնի է $P(x, y)$ հավանականությունը: Օրինակ՝

$$P(X = x|Y = 1) = \frac{P(X = x, Y = 1)}{P(Y = 1)}:$$

Օգտվելով Բայեսի բանաձևից՝ կարելի է որոշել օբյեկտի այս կամ այն դասին պատկանելու հավանականությունները: Օրինակ՝

$$P(Y = 1|X = x) = \frac{P(Y = 1)P(X = x|Y = 1)}{P(X = x|Y = 1)P(Y = 1) + P(X = x|Y = -1)P(Y = -1)}:$$

Կամայական $g : A \rightarrow B$ կլասիֆիկատորի համար սահմանենք նրա հավանականային սխալանքը հետևյալ բանաձևով՝

$$\text{error}_P(g) = P(g(X) \neq Y):$$

Եռանակենք

$$\eta_1(x) = P(Y = 1|X = x):$$

Սա հավանականությունն է այն բանի, որ x օբյեկտը պատկանում է առաջին դասին: Բայեսյան դասակարգիչը որոշվում է հետևյալ բանաձևով՝

$$h_1(x) = \begin{cases} 1, & \text{եթե } \eta_1(x) > 1/2, \\ -1, & \text{հակառակ դեպքում:} \end{cases}$$

Հետևյալ պնդումը ցույց է տալիս, որ բայեսյան դասակարգիչը այդ դասի համար

մինիմիզացնում է բայեսյան սխալանքի $\text{error}_P(g)$ ֆունկցիան, այսինքն՝

$$\min_g \text{error}_P(g) = \text{error}(h_1):$$

Թեորեմ 8.1. *Տեղի ունի հետևյալ անհավասարությունը՝*

$$P\{h_1(X) \neq Y\} = \int_A P(h_1(X) \neq Y|X=x)dx \leq$$

$$P\{g(X) \neq Y\} = \int_A P(g(X) \neq Y|X=x)dx \quad \forall g: A \rightarrow B:$$

Ապացույց: Կամայական x օբյեկտի համար ունենք

$$\begin{aligned} P\{g(X) \neq Y|X=x\} &= \\ &= 1 - P\{g(X) = Y|X=x\} = 1 - (P\{Y=1, g(X)=1|X=x\} + \\ &\quad P\{Y=-1, g(X)=-1|X=x\}) = \\ &= 1 - (\mathbb{1}_{\{g(x)=1\}}P\{Y=1|X=x\} + \mathbb{1}_{\{g(x)=-1\}}P\{Y=-1|X=x\}) = \\ &= 1 - (\mathbb{1}_{\{g(x)=1\}}\eta(x) + \mathbb{1}_{\{g(x)=-1\}}(1-\eta(x))), \end{aligned}$$

որտեղ

$$\mathbb{1}_{R(x)} = 1,$$

եթե $R(x)$ պայմանը ճիշտ է, հակառակ դեպքում $\mathbb{1}_{R(x)} = 0$: Նկատենք նաև, որ

$$\mathbb{1}_{g(x)=-1} = 1 - \mathbb{1}_{g(x)=1}:$$

Այսպիսով, ցանկացած x օբյեկտի համար կունենանք

$$\begin{aligned} P\{g(X) \neq Y|X=x\} - P\{h_1(X) \neq Y|X=x\} &= \\ &= \eta_1(x)(\mathbb{1}_{h_1(x)=1} - \mathbb{1}_{g(x)=1} + (1-\eta_1(x))(\mathbb{1}_{h_1(x)=-1} - \mathbb{1}_{g(x)=-1})) = \\ &= (2\eta_1(x) - 1)(\mathbb{1}_{h_1(x)=1} - \mathbb{1}_{g(x)=1}) \geq 0: \end{aligned}$$

Այնուհետև ինտեգրելով այս անհավասարությունը՝ կստանանք թեորեմի պնդումը:

Անալոգիայով սահմանվում է Բայեսի պրեդիկտորը երկրորդ դասի համար՝

$$h_{-1}(x) = \begin{cases} -1, & \text{եթե } \eta_{-1}(x) > 1/2, \\ 1, & \text{հակառակ դեպքում,} \end{cases}$$

որտեղ

$$\eta_{-1}(x) = P(Y = 1|X = x):$$

Հեշտ է նկատել, որ

$$\max\{h_1(x), h_{-1}(x)\} = h_{baies}(x) = \arg \max_{y \in B} P(Y = y|X = x)$$

Ֆունկցիան բավարարում է թեորեմ 8.1-ի անհավասարությանը:

Դիցուք այժմ $B = \{1, 2, \dots, K\}$ և $B \times B$ բազմության վրա տրված է $L(y, s) = [y \neq s]$ ֆունկցիան: Միջին ռիսկի ֆունկցիա $h(x)$ պրեդիկտորի (ալգորիթմի) դեպքում կոչվում է հետևյալ ֆունկցիոնալը՝

$$R(h) = \sum_{i=1}^K \int_A L(y, h(x)) P(Y = y_k|X = x) dx:$$

Եթե հայտնի է հավանականությունների բաշխման $P(Y = y|X = x)$ ֆունկցիան, ապա կարելի է գտնել այն օպտիմալ h^* պրեդիկտորը, որը մինիմիզացնում է այդ միջին ռիսկը: Ապացուցենք, որ միջին ռիսկը մինիմիզացվում է

$$h^*(x) = \arg \max_{y \in B} P(Y = y|X = x)$$

հիպոթեզի վրա: Իրոք, ունենք

$$\begin{aligned} R(h) &= \sum_{y=1}^K \int_A L(y, h(x)) P(Y = y|X = x) dx = \\ &= \sum_{y=1}^K \int_A [y \neq h(x)] P(Y = y|X = x) dx = \\ &= \int_A \sum_{y \neq h(x)} P(Y = y|X = x) dx = \\ &= 1 - \int_A P(x, h(x)) dx \geq 1 - \int_A \max_{s \in Y} P(s|X = x) dx = \\ &= 1 - \int_A P(Y = h^*(x)|X = x) dx = R(h^*): \end{aligned}$$

Այսպիսով ապացուցվեց, որ օպտիմալ բայեսյան դասակարգիչը ընտրում է այն դասը, որը ունի մաքսիմալ ապոստորիոր հավանականություն: Օպտիմալ այս ալգորիթմը ունի նաև հետևյալ համարժեք տեսքը՝

$$h^*(x) = \arg \max_{y \in B} P(Y = y)P(X = x|Y = y):$$

Բայեսյան օպտիմալ ալգորիթմի կառուցման հիմնական պրոբլեմն այն է, որ անհայտ է $P(X = x|Y = y)$ հավանականային բաշխման ֆունկցիան: Սովորաբար այդ հավանականությունը վերականգնում են տրված հավաքածուի միջոցով: Դիցուք ունենք S նմուշների հավաքածու: Նշանակենք

$$S_y = \{(x, y_x) \in S / y_x = y\}:$$

Այդ դեպքում որպես հավանականությունների մոտավոր արժեքներ կարելի է վերցնել հետևյալ արտահայտությունները՝

$$P(Y = y) \approx \frac{|S_y|}{|S|}, \quad P(x|y) \approx \frac{1}{m} \sum_{i=1}^m [x_i = x],$$

որտեղ $x_1, \dots, x_m$ էլեմենտները S_y բազմության մեջ մտնող գույգերի առաջին կոմպոնենտներն են:

Օրինակ՝ ենթադրենք, որ կա գծային կախվածություն օբյեկտների և դրանց պատասխանների միջև: Այսինքն՝

$$y_i = (w, x_i) + \epsilon_i, \quad i = 1, 2, \dots, l,$$

որտեղ ϵ_i նորմալ բաշված աղմուկ է: Այսինքն՝

$$L(w) = \prod_{i=1}^l P(y_i|x_i, w) = \prod_{i=1}^l \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - (w, x_i))^2}{2\sigma^2}\right):$$

Լոգարիթմելով այս արտահայտությունը՝ կստանանք w պարամետրի որոշման հետևյալ օպտիմիզացիոն խնդիրը՝

$$\ln(L(w)) = -l \ln(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2} \sum_{i=1}^l (y_i - (w, x_i))^2 \rightarrow \max:$$

Անտեսելով այն անդամները, որոնք w -ից կախված չեն, կստանանք փոքրագույն քառակուսիների մեթոդի հետևյալ խնդիրը՝

$$\sum_{i=1}^l (y_i - (w, x_i))^2 \rightarrow \min_w:$$

8.3 Բայեսի պարզ դասակարգիչ

Ինչպես արդեն ցույց տրվեց, Բայեսի օպտիմալ պրեդիկտորն է

$$h_{bajes}(x) = \arg \max_{y \in B} P(Y = y|X = x)$$

արտապատկերումը: Ենթադրենք՝ $x = (x_1, x_2, \dots, x_n)$, իսկ $X = (X_1, X_2, \dots, X_n)$ պատահական վեկտորը բավարարում է հետևյալ պայմանը՝

$$P(X = x|Y = y) = \prod_{i=1}^n P(X_i = x_i|Y = y):$$

Սա նշանակում է, որ միևնույն նշիչ (մետկա) ունեցող օբյեկտները իրարից անկախ են: Հաշվի առնելով սա և Բայեսի բանաձևը օպտիմալ պրեդիկտորի համար՝ կստանանք՝

$$h_{bajes}(x) = \arg \max_{y \in B} P(Y = y|X = x) =$$

$$\arg \max_{y \in B} P(Y = y)P(X = x|Y = y) =$$

$$\arg \max_{y \in B} P(Y = y) \prod_{i=1}^n P(X_i = x_i|Y = y):$$

Այժմ երեք դասերի համար դիտարկենք Բայեսի դասակարգիչը՝

$$P(classA|Feature1, Feature2, Feature3) = \frac{P(Feature1|A)P(Feature2|A)P(Feature3|A)P(class(A))}{P(Feature1)P(Feature2)P(Feature3)}:$$

Աղյուսակ 8.1, տվյալների բազա				
Դասեր(Class)	Երկար (Long)	Քաղցր (Sweet)	Դեղին (Yellow)	Ընդհ.(Total)
բանան(banana)	400	350	450	500
նարինջ (orange)	0	150	300	300
ուրիշ (other)	100	150	50	200
ընդհանուր (total)	500	650	800	1000

Ունենք տվյալների հավաքածու, որը տրված է աղյուսակ 8.1-ում:

- Կա 1000 հատ միրգ:
- Բանան, նարինջ և ուրիշ, այսինքն՝ երեք դասի:
- Սրգերը երկար են, քաղցր և դեղին: Սրանք հատկանիշներ են (Feature 1, Feature 2, Feature 3):

Ինչպես երևում է աղյուսակից

- 500 բանաններից 400-ը երկար է, 350-ը քաղցր է, 450-ը՝ դեղին:
- 300 նարինջից 0 հատը երկար է, 150-ը՝ քաղցր, 300-ը՝ դեղին:
- 200 մնացած սրգերից 100-ը երկար է, 150-ը քաղցր է, 50-ը՝ դեղին:

Դիցուք անհայտ միրգը (երկար, քաղցր, դեղին) հատկանիշներ ունի: Պետք է որոշել, թե որ դասին է նա պատկանում՝ բանան, նարինջ և ուրիշ միրգ: Հաշվենք համապատասխան հավանականությունները.

$$P(Banana|long, Sweet, Yellow), P(Orange|Long, Sweet, Yellow),$$

$$P(Other|Long, Sweet, Wellow):$$

Սկզբից գտնենք հետևյալ հավանականությունները՝

- $P(Long|Banana) = 400/500 = 0.8$,
- $P(Sweet|Banana) = 350/500 = 0.7$,
- $P(Yellow|Banana) = 450/500 = 0.9$,
- $P(Banana) = 500/1000 = 0.5$:

Բազմապատկելով այս անհավասարությունները՝ կստանանք՝

$$0.8 \times 0.7 \times 0.9 \times 0.2 = 0.252:$$

Այս թիվը արտահայտում է Բայեսի դասակարգչի համարիչը: Նույն ձևով հաշվարկ կատարելով մնացած դասերի համար՝ կստանանք

- $P(Orange|Long, Sweet, Yellow) = 0$,
- $P(Other|Long, Sweet, Yellow) = 0.01875$:

Բայեսի կլասիֆիկատորի հայտարարը բոլոր երեք դեպքերում էլ նույնն է, հետևաբար համեմատելով ստացված արդյունքները, ստանում ենք, որ անհայտ միրգը պատկանում է *Banana* դասին, քանի որ նրան պատկանելու հավանականությունը ավելի մեծ է:

8.4 Առավելագույն հավանականության մեթոդը (Maximum likelihood function)

Դիցուք դեղագործական մի ընկերություն մշակել է դեղամիջոց մահացու ինչ-որ մի հիվանդության բուժման և ցանկանում է գնահատել արդյունավետությունը, երբ բուժառուն (պացիենտը) ընդունի այն: Այդ պատճառով ընկերությունն ընտրում է l կամավոր: Թող $S = (x_1, x_2, \dots, x_l)$ բազմությունը լինի ուսուցանող հավաքածուն, որում $x_i = 1$, եթե i -րդ բուժառուն ողջ է դեղամիջոցն ընդունելուց հետո, և $x_i = 0$ ՝ հակառակ դեպքում: θ -ով նշանակենք ողջ մնալու հավանականությունը: Այժմ ցանկանում ենք գնահատել այդ մեծությունը S հավաքածուի միջոցով: Բնական է որպես այդ պարամետրի գնահատական պետք է վերցնել մեկերի միջին քանակությունը այդ հավաքածուում՝

$$\hat{\theta} = \frac{1}{l} \sum_{i=1}^l x_i:$$

Ակնհայտ է, որ $E_S(\hat{\theta}) = \theta$: Վիճակագրությունում ասում են, որ $\hat{\theta}$ -ն θ պարամետրի **անշեղելի** գնահատականն է: Բացի դրանից, քանի որ $\hat{\theta}$ -ն իրարից անկախ, միևնույն կանոնով բաշխված պատահական մեծությունների միջինն է, ապա համաձայն Հոֆդինգի թեորեմի, առնվազն $1 - \delta$ հավանականությամբ S

հավաքածուի համար տեղի ունի

$$|\hat{\theta} - \theta| \leq \sqrt{\frac{\ln(2/\delta)}{2l}}$$

անհավասարությունը:

Այժմ տանք $\hat{\theta}$ գնահատականի ևս մի մեկնաբանություն, որը կոչվում է մաքսիմալ ճշմարտանմանության եղանակ կամ առավելագույն հավանականությունների մեթոդ: Հաշվենք S հավաքածուի գեներացման հավանականությունը՝

$$P(S = (x_1, x_2, \dots, x_l)) = \prod_{i=1}^l \theta^{x_i} (1 - \theta)^{1-x_i} = \theta^{\sum_i x_i} (1 - \theta)^{\sum_i (1-x_i)} :$$

Լոգարիթմելով այս արտահայտությունը՝ կստանանք՝

$$L(S, \theta) \equiv \ln(P(S = (x_1, x_2, \dots, x_l))) = \ln(\theta) \sum_i x_i + \ln(1 - \theta) \sum_i (1 - x_i) :$$

θ պարամետրի մաքսիմալ հավանականային գնահատական է կոչվում այդ պարամետրի այն արժեքը, որը մաքսիմում է հաղորդում $L(S, \theta)$ ֆունկցիային: Այսինքն՝

$$\hat{\theta} = \arg \max_{\theta} L(S, \theta) :$$

Ածանցելով L ֆունկցիան ըստ θ պարամետրի և հավասարեցնելով այն զրոյի՝ կստանանք՝

$$\frac{\sum_i x_i}{\theta} - \frac{\sum_i (1 - x_i)}{(1 - \theta)} = 0 :$$

Այստեղից՝

$$\theta = \frac{1}{l} \sum_i x_i,$$

որը համընկնում է վերևի գնահատականի հետ:

Այժմ քննարկենք անընդհատ պատահական մեծությունների դեպքը: Դիցուք այժմ հավաքածուի յուրաքանչյուր էլեմենտ անկախ ընտրվել է մի P_{θ} հավանականային խտությամբ բաշխված կետերի բազմությունից, որտեղ θ -ն պարամետր է: Հավաքածուի «ճշմարտանմանության ֆունկցիա» (likelihood function) սահմանվում է հետևյալ բանաձևով՝

$$L(S, \theta) = \ln\left(\prod_{i=1}^l P_{\theta}(x_i)\right) = \sum_{i=1}^l \ln(P_{\theta}(x_i)) :$$

$\hat{\theta} = \arg \max_{\theta} L_S(S, \theta)$ պայմանը բավարարող պարամետրը կոչվում է մաքսիմալ ճշմարտանմանության գնահատական:

Այժմ ցույց տանք, որ մաքսիմալ ճշմարտանմանության գնահատականը կապված է Էմպիրիկ ռիսկի ֆունկցիայի մինիմիզացիայի հետ: Սահմանենք

օբյեկտի վրա կորստի ֆունկցիա հետևյալ կերպ՝

$$l(\theta, x) = -\ln(P_\theta(x)):$$

Այս սահմանումից հետևում է, որ առավելագույն հավանականության սկզբունքը համարժեք է էմպիրիկ ռիսկի հետևյալ ֆունկցիայի մինիմիզացիային՝

$$L_S(S, \theta) = \sum_{i=1}^l -\ln(P_\theta(x_i)) \rightarrow \min_\theta:$$

Այժմ ցույց տանք, որ որոշ դեպքերում մաքսիմալ ճշմարտանմանության գնահատականը ապահովում է ցածր ռիսկ: Որպես օրինակ դիտարկենք միավոր դիսպերսիայով նորմալ բաշխման մաթեմատիկական սպասման ճշմարտանմանության գնահատականը: Այս դեպքում հայտնի է, որ այդպիսի պատահական մեծության հավանականային խտության ֆունկցիան ունի հետևյալ տեսքը՝

$$P_\theta = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\theta)^2}{2}\right):$$

Այս դեպքում ճշմարտանմանության ֆունկցիան ունի հետևյալ տեսքը՝

$$\begin{aligned} L(S, \theta) &= \sum_{i=1}^l \ln P_\theta(x_i) = \\ &= -l \ln \sqrt{2\pi} - \frac{1}{2} \sum_{i=1}^l (x_i - \theta)^2: \end{aligned}$$

Ածանցելով այս ֆունկցիան ըստ θ -ի և հավասարեցնելով այն զրոյի՝ կստանանք θ պարամետրի (մաթեմատիկական սպասման) ճշմարտանման արժեքը՝

$$\hat{\theta} = \frac{1}{l} \sum_{i=1}^l x_i:$$

Այժմ գնահատենք հետևյալ մեծությունը՝

$$\begin{aligned} E_{x \sim N(\theta, 1)}[l(\hat{\theta}, x) - l(\theta, x)] &= \\ &= E_{x \sim N(\theta, 1)}\left[\ln\left(\frac{P_{\theta(x)}}{P_{\hat{\theta}}(x)}\right)\right] = E_{x \sim N(\theta, 1)}\left[-\frac{1}{2}(x - \theta)^2 + \frac{1}{2}(x - \hat{\theta})^2\right] = \\ &= \frac{\hat{\theta}^2}{2} - \frac{\theta^2}{2} + (\theta - \hat{\theta})E_{x \sim N(\theta, 1)}(x) = \\ &= \frac{\hat{\theta}^2}{2} - \frac{\theta^2}{2} + (\theta - \hat{\theta})\theta = \frac{1}{2}(\hat{\theta} - \theta)^2: \end{aligned}$$

Նկատենք, որ $\hat{\theta}$ -ն է միևնույն θ մաթեմատիկական սպասումը ունեցող նորմալ բաշխված l հատ պատահական մեծությունների միջին արժեքն է: Միևնույն ժամանակ այդ միջին պատահական մեծության դիսպերսիան հավասար է $\sigma =$

1: Հետևաբար, կիրառելով Հոֆդինգի թեորեմը՝ կստանանք, որ առնվազն $1 - \delta$ հավանականությամբ տեղի ունի

$$|\hat{\theta} - \theta| \leq \epsilon$$

անհավասարությունը, որտեղ ϵ -ը կախված է δ և σ պարամետրերից: Հետևաբար՝

$$P\left(|E_{x \sim N(\theta, 1)}(l(\hat{\theta}, x)) - E_{x \sim N(\theta, 1)}(l(\theta, x))| \leq \epsilon\right) \geq 1 - \delta:$$

Այս դեպքում գրականության մեջ ասում են, որ չկա գերուսուցում (**overfitting**), Էֆեկտիվ չէ այգորիթմը:

Դիտարկենք այժմ X պատահական մեծությունը, որ ունի Բեռնուլի բաշխում, և դիցուք $P(X = 1) = \theta$: Ենթադրենք θ -ն բավականաչափ փոքր է: Դիտարկենք հավաքածու, որի բոլոր էլեմենտները հավասար են գրոյի: Այդ դեպքում մաքսիմալ հավանականային գնահատականը կլինի՝

$$\hat{\theta} = \frac{1}{l} \sum_{i=1}^l x_i = 0:$$

Հաշվենք դրա ռիսկը:

$$\begin{aligned} E_{x \sim \theta}(l(\hat{\theta}, x)) &= \theta l(\hat{\theta}, 1) + (1 - \theta)l(\hat{\theta}, 0) = \\ &= \theta \ln(1/\hat{\theta}) + (1 - \theta) \ln(1/(1 - \hat{\theta})) = \\ &= \theta \ln(1/0) = \infty: \end{aligned}$$

Այս պարզ օրինակը ցույց է տալիս, որ առավելագույն հավանականության մոդելով ծնված պրեդիկտորը **Էֆեկտիվ** չէ որոշ դեպքերում:

Դիտարկենք բինոմական բաշխման դեպքը: Դիցուք ունենք $X_1, X_2, \dots, X_n$ պատահական անկախ մեծություններ, որոնք θ հավանականությամբ ընդունում են 1 արժեք, իսկ $1 - \theta$ հավանականությամբ՝ 0 արժեք: Այդ դեպքում ակնհայտ է, որ

$$P(X_i = x_i) = \theta^{x_i} (1 - \theta)^{1-x_i}:$$

Դիցուք $X = \sum_{i=1}^n X_i$: Քանի որ X_i , $i = 1, 2, \dots, n$ պատահարները անկախ են, ապա դժվար չէ համոզվել, որ

$$P(X = m | n, p) = C_n^m \theta^m (1 - \theta)^{n-m}:$$

Այս հավանականությունը կոչվում է բինոմական բաշխման ֆունկցիա:

Օրինակ: Առավելագույն հավանականության մեթոդով պետք է գնահատել θ պարամետրը բինոմական բաշխման դեպքում, եթե n_1 անկախ փորձերի ընթացքում պատահարը հանդես է եկել m_1 անգամ, իսկ n_2 փորձերի ընթացքում՝ m_2 անգամ:

Լուծում: Բինոմական բաշխման ֆունկցիան ունի հետևյալ տեսքը՝

$$P_n(k) = C_n^k \theta^k (1 - \theta)^{n-k}:$$

Կազմենք, այսպես ասած, ճշմարտանմանության ֆունկցիան (երբեմն այն անվանում են « վստահելիության ֆունկցիա »).

$$L \equiv P_{n_1}(m_1)P_{n_2}(m_2) = C_{n_1}^{m_1} C_{n_2}^{m_2} \theta^{m_1+m_2} (1 - \theta)^{n_1+n_2-(m_1+m_2)}:$$

Լոգարիթմելով այս արտահայտությունը և այն հավասարեցնելով զրոյի՝

$$\frac{d \ln L}{d\theta} = \frac{m_1 + m_2}{\theta} - \frac{(n_1 + n_2) - (m_1 + m_2)}{1 - \theta} = 0:$$

Այստեղից,

$$\theta = \frac{m_1 + m_2}{n_1 + n_2}:$$

Այս կետում երկրորդ կարգի ածանցյալը բացասական է, և հետևաբար այն մաքսիմումի կետ է:

Թեորեմ 8.2. Դիցուք ունենք

$$P(X = x|n, \theta) = C_n^x \theta^x (1 - \theta)^{n-x}, \quad x = 0, 1, 2, \dots, n$$

բինոմական բաշխումը: Այդ դեպքում

$$P(X = x|n, \theta) \rightarrow \frac{\lambda^x e^{-\lambda}}{x!}, \quad \text{երբ } n \rightarrow \infty, \theta \rightarrow 0, n\theta = \lambda = \text{const}:$$

Ապացույց: Ունենք

$$\begin{aligned} P\{X = x|n, \theta\} &= \frac{n(n-1)\dots(n-x+1)}{x!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} = \\ &= \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)\dots\left(1 - \frac{x-1}{n}\right)\left(1 - \frac{\lambda}{n}\right)^{-x} \left(1 - \frac{\lambda}{n}\right)^n \frac{\lambda^x}{x!}: \end{aligned}$$

Այստեղ, անցնելով սահմանի, հաշվի առնելով, որ

$$\left(1 - \frac{\lambda}{n}\right)^n \rightarrow e^{-\lambda},$$

կստանանք՝

$$P(X = x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \dots:$$

Սա կոչվում է Պուանսոնի բաշխում:

Օրինակ: Տրված հավաքածուի վրա գտնել Պուանսոնի բաշխման ֆունկցիայի պարամետրի առավելագույն գնահատականը:

Լուծում: Ունենք

$$\begin{aligned} L(\lambda) &= P(X = x_1|\lambda)P(X = x_2|\lambda)\dots P(X_n = x_n|\lambda) = \\ &= \frac{e^{-\lambda}\lambda^{x_1}}{x_1!}\dots \frac{e^{-\lambda}\lambda^{x_n}}{x_n!} = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}}{x_1!\dots x_n!} \rightarrow \max: \end{aligned}$$

Այս խնդիրը համարժեք է հետևյալին՝

$$l(\lambda) = -n\lambda + \ln \lambda \sum_{i=1}^n x_i \rightarrow \max:$$

Այստեղից, ածանցելով այս ֆունկցիան և հավասարեցնելով այն զրոյի՝ կստանանք

$$\lambda = \frac{\sum_{i=1}^n x_i}{n} = \bar{X}:$$

Քանի որ

$$l''(\lambda) = -\frac{\sum_{i=1}^n x_i}{\lambda^2} < 0,$$

ապա $\lambda = \bar{X}$ կետը մաքսիմումի կետն է:

8.5 EM (expectation maximization) ալգորիթմը

Այս ալգորիթմը հետապնդում է պայմանական $p(x|y)$ հավանականության համար լավագույն գնահատական ստանալու նպատակ, երբ հայտնի է, որ այն հավանականային խտությունների «խառնուրդ» է: Այսինքն՝ այն ունի հետևյալ տեսքը՝

$$p(x|y) = \sum_{j=1}^k w_j \phi(x, \theta_j),$$

որտեղ ϕ -ին հայտնի ֆունկցիա է, իսկ $w_j, \theta_j; j = 1, 2, \dots, k$ պարամետրերը անհայտ են, ընդ որում

$$\sum_{j=1}^k w_j = 1, \quad w_j \geq 0, \quad j = 1, 2, \dots, k:$$

Նշանակենք

$$\Theta = (w_1, \dots, w_k, \theta_1, \dots, \theta_k):$$

Ունենք տվյալների $X = \{x_1, x_2, \dots, x_m\}$ հավաքածու և այդ հավաքածուի համար կազմենք *likelihood* ֆունկցիան և այն մաքսիմիզացնենք: Այսինքն՝ ունենք օպտիմիզացիայի հետևյալ խնդիրը՝

$$\prod_{i=1}^m \sum_{j=1}^k w_j \phi(x_i, \theta_j) \rightarrow \max_{\Theta},$$

$$\sum_{j=1}^k w_j = 1, \quad w_j \geq 0, \quad j = 1, 2, \dots, k:$$

Այս խնդիրը նույնն է, ինչ մաքսիմիզացնել

$$\sum_{i=1}^m (\ln \sum_{j=1}^k w_j \phi(x_i, \theta_j))$$

ֆունկցիան նշված սահմանափակումներով: Այժմ այն լուծենք Լագրանժի անորոշ գործակիցների մեթոդով: Լագրանժյանը ունի հետևյալ տեսքը՝

$$L = \sum_{i=1}^m (\ln \sum_{j=1}^k w_j \phi(x_i, \theta_j)) - \lambda (\sum_{j=1}^k w_j - 1):$$

Հետևաբար՝ ըստ Էքստրեմումի անհրաժեշտ պայմանի՝ ունենք

$$\frac{\partial L}{\partial w_j} = \sum_{i=1}^m \frac{\phi(x_i, \theta_j)}{p(x_i)} - \lambda = 0,$$

որտեղ

$$p(x_i) = \sum_{j=1}^k w_j \phi(x_i, \theta_j):$$

Այստեղից հետևում է, որ

$$\sum_{i=1}^m \sum_{j=1}^k w_j \frac{\phi(x_i, \theta_j)}{p(x_i)} = \lambda \sum_{j=1}^k w_j = \lambda:$$

Մյուս կողմից այս հավասարության ձախ մասը հավասար է m -ի: Ուրեմն՝

$$\lambda = m, \quad w_j = \frac{1}{m} \sum_{i=1}^m \frac{w_j \phi(x_i, \theta_j)}{p(x_i)}: \quad (8.1)$$

Ինչպես տեսնում ենք, $w_j, \quad j = 1, 2, \dots, k$ պարամետրերը ոչ բացասական են: Ըստ $\theta_j, \quad j = 1, 2, \dots, k$ պարամետրերի օպտիմալության՝ անհրաժեշտ պայմանները հետևյալներն են՝

$$\frac{\partial L}{\partial \theta_j} = \frac{\partial}{\partial \theta_j} \sum_{i=1}^m \frac{w_j \phi(x_i, \theta_j)}{p(x_i)} \ln \phi(x_i, \theta_j) = 0,$$

ինչը նշանակում է, որ

$$\theta_j = \arg \max_{\theta} \sum_{i=1}^m \frac{w_j \phi(x_i, \theta)}{p(x_i)} \ln \phi(x_i, \theta), \quad j = 1, 2, \dots, k: \quad (8.2)$$

Նորմալ բաշխման դեպքում, երբ

$$\phi(x, \theta_j) = N(x, \mu_j, \Sigma_j),$$

ապա $\theta_j = (\mu_j, \Sigma_j)$ արժեքները արտահայտվում են հետևյալ բանաձևերով՝

$$\mu_j = \frac{1}{m_j} \sum_{i=1}^m g_{ij} x_i, \quad \Sigma_j = \frac{1}{m w_j} \sum_{i=1}^m g_{ij} (x_i - \mu_j)(x_i - \mu_j)^T: \quad (8.3)$$

Այժմ նկարագրենք EM ալգորիթմը, որը գեներացնում է $((w_j^{(n)}, \theta_j^{(n)}, j = 1, 2, \dots, k)$ հաջորդականությունը, և որը գուգամիտում է խնդրի օպտիմալ լուծմանը:

- $(w_j, \theta_j), j = 1, 2, \dots, k$ պարամետրերին վերագրվում են կամայական սկզբնական արժեքներ՝

$$w_j = w_j^{(0)} \geq 0, \quad \theta_j = \theta_j^{(0)}, \quad j = 1, 2, \dots, k, \quad \sum_{j=1}^k w_j^{(0)} = 1:$$

- Կառուցում ենք $G = \{g_{ij}\}$ մատրիցը:
- (8.1) բանաձևով հաշվարկվում է $w_j, j = 1, 2, \dots, k$ պարամետրերի նոր արժեքներ:
- (8.2) բանաձևով հաշվարկվում է $\theta_j, j = 1, 2, \dots, k$ պարամետրերի նոր արժեքներ:
- Եթե $\max_{i,j} |g_{i,j} - g'_{i,j}| < \delta$, ապա ավարտել աշխատանքը, հակառակ դեպքում՝ վերադառնալ երկրորդ կետին: Այստեղ $g_{i,j}, g'_{i,j}$ թվերը հաջորդ և նախորդ իտերացիաներում կառուցված G և G' մատրիցների էլեմենտներն են, իսկ δ -ն նախօրոք տրված ճշտություն է:

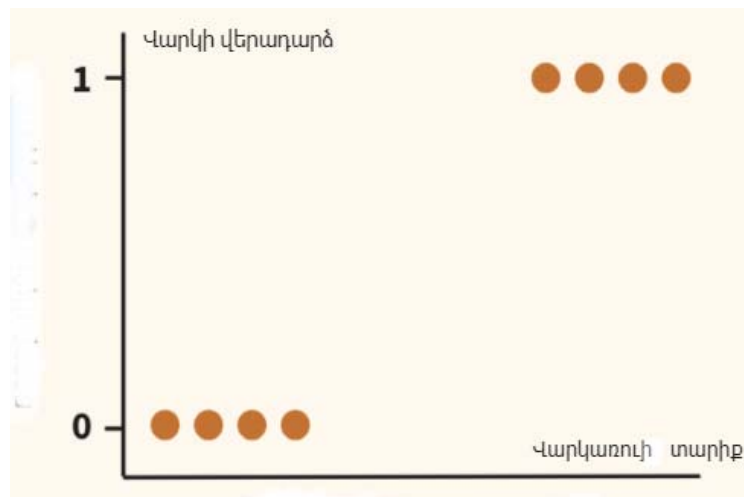
Այս ալգորիթմի գուգամիտության ապացույցի հետ կարելի է ծանոթանալ [21] աշխատանքում:

8.6 Լոգիստիկ ռեգրեսիայի մեթոդը

Լոգիստիկ ռեգրեսիայի մեթոդը բինար դասակարգման հավանականային մեթոդներից է: Քննարկենք դա օրինակների վրա: Նորից անդրադառնանք վարկավորման խնդրին: Ենթադրենք՝ մենք հավաքել ենք տվյալներ վարկառուների տարիքների և վարկը վերադարձնելու վերաբերյալ: Դա գրաֆիկորեն ներկայացվում է նկ. 8.2-ում: Ինչպես երևում է, ութ վարկառուներից չորսը, որոնք ավելի երիտասարդ են, չեն վերադարձրել վարկը, իսկ մյուս չորսը՝ ավելի բարձր տարիքով, վերադարձրել են այն: Հարց է առաջանում, թե ինչպիսի մոդելով կարելի է նկարագրել այս կախվածությունը:

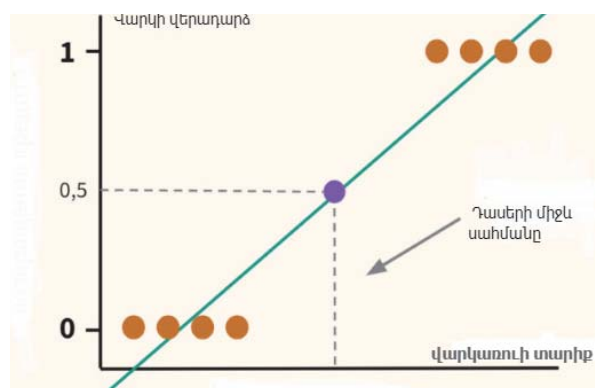
Թվում է, թե կարելի է կառուցել գծային ռեգրեսիայի ֆունկցիա՝ $(\mathbf{w}, \mathbf{x}) \equiv \sum_{i=1}^n x^i w_i - w_0$ այնպես, որ եթե մի օբյեկտի վրա նրա արժեքը փոքր լինի որոշակի ֆիքսված արժեքից, դասել մի խմբում, հակառակ դեպքում օբյեկտը դասել մյուս խմբում: Օրինակ հետևյալ կերպ՝

- $(\mathbf{w}, \mathbf{x}) < 0.5 \Rightarrow y = 0,$
- $(\mathbf{w}, \mathbf{x}) \geq 0.5 \Rightarrow y = 1:$



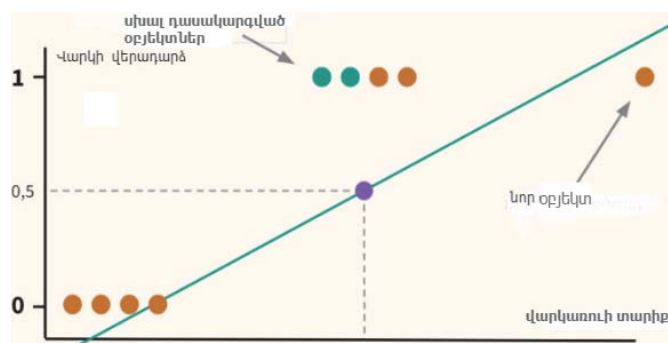
Նկ. 8.2: Էմպիրիկ տվյալներ

Այնուամենայնիվ, նույնիսկ եթե ենթադրենք, որ մենք հաջողությամբ գծեցինք ռեգրեսիոն գիծը (նկար 8.3-ում մենք իրականում բավականին դա հաջող գծեցինք), և մեր մոդելը կարող է բարձրորակ կանխատեսում անել: Նոր տվյալների ի հայտ գալը կփոխի այս սահմանը, և. արդյունքում ոչինչ չի ավելացնի, այլ միայն կվատթարացնի մոդելի ճշգրտությունը:



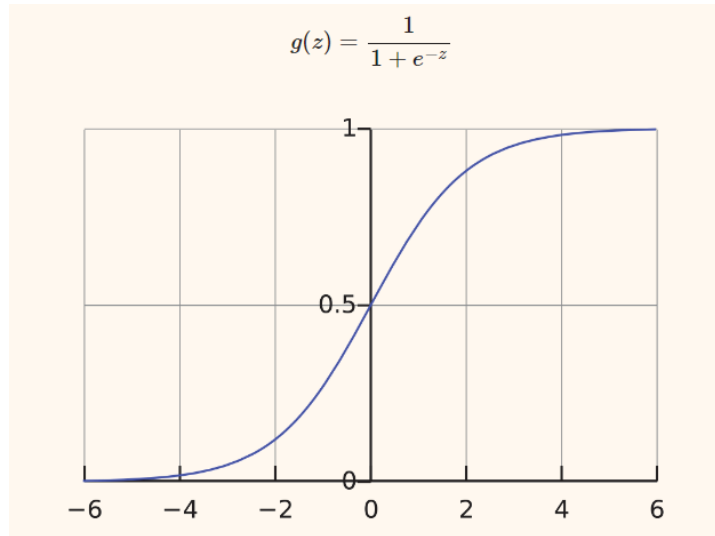
Նկ. 8.3: Տվյալների բաժանումը գծային ֆունկցիայով

Այստեղ մոդելը երկու օբյեկտի վրա կատարել է սխալ կանխատեսում: Բացի դրանից գծային ֆունկցիայի արժեքները դուրս են գալիս $[0, 1]$ միջակայքից (տե՛ս նկ. 8.4):



Նկ. 8.4: Նոր տվյալների վրա գծային մոդելը տալիս է սխալ կանխատեսում

Վերը նշված դժվարությունների հնարավոր լուծումը գծային ռեգրեսիայի փոխարինում են սիգմոյդ ֆունկցիայով, որի արժեքները արդեն դուրս չեն գալիս նշված տիրույթից: Հիշենք սիգմոյդի բանաձևն ու գրաֆիկը (նկ. 8.5):



Նկ. 8.5: Սիգմոյդի գրաֆիկը

Այժմ մենք կարող ենք կառուցել գծային $(\mathbf{w}, \mathbf{x}) \equiv \sum_{i=1}^n x^i w_i - w_0$, մոդելը և այն փոխանցել սիգմոյդի

$$z = (\mathbf{w}, \mathbf{x}) \rightarrow h_{\mathbf{w}}(x) = \frac{1}{1 + e^{-(\mathbf{w}, \mathbf{x})}}:$$

Այժմ մենք $h_{\mathbf{w}}(x)$ արժեքը կարող ենք համարել այս կամ այն դասին պատկանելու հավանականություն՝

$$h_{\mathbf{w}}(x) = P(y = 1 | x, \mathbf{w}):$$

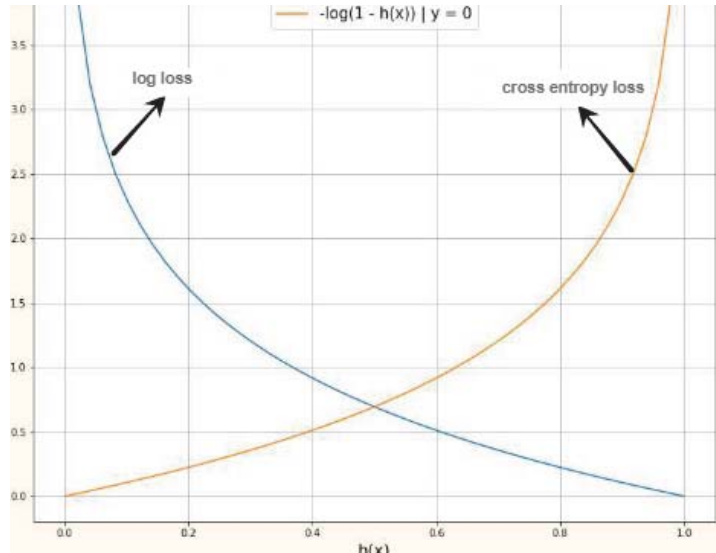
Սա x օբյեկտի «1» դասին պատկանելու հավանականությունն է w պարամետրով գծային մոդելի դեպքում: Այստեղ, եթե որպես կորստի ֆունկցիա վերցնենք միջին քառակուսային շեղման ֆունկցիան, ապա արդյունքում կստացվի ոչ ուռուցիկ ֆունկցիա, որի գլոբալ մինիմումի կետ գտնելը բավականին դժվար խնդիր է:

Միջին քառակուսային շեղման ֆունկցիայի փոխարեն մենք կօգտագործենք լոգիստիկ սխալի ֆունկցիան, որը կոչվում է նաև *log loss* և Էնտրոպիայի կորստի (binary cross-entropy loss) ֆունկցիա (նկ. 8.6) :

Այստեղ ներկայացված են *log loss* և *binary cross-entropy loss* ֆունկցիաների գրաֆիկները:

Ենթադրենք՝ տվյալների բազայում վարկառուն մարել է վարկը (նրա թիրախային փոփոխականը հավասար է մեկի): Այնուհետև գործարկվում է *log-loss* ֆունկցիան, և սխալը չափվում է դրա միջոցով: Ըստ այդմ, որքան մոտ է մոդելի տված հավանականությունը մեկին, այնքան փոքր է սխալը:

Եվ եթե վարկառուն չի մարել վարկը (նրա թիրախային փոփոխականը հավասար է զրոյի) գործարկվում է *entropy loss* ճյուղը: Մոդելի սխալը նվազագույն կլինի զրոյին մոտ արժեքների համար:



Նկ. 8.6: log loss and binary cross-entropy loss ֆունկցիաների գրաֆիկները

Այսպիսով հանրագումարելով այս երկուսը, եզրակացնում ենք, որ x օբյեկտը $y \in \{0, 1\}$ դասին պատկանելու log loss-ը հավասար կլինի՝

$$-y \ln h_{\mathbf{w}}(x) - (1 - y) \ln(1 - h_{\mathbf{w}}(x)):$$

Այժմ կարող ենք կառուցել մեր տվյալների բազայի միջին log loss-ի ֆունկցիան կախված w պարամետրից և մինիմիզացնել այն, այսինքն՝

$$L(w) = -\frac{1}{l} \sum_{i=1}^l (y_i \ln h_{\mathbf{w}}(x_i) + (1 - y_i) \ln(1 - h_{\mathbf{w}}(x_i))) \rightarrow \min_{\mathbf{w}}:$$

Դժվար չէ համոզվել, որ այս խնդիրը համարժեք է

$$L(\mathbf{w}) = \sum_{i=1}^l \ln(1 + e^{-z_i(\mathbf{w}, \mathbf{x}_i)}) \rightarrow \min_{\mathbf{w}}$$

խնդրին, որտեղ $z_i = 1$, եթե x_i օբյեկտի նշիչը հավասար է 1, իսկ $z_i = -1$, եթե x_i օբյեկտի նշիչը 0 է:

Այժմ ևս մի կոնկրետ օրինակով, առավելագույն հավանականությունների մեթոդով դուրս բերենք լոգիստիկ ռեգրեսիայում կորստի $L(w)$ ֆունկցիայի բանաձևը:

Դիցուք բանկում կա բլանկ, որում նշված են հաճախորդի (կլիենտի) տվյալները վարկ ստանալու վերաբերյալ: Հաճախորդի տվյալներ են օրինակ՝ եկամուտը, բնակարան ունենալը, աշխատանք ունենալը, երեխաների քանակը և վարկի վերադարձնելու աղյուսակը, որտեղ նշված են 1, որը նշանակում է վարկը վերադարձրել է և 0՝ այն չի վերադարձրել:

Աղյուսակ 8.2, տվյալների բազա					
N	Եկամուտ	Բն.	Աշխ.	Երեխ.	Վարկ. վերադ.
1	10000	1	1	2	1
2	10000	1	1	2	0
3	10000	1	1	2	0
4	10000	1	1	2	1
5	10000	1	1	2	0
6	10000	1	1	2	1
7	15000	1	0	3	0
8	15000	1	0	3	0
9	15000	1	0	3	0
10	15000	1	0	3	0
11	15000	1	0	3	0
12	15000	1	0	3	1

Աղյուսակ 8.2-ից երևում է, որ միևնույն տվյալներով վարկառուները երբեմն վերադարձնում են վարկը, երբեմն էլ՝ ոչ: Եթե կատարենք դասակարգում, վերցնելով որոշիչ ֆունկցիան $a(\mathbf{x}, \mathbf{w}) = \text{sign}(\mathbf{w}, \mathbf{x})$ տեսքով, ապա այն միևնույն կոորդինատ ունեցող վեկտորի վրա պետք է տա երբեմն 1 արժեք, երբեմն էլ՝ -1: Պարզ է, որ դետերմինացված մոդելը նման հատկություն ունենալ չի կարող:

Այս դեպքերում որոշիչ ֆունկցիան կառուցվում է հետևյալ սկզբունքով: Դիտարկենք $P(y|x, w)$ պայմանական հավանականությունը, որը y -ի հանդես գալու հավանականությունն է, երբ մուտքում տրված է x վեկտորը: Օրինակ՝ եթե $x = (10000, 1, 1, 2)$, ապա

$$p(y = +1|x, \mathbf{w}) = \frac{3}{6} = 0.5, \quad p(y = 0|x, \mathbf{w}) = \frac{3}{6} = 0.5:$$

Լոգիստիկ ռեգրեսիայում վերցնում են

$$P(y|x, \mathbf{w}) \equiv \frac{1}{1 + e^{-(\mathbf{w}, \mathbf{x})y}} = \sigma(y(\mathbf{w}, \mathbf{x})):$$

Այժմ ենթադրենք, որ աղյուսակի բոլոր տողերը անկախ են հավանականության իմաստով: Այդ դեպքում կունենանք՝

$$P(x_1, x_2, \dots, x_l | y_1, y_2, \dots, y_l, w) = \prod_{i=1}^l P(x_i | y_i, w):$$

Առավելագույն հավանականության մեթոդով ընտրվում է w պարամետրը: Այսինքն՝ պետք է գտնել այնպիսի w , որ վերը նշված արտադրյալը ընդունի մաքսիմալ արժեք:

Այս խնդիրը համարժեք է մինիմիզացիայի հետևյալ խնդրին՝

$$\sum_{i=1}^l \ln(1 + e^{-(\mathbf{w}, \mathbf{x}_i)z_i}) \rightarrow \min_{\mathbf{w}} :$$

Լուծելով այս խնդիրը և գտնելով օպտիմալ w վեկտորը՝ $\sigma((\mathbf{w}, \mathbf{x})z)$ արտահայտության արժեքով գտնում ենք, թե ինչ հավանականությամբ x վեկտորը պատկանում է տվյալ դասին: Քանի որ $\sigma((\mathbf{w}, \mathbf{x})z)$ հավանականությունը մեծագույն արժեք ընդունում է $z = 1$ դեպքում, ապա x օբյեկտը դասվում է վարկ վերադարձնողների դասին: Այժմ քննարկենք ևս մի օրինակ:

Աղյուսակ 8.3, տվյալների բազա										
Ժամ	0.5	0.75	1	1.25	1.5	1.75	2	2.25	2.50	2.75
քնն. հանձ.	0	0	0	0	0	0	1	0	1	1
Ժամ	3	3.25	3.50	4	4.25	4.5	4.75	5	5.5	5.75
քնն. հանձ.	0	1	1	1	1	1	1	1	1	1

Աղյուսակի 8.3-ի առաջին տողում տրված է ուսանողի պարապելու ժամերի քանակը: Երկրորդում 1-ը նշանակում է, որ ուսանողը հանձնել է քննությունը, 0-ն՝ չի հանձնել:

Այստեղ ունենք երկու պարամետր (w_0, w_1) և մի հայտանիշ՝ x :

Այսինքն՝

$$h_w(x) = \frac{1}{1 + e^{-(w_0 + w_1 x)}} :$$

Կազմենք Էմպիրիկ կորստի ֆունկցիան՝

$$L(w_0, w_1) = \sum_{i=1}^{20} (y_i \ln p_i + (1 - y_i) \ln(1 - p_i)),$$

որտեղ

$$p_i = \frac{1}{1 + e^{-(w_0 + w_1 x_i)}} :$$

Համաձայն մինիմումի անհրաժեշտ և բավարար պայմանի՝

$$\frac{\partial L}{\partial w_0} = \sum_{i=1}^{20} (y_i - p_i) = 0,$$

$$\frac{\partial L}{\partial w_1} = \sum_{i=1}^{20} (y_i - p_i) x_i = 0:$$

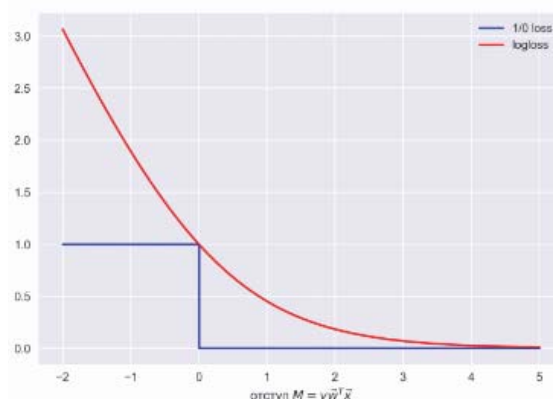
Լուծելով այս համակարգը w_0, w_1 պարամետրերի նկատմամբ՝ կստանանք՝

$$w_0 \approx -4.1, \quad w_1 \approx 1.5:$$

Օրինակ՝ երկու ժամ հետո քննությունը հանձնելու հավանականությունը հավասար է $p = \frac{1}{1+e^{-(w_0+2w_1)}} \approx 0.25$, իսկ չորս ժամ հետո՝

$$p = \frac{1}{1+e^{-(w_0+4w_1)}} \approx 0.87:$$

Վերջում նշենք, որ $\ln(1+e^{-M})$ ֆունկցիան հենքային $[M < 0]$ կորստի ֆունկցիային վերևից մոտարկող լոգարիթմական ողորկ ֆունկցիան է (նկ. 8.7):



Նկ. 8.7: $\frac{1}{2} \ln(1+e^{-M})$ և $[M < 0]$ ֆունկցիաների գրաֆիկները

8.7 Գծային ռեգրեսիայի մոդելի կայունության պայմանները

Այս պարագրաֆում քննարկվում է, թե ինչ պայմանների դեպքում գծային ռեգրեսիայի մոդելը կայուն է և միջին քառակուսային շեղման ալգորիթմը կառուցում է օպտիմալ պրեդիկտոր:

Դիցուք (x_i, y_i) , $i = 1, 2, \dots, l$ տվյալները գեներացվում են հետևյալ օրենքով՝

$$y_i = (w, x_i) + \epsilon_i, \quad i = 1, 2, \dots, l,$$

որտեղ ϵ_i , $i = 1, 2, \dots, l$ իրարից անկախ պատահական մեծությունները բավարարում են

$$E[\epsilon_i] = 0, \quad Var(\epsilon_i) = \sigma^2, \quad i = 1, 2, \dots, l$$

պայմանները: Իսկ x_i , $i = 1, 2, \dots, l$ օբյեկտները պատահական մեծություններ չեն: Կիրառենք միջին քառակուսային շեղման կորստի էմպիրիկ ֆունկցիան և գտնենք

$$\hat{w} = (XX^T)^{-1}X^Ty$$

վեկտորը: Բնականաբար այս վեկտորը պատահական մեծություն է: Հաշվենք դրա մաթեմատիկական սպասումը.

$$E[\hat{w}] = E[(XX^T)^{-1}]X^Ty = (XX^T)^{-1}XE[y] =$$

$$= (XX^T)^{-1}X^TE[Xw + \epsilon] = (XX^T)^{-1}XX^TE[w] = E[w]:$$

Վիճակագրությունում այս դեպքում ասում են, որ $\hat{w}$ վեկտորը w պարամետրի անշեղելի գնահատականն է:

Այժմ հաշվենք սպասվող սխալանքը ամբողջ հավաքածուի վրա՝

$$E[\text{error}_{in}] \equiv \left[\frac{1}{l} \sum_{i=1}^l (y_i - \hat{y}_i)^2 \right], \quad \hat{y}_i = (\hat{w}, x_i) = X\hat{w} = X(XX^T)^{-1}X^Ty = Hy:$$

Քանի որ H -ը պրոյեկտման օպերատոր է, ապա

$$(I - H)^2 = I - H - H - H^2 = I - H:$$

Այստեղից՝ կամայական $z = (z_1, z_2, \dots, z_l)$ վեկտորի համար ունենք

$$\|z - \hat{y}\|^2 = \|z - Hz\|^2 = \langle (I - H)z, (I - H)z \rangle = \langle (I - H)z, z \rangle:$$

Դիցուք այժմ $\xi = (\xi_1, \dots, \xi_l)$ վեկտորի կոմպոնենտները իրարից անկախ պատահական մեծություններ են, այնպիսին, որ $\text{Var}(\xi_i) = \sigma^2$, $i = 1, 2, \dots, l$: Դիցուք $D(l \times l)$ ֆիքսված մատրից է: Քանի որ

$$E[\xi_i \xi_j] = E[\xi_i]E[\xi_j], \quad i \neq j, \quad E[\xi_i^2] = \sigma^2 + (E[\xi_i])^2,$$

ապա

$$\begin{aligned} \langle D\xi, \xi \rangle &= \sum_{i=1}^m d_{ii}E[\xi_i^2] + \sum_{i \neq j} d_{ij}E[\xi_i \xi_j] = \\ &= \sigma^2 \sum_{i=1}^m d_{ii} + \sum_{i,j=1}^l d_{ij}E[\xi_i]E[\xi_j] = \\ &= \sigma^2 \text{Tr}(D) + (DE[\xi], E[\xi]): \end{aligned}$$

Տեղադրելով այս բանաձևում

$$D = I - H, \quad \xi = y$$

արժեքները՝ կստանանք

$$E[\|y - \hat{y}\|^2] = \sigma^2 \text{Tr}(I - H) + ((I - H)E[y], E[y]):$$

Ունենք նաև

$$\text{Tr}(I - A) = l - \text{Tr}(H) = l - \text{Tr}(X(XX^T)^{-1}X^T) = l - \text{Tr}(I_{d+1}) = l - d - 1,$$

$$(I - H)Ey = (I - H)Xw = (I - X(XX^T)^{-1}X^T)Xw = 0:$$

Այսպիսով կստանանք

$$E[\text{error}_{in}] = \frac{1}{l} \sigma^2 (l - d - 1) = \sigma^2 - \frac{d+1}{l} \sigma^2:$$

Գրականության մեջ այս գնահատականը կոչվում է insample գնահատական գծային ռեգրեսիայում: Այս գնահատականից հետևում է, որ նման շեղումը ավելի փոքր է, քան պատահական ξ մեծությունների դիսպերսիան: Երբ $l \rightarrow \infty$, ապա այս շեղումը ձգտում է այդ դիսպերսիային, այսինքն՝ σ^2 -ին: Այժմ հաշվենք սպասվող սխալանքը հավաքածուից դուրս (outsample): Դիտարկվում է թեստային հավաքածու, որի օբյեկտները նույնն են, բայց պատասխանները գեներացվում են հետևյալ բանաձևով՝

$$y'_i = (w, x_i) + \epsilon'_i,$$

որտեղ ϵ'_i պատահական մեծությունները իրարից անկախ են, և

$$E[\epsilon_i] = E[\epsilon'_i], \text{Var}(\epsilon_i) = \text{Var}(\epsilon'_i) = \sigma^2:$$

Սպասվող սխալանքը թեստայինի վրա որոշվում է հետևյալ բանաձևով՝

$$E[\text{error}_{out}] = \frac{1}{l} E\left[\sum_{i=1}^l (y'_i - \hat{y}_i)^2\right], \hat{y}_i = (\hat{w}_i, x_i):$$

Քանի որ

$$\text{Cov}(y'_i, y_i) = 0, E y_i = E y'_i, \text{Var}(y'_i) = \text{Var}(y_i),$$

ապա

$$\begin{aligned} E[(y_i - \hat{y}_i)^2] &= \text{Var}(y_i - \hat{y}_i) + (E[y_i] - E[\hat{y}_i])^2 = \\ &= \text{Var}(y_i) + \text{Var}(y'_i) - 2\text{Cov}(y_i, \hat{y}_i) + (E[y_i] - E[\hat{y}_i])^2, \\ (E[y'_i] - E[\hat{y}_i])^2 &= \text{Var}(y'_i - \hat{y}_i) + (E[y'_i] - E[\hat{y}_i])^2 = \\ &= \text{Var}(y'_i) + \text{Var}(\hat{y}_i) - 2\text{Cov}(y'_i, \hat{y}_i) + (E[y'_i] - E[\hat{y}_i])^2 = \\ &= \text{Var}(y_i) + \text{Var}(\hat{y}_i) + (E[y_i] - E[\hat{y}_i])^2: \end{aligned}$$

Հետևաբար՝

$$E[(y'_i - \hat{y}_i)^2] = E[(y_i - \hat{y}_i)^2] + 2\text{Cov}(y_i, \hat{y}_i):$$

Մյուս կողմից ունենք

$$\text{Cov}(y_i, \hat{y}_i) = \text{Cov}(y_i, (HY)_i) = \sum_{j=1}^l H_{ij} \text{Cov}(y_i, y_j) = H_{ii} \text{Var}(y_i) = H_{ii} \sigma^2:$$

Այսպիսով,

$$\begin{aligned} E[\text{error}_{out}] &= E\left[\frac{1}{l} \sum_{i=1}^l (y'_i - \hat{y}_i)^2\right] = \\ &= E\left[\frac{1}{l} \sum_{i=1}^l (y_i - \hat{y}_i)^2\right] + \frac{2}{l} \sigma^2 \text{Tr}(H) = E[R_{in}] + \frac{2}{l} \sigma^2 (d+1): \end{aligned}$$

Ուստի՝ եթե ցանկացած $\epsilon > 0$ թվի համար, վերցնենք l երկարությամբ կամայական

երկու հավաքածու՝ (In) (ուսուցանող) և (Out) (թեստային), այնպիսին, որ

$$l > \frac{2}{\epsilon} \sigma^2 (d + 1),$$

ապա

$$E[\text{error}_{Out}] \leq E[\text{error}_{In}] + \epsilon:$$

Այս դեպքում ասում են, որ գծային ռեգրեսիա խնդրում $h(x) \equiv (\hat{w}, x)$ պրեդիկտորը **կայուն** է:

Այսպիսով, դիտարկվող այս խնդրում կայուն պրեդիկտորի գոյության բավարար պայման է XX^T մատրիցայի հակադարձելիությունը: :

8.8 Հավանականային խտության և բաշխման ֆունկցիաների գնահատման մեթոդները

Երբ անհրաժեշտ է կառուցել որևէ հավանականային դասակարգիչ, ապա պետք է իմանալ յուրաքանչյուր դասի հանդես գալու $p(A)$ հավանականությունը և այդ դասերում հավանականության խտության $f(x|A)$ ֆունկցիան: Հայտնի է նաև՝ նմուշների որևէ հավաքածու $X = \{x_1, x_2, \dots, x_N\}$, որի էլեմենտները հավանականային $f(x|A)$ խտությունն ունեցող պատահական մեծության արժեքներ են: Պահանջվում է նմուշների տրված հավաքածուով գնահատել այդ հավանականությունը:

Այստեղ հնարավոր են հետևյալ երկու դեպքերը.

- Հայտնի է խտության ֆունկցիայի ընդհանուր տեսքը, և այն կախված է որոշ պարամետրերից: Պետք է գնահատել այդ պարամետրերը:
- Խտության ֆունկցիայի ընդհանուր տեսքը հայտնի չէ, բայց հայտնի են նրա որոշ հատկություններ: Օրինակ՝ անընդհատություն, դիֆերենցելիություն և այլն:

Դիտարկենք յուրաքանչյուրից մեկական օրինակ:

Դիցուք հայտնի է, որ պատահական մեծությունը ունի նորմալ բաշխում՝

$$f(x) = \frac{1}{(2\pi)^{n/2} \sigma^n} e^{-\frac{1}{2\sigma^2} \|x-m\|^2}:$$

Այստեղ մասնակցող σ , m պարամետրերի գնահատումը կատարենք մաքսիմալ հավանականության մեթոդով: Կազմենք առավելագույն հավանականության ֆունկցիան՝

$$\begin{aligned} L(m, \sigma | X) &= \prod_{i=1}^N f(x_i, |m, \sigma) = \\ &= \left(\frac{1}{(2\pi)^{n/2} \sigma^n} \right)^N \prod_{i=1}^N e^{-\frac{1}{2(\sigma)^2} \|x_i - m\|^2}: \end{aligned}$$

Լոգարիթմելով այս ֆունկցիան և հավասարեցնելով նրա գրադիենտը (ըստ σ և m պարամետրերի) գրոյի՝ կստանանք՝

$$\tilde{m} = \frac{1}{N} \sum_{i=1}^N x_i,$$

$$\tilde{\sigma}^2 = \frac{1}{nN} \sum_{i=1}^N \|x_i - \tilde{m}\|^2:$$

Այժմ դիտարկենք երկրորդ տիպի օրինակ:

Եթե η պատահական մեծությունը ունի $f(x)$ հավանականային խտություն, ապա D տիրույթում նրա արժեքների ընկնելու հավանականությունը հաշվում են հետևյալ բանաձևով՝

$$P(D) = \int_D f(x) dx:$$

Ենթադրենք՝ ունենք նմուշների $X_N = \{x_1, x_2, \dots, x_N\}$ հավաքածու, որի էլեմենտները η պատահական մեծության արժեքներն են: Այդ հավաքածուի օգնությամբ ուզում ենք գնահատել այդ պատահական մեծության խտության ֆունկցիան ինչ-որ մի կետի շրջակայքում: Դիցուք D -ն x_0 կետը պարունակող մի տիրույթ է, իսկ $V(D)$ -ն նրա ծավալը: Այդ դեպքում, ըստ Բեռնուլիի բանաձևի, կունենանք՝

$$P(X_N = k) = C_N^k P(D)^k (1 - P(D))^{N-k},$$

ընդ որում՝ $E[X_N] = NP(D)$: Հետևաբար՝

$$P(D) = \frac{E[X_N]}{N} \approx \frac{k}{N}:$$

Մյուս կողմից, համաձայն միջին արժեքի վերաբերյալ թեորեմի՝

$$P(D) = \int_D f(x) dx = f(\hat{x})V(D),$$

որտեղ $V(D)$ -ն D տիրույթի ծավալն է, իսկ $\hat{x}$ -ն՝ այդ տիրույթի ինչ-որ կետ:

Այսպիսով, որպես $f(x)$ ֆունկցիայի գնահատական կարող է հանդես գալ

$$\frac{k}{NV(D)}:$$

մեծությունը:

Այժմ քննարկենք, թե որքանով է ճշգրիտ այդ գնահատականը:

Տեղի ունի խտության ֆունկցիայի լոկալ գնահատման հետևյալ թեորեմը:

Թեորեմ 8.3. Դիցուք

- $f(x)$ խտության ֆունկցիան անընդհատ է x_0 կետում:
- x_0 կետը պարունակող D_N , տիրույթների հաջորդականության համար տեղի

ունեն հետևյալ պայմանները՝

$$\lim_{N \rightarrow \infty} V(D_N) = 0, \quad \lim_{n \rightarrow \infty} NV(D_N) = \infty:$$

• **Դիցուք**

$$\tilde{f}^N(x) \equiv \frac{k_N}{NV(D_N)}, \quad x \in D_N$$

Այդ դեպքում

$$P(|\tilde{f}^N(x) - f(x)| > \epsilon) \rightarrow 0, \quad \text{երբ } N \rightarrow \infty:$$

Ապացույց: X_N հավաքածուի արժեքներից k_N հատը D_N տիրույթում հայտնվելու հավանականությունը կարելի է ներկայացնել հետևյալ կերպ՝

$$k_N = \sum_{i=1}^N \varphi_N(x_i),$$

որտեղ $\varphi_N(x)$ -ն D_N տիրույթի ինդիկատոր ֆունկցիան է: Այդ դեպքում

$$\tilde{f}^N(x) = \frac{1}{NV(D_N)} \sum_{i=1}^N \varphi_N(x_i):$$

Քանի որ բոլոր $\varphi_N(x_i)$ պատահական մեծությունները ունեն նույն բաշխման ֆունկցիան, ապա

$$\begin{aligned} E[\tilde{f}^N] &= \frac{1}{NV(D_N)} \sum_{i=1}^N E[\varphi_N(x_i)] = \frac{E[\varphi_N(\eta)]}{V(D_N)} = \\ &= \frac{1}{V(D_N)} \int_{\mathbb{R}^n} \varphi_N(v) f(v) \mathbf{d}v = \frac{1}{V(D_N)} \int_{D_N} f(v) \mathbf{d}v = f(\tilde{x}_N) \rightarrow f(x_0), \quad \text{երբ } N \rightarrow \infty: \end{aligned}$$

Այսպիսով $\tilde{f}^N$ -ը f -ի անշեղելի գնահատականն է:

Գնահատենք դիսպերսիան

$$\begin{aligned} \sigma^2(\tilde{f}^N) &= \frac{1}{N^2 V^2(D_N)} \sum_{i=1}^N \sigma^2(\varphi_N(x_i)) = \\ &= \frac{1}{NV^2(D_N)} \sigma^2(\varphi_N(\eta)) = \\ &= \frac{1}{NV^2(D_N)} \left(\int_{D_N} f(v) \mathbf{d}v - \left(\int_{D_N} f(v) \mathbf{d}v \right)^2 \right) = \\ &= \frac{f(\tilde{x}_N)}{NV(D_N)} (1 - V(D_N) f(\tilde{x}_N)) \rightarrow 0: \end{aligned}$$

Այսպիսով $\tilde{f}^N$ -ը էֆեկտիվ ասիմպտոտիկ գնահատական է խտության f ֆունկցիայի համար: Համաձայն մեծ թվերի օրենքի՝ կարելի է հաշվել, որ մեծ N -երի դեպքում $\tilde{f}^N$ պատահական մեծությունը նորմալ է բաշխված: Հաշվի առնելով

նաև միջին արժեքի և դիսպերսիայի ստացված բանաձևերը՝ կատանանք՝

$$\begin{aligned} P(|\tilde{f}^N(x) - f(x_0)| > \epsilon) &= 1 - P(\tilde{f}^N(x) < f(x_0) + \epsilon) + P(\tilde{f}^N(x) < f(x_0) - \epsilon) = \\ &= 1 - \Phi\left(\frac{(f(x_0) + \epsilon - f(\tilde{x}_N))\sqrt{NV(D_N)}}{\sqrt{f(\tilde{x}_N)(1 - V(D_N)f(\tilde{x}_N))}}\right) + \\ &\quad + \Phi\left(\frac{(f(x_0) - \epsilon - f(\tilde{x}_N))\sqrt{NV(D_N)}}{\sqrt{f(\tilde{x}_N)(1 - V(D_N)f(\tilde{x}_N))}}\right), \end{aligned}$$

որտեղ

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt:$$

Այստեղից, հաշվի առնելով թեորեմի պայմանները՝ կատանանք

$$P(|\tilde{f}^N(x) - f(x_0)| > \epsilon) \rightarrow 0, \text{ երբ } N \rightarrow \infty:$$

Այժմ դիտարկենք մասնավոր դեպք: Որպես x_0 կետը պարունակող D_N տիրույթների վերցնենք այդ կենտրոնով ϵ_N շառավղով գնդեր՝

$$B_{\epsilon_N}(x_0) = \{x \in \mathbb{R}^n : \|x - x_0\| \leq \epsilon_N\}:$$

Դիցուք $\varphi(v)$ -ն այդ հիպերգնդի բնութագրիչ ֆունկցիան է: Այդ դեպքում

$$k(x_0, \epsilon_N) = \sum_{i=1}^N \varphi\left(\frac{x_0 - x_i}{\epsilon_N}\right)$$

արտահայտությունը ցույց կտա հավաքածուի այն էլեմենտների քանակը, որոնք ընկած են այդ շրջակայքում: Ուստի որպես գնահատող ֆունկցիա կարելի է դիտարկել հետևյալը՝

$$\tilde{f}^N(x) = \frac{1}{NV_N} \sum_{i=1}^N \varphi\left(\frac{x - x_i}{\epsilon_N}\right), \quad x \in \mathbb{R}^n:$$

Համաձայն n -չափանի գնդի ծավալի բանաձևի՝

$$V_N = C(n)\epsilon_N^n,$$

որտեղ $C(n)$ -ը ինչ-որ հաստատուն է: Հետևաբար՝ թեորեմ 8.3-ի պայմանները կկատարվեն, եթե, օրինակ՝ վերցնենք

$$\epsilon_N = \frac{a}{\ln(N+1)}, \quad a > 0:$$

Հավանականային խտության ֆունկցիայի լոկալ գնահատման այս եղանակը կոչվում է պարզենովյան մեթոդ:

Այժմ ցույց տանք, որ այսպես ասած Էմպիրիկ բաշխման ֆունկցիան իրական բաշխման ֆունկցիայի գլոբալ գնահատականն է: Դա արտահայտվում է Գլիվենկոյի հայտնի թեորեմով: Դիցուք ունենք X պատահական մեծություն է, և N անկախ փորձերի ելքերում ստացվել են $x_1 < x_2 < \dots < x_k$ տարբեր թվերը:

Ընդ որում ենթադրենք այս վեկտորում x_i արժեքի հանդես գալու հաճախությունը հավասար է n_i -ի: Հետևաբար՝ $\sum_{i=1}^k n_i = N$: Կառուցենք աստիճանաձև հետևյալ ֆունկցիան՝

$$\tilde{F}^N(x) = \begin{cases} 0, & x < x_1 \\ \frac{n_1+n_2+\dots+n_i}{N}, & x \in (x_i, x_{i+1}], \quad i = 0, \dots, k-1, \\ 1, & x > x_k, \end{cases}$$

որը կոչվում է Էմպիրիկ բաշխման ֆունկցիա:

Օրինակ: 8.4 աղյուսակում տրված են X պատահական մեծության արժեքները իրենց համապատասխան հաճախություններով: Կառուցենք Էմպիրիկ բաշխման ֆունկցիան: Քանի որ

$$N = 4 + 10 + 6 + 8 + 7 + 5 = 40,$$

ապա

$$\tilde{F}(x \leq 1) = \frac{0}{40},$$

$$\tilde{F}(1 < x \leq 2) = \frac{0+4}{40} = 0.1,$$

$$\tilde{F}(2 < x \leq 3) = \frac{4+10}{40} = 0.35,$$

$$\tilde{F}(3 < x \leq 4) = \frac{4+10+6}{40} = 0.5,$$

$$\tilde{F}(4 < x \leq 5) = \frac{4+10+6+8}{40} = 0.7,$$

$$\tilde{F}(5 < x \leq 6) = \frac{4+10+6+7+5}{40} = 0.875,$$

$$\tilde{F}(x > 6) = 1:$$

Աղյուսակ 8.4, արժեքներ և հաճախություններ

x_i	1	2	3	4	5	6
n_i	4	10	6	8	7	5

Թեորեմ 8.4. (Գլիվենկոյի թեորեմը):

Դիցուք $F(x)$ -ը X պատահական X մեծության բաշխման ֆունկցիան է, իսկ $\tilde{F}^N(x)$ -ը՝ Էմպիրիկ բաշխման ֆունկցիան: Եթե $F(x)$ -ը անընդհատ է, ապա ցանկացած $\epsilon > 0$ թվի համար

$$\lim_{N \rightarrow \infty} P(\sup_x |\tilde{F}^N(x) - F(x)| < \epsilon) = 1:$$

Ապացույց: Դիցուք $\epsilon > 0$ նախապես տրված դրական թիվ է, իսկ S -ը այնպիսի բնական թիվ է, որ

$$\frac{1}{S} < \frac{\epsilon}{2}:$$

Թող $\bar{x}_i$ -ին լինի $F(x_i) = i/S$, $i = 1, 2, \dots, S$ հավասարման արմատը: Այդպիսի արմատ գոյություն ունի, քանի որ $F(x)$ բաշխման ֆունկցիան ըստ ենթադրության անընդհատ է և մոնոտոն՝ համաձայն սահմանման: Ուստի՝

$$F(\bar{x}_i) - F(\bar{x}_{i-1}) = \frac{1}{S} < \frac{\epsilon}{2}:$$

Դիցուք

$$\bar{x}_i \leq x \leq \bar{x}_{i-1}:$$

Այդ դեպքում, հաշվի առնելով, որ $\tilde{F}^N(x)$ և $F(x)$ ֆունկցիաները չնվազող են, կունենանք՝

$$\begin{aligned} \tilde{F}^N(\bar{x}_{i-1}) - F(x_i) &\leq \tilde{F}^N(x) - F(x) \leq \\ &\leq \tilde{F}^N(\bar{x}_i) - F(\bar{x}_{i-1}): \end{aligned}$$

Այստեղից, հաշվի առնելով նաև վերևի անհավասարությունը, կունենանք՝

$$\begin{aligned} \tilde{F}^N(\bar{x}_{i-1}) - F(\bar{x}_{i-1}) - \frac{\epsilon}{2} &\leq \\ &\leq \tilde{F}^N(x) - F(x) \leq \\ &\leq \tilde{F}^N(\bar{x}_i) - F(\bar{x}_i) + \frac{\epsilon}{2}: \end{aligned}$$

Հետևաբար՝

$$\sup_x |\tilde{F}^N(x) - F(x)| \leq \max_{1 \leq i \leq S} |\tilde{F}^N(\bar{x}_i) - F(\bar{x}_i)| + \frac{\epsilon}{2}:$$

Նշանակենք՝ $P_i = P(X < \bar{x}_i)$: Թող m_i -ին լինի X պատահական մեծության այն արժեքների թիվը, որոնք փոքր են $\bar{x}_i$ -ից: Պարզ է, որ

$$P_i = F(\bar{x}_i), \quad \frac{m_i}{N} = \tilde{F}^N(\bar{x}_i):$$

Ուստի կունենանք

$$\sup_x |\tilde{F}^N(x) - F(x)| \leq \max_{1 \leq i \leq S} \left| \frac{m_i}{N} - P_i \right| + \frac{\epsilon}{2}:$$

Դիտարկենք

$$\max_{1 \leq i \leq S} \left| \frac{m_i}{N} - P_i \right| \geq \frac{\epsilon}{2}$$

պատահաբար: Պարզ է, որ այն տեղի ունի այն և միայն այն դեպքում, երբ տեղի ունի

$$\left| \frac{m_i}{N} - P_i \right| \geq \frac{\epsilon}{2}, \quad i = 1, 2, \dots, S$$

պատահարներից գոնե մեկը:

Հետևաբար՝

$$P(\max_{1 \leq i \leq S} |\frac{m_i}{N} - P_i| \geq \frac{\epsilon}{2}) \leq \sum_{i=1}^S P(|\frac{m_i}{N} - P_i| \geq \frac{\epsilon}{2}):$$

Բեռնուլի թեորեմի համաձայն՝ N -ը վերցնենք այնքան մեծ, որ աջ մասի գումարի յուրաքանչյուր հավանականություն լինի փոքր ցանկացած α/S թվից: Ուստի՝

$$P(\max_{1 \leq i \leq S} |\frac{m_i}{N} - P_i| < \frac{\epsilon}{2}) > 1 - \alpha:$$

Արդյունքում ցանկացած ϵ և α թվերի համար կստանանք

$$P(\sup_x |\tilde{F}^N(x) - F(x)| < \epsilon) > 1 - \alpha,$$

կամ

$$\lim_{N \rightarrow \infty} P(\sup_x |\tilde{F}^N(x) - F(x)| < \epsilon) = 0:$$

Օգտագործելով Հոֆդինգի անհավասարությունը կարելի է ստանալ հետևյալ գնահատականը.

Թեորեմ 8.5. Դիցուք $\{X_i\}_{i=1}^\infty$ միևնույն բաշխման ֆունկցիա ունեցող պատահական մեծությունների հաջորդականություն է: Այսինքն՝ $F(x) = P(X_i \leq x)$, $\forall i$:

Դիցուք

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{[X_i \leq x]}$$

Էմպիրիկ բաշխման ֆունկցիան է:

Այդ դեպքում կամայական $\epsilon > 0$ թվի համար տեղի ունի հետևյալ անհավասարությունը՝

$$P\left(\sup_{x \in R} |f(x) - F_n(x)| > \epsilon\right) \leq 8(n+1) \exp\left(-\frac{n\epsilon^2}{32}\right):$$

Ուստի, եթե փորձերի քանակը մեծ է, ապա էմպիրիկ բաշխման ֆունկցիան կարող է փոխարինել իրական բաշխման ֆունկցիային:

8.9 Գծային տարբերակիչ (դիսկրիմինանտ) անալիզ

Այժմ քննարկենք թե ինչպես կարելի է կառուցել Բայեսի օպտիմալ $h^*(x)$ պրեդիկտորը այն դեպքում, երբ օբյեկտները դասերում ունեն բազմաչափ գաուսյան նորմալ բաշխում: Այսինքն՝

$$P(x|y) = \frac{1}{\sqrt{(2\pi)^d}(\det \Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_y)^T \Sigma^{-1}(x - \mu_y)\right), \quad y \in Y:$$

Ինչպես արդեն գիտենք, $h^*(x) = \max_{y \in Y} P(x, y)$: Դիտարկենք բինար դասակարգման դեպքը՝ $Y = \{0, 1\}$: Ցույց տանք, որ այս պարագայում կառուցված Բայեսի օպտիմալ

կլասիֆիկատորը դասերը բաժանող հիպերհարթությունն է:

Դա կատարվում է հետևյալ ընթացակարգով: Երկու տարրեր $y, y' \in Y$ նշիչների համար համեմատվում են ապոստրոփոր հավանականությունները՝

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} > P(y'|x) = \frac{P(x|y')P(y')}{P(x)}:$$

Տեղադրելով այստեղ պայմանական հավանականությունների նորմալ բաշխման բանաձևը և լոգարիթմելով այն՝ կստանանք՝

$$(\mu_y - \mu_{y'})^T \Sigma^{-1} x > \frac{1}{2} \mu_y^T \Sigma^{-1} \mu_y - \frac{1}{2} \mu_{y'}^T \Sigma^{-1} \mu_{y'} - \ln P(y) + \ln P(y') \equiv c:$$

Այսպիսով կարող ենք եզրակացնել, որ այդ դասերը անջատվում են $(w, x) = c$ հիպերհարթությամբ, որտեղ $w = (\mu_y - \mu_{y'})^T \Sigma^{-1}$: Յուրաքանչյուր դասի համար ներմուծենք գծային տարբերակիչ ֆունկցիայի գաղափարը՝

$$\delta_y(x) = -\frac{1}{2}(x - \mu_y)^T \Sigma^{-1}(x - \mu_y) + \frac{1}{2} x^T \Sigma^{-1} x + \ln P(y) =$$

$$\mu_y^T \Sigma^{-1} x - \frac{1}{2} \mu_y^T \Sigma^{-1} \mu_y + \ln P(y):$$

Օպտիմալ կլասիֆիկատորը կլինի $h^*(x) = \arg \max_y \delta_y(x)$ ֆունկցիան, իսկ դասերը անջատող հիպերհարթությունը կլինի $\delta_1(x) = \delta_0(x)$ հավասարումով տրվող հիպերհարթությունը:

Գրականության մեջ նման կերպ կառուցված պրեդիկտորը կոչվում է գծային տարբերակիչ անալիզ:

Պրակտիկայում, ընդհանրապես ասած, նորմալ բաշխման պարամետրերը հայտնի չեն: Դրանք գնահատվում են ուսուցանող հավաքածուի օգնությամբ՝ հետևյալ բանաձևերով՝

$$\hat{P}(y) = \frac{N_y}{N}, \quad \hat{\mu}_y = \sum_{y^i=k} \frac{x^{(i)}}{N_y},$$

$$\hat{\Sigma} = \frac{1}{N-2} \sum_{k=0}^1 \sum_{y^{(i)}=k} (x^{(i)} - \hat{\mu}_k)^T (x^{(i)} - \hat{\mu}_k):$$

Այս բանաձևերում N -ը օբյեկտների քանակն է, իսկ N_y -ն այն օբյեկտների քանակն է, որոնք ունեն y նշիչը:

Այժմ օրինակով մեկնաբանենք այս ալգորիթմը:

Դիցուք տրված է ուսուցանող հավաքածու (աղյուսակ 8.8): Յուրաքանչյուր դասի համար կառուցել տարբերակիչ ֆունկցիա և գտնել անջատող հիպերհարթության հավասարումը: Հաշվենք դասերի նմուշային միջինները և նմուշային կովարիացիոն մատրիցը՝

$$P(Y=0) = P(Y=1) = \frac{1}{2},$$

$$\hat{\mu}_0 = (1, 4), \quad \hat{\mu}_1 = (4, 1),$$

x_1	0	2	1	1	1	2	4	4	4	6
x_2	4	4	3	5	4	1	0	2	1	1
y	0	0	0	0	0	1	1	1	1	1

Նկ. 8.8: Ուսուցանող հավաքածու

$$\begin{aligned}\hat{\Sigma} &= \frac{1}{8} \sum_k \sum_{y^i=k} (x^{(i)} - \hat{\mu}_k)^T (x^{(i)} - \hat{\mu}_k) = \\ &= \frac{1}{8} \left(\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} + \begin{pmatrix} 8 & 0 \\ 0 & 2 \end{pmatrix} \right) = \begin{pmatrix} 5/4 & 0 \\ 0 & 1/2 \end{pmatrix}:\end{aligned}$$

Կառուցենք յուրաքանչյուր դասի տարբերակիչ ֆունկցիան՝

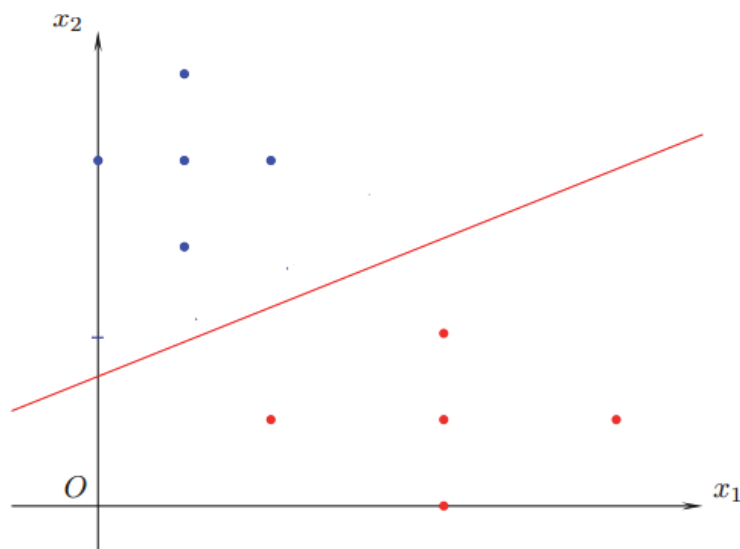
$$\delta_0(x) = x^T \hat{\Sigma}^{-1} \hat{\mu}_0 - \frac{1}{2} \hat{\mu}_0^T \hat{\Sigma}^{-1} \hat{\mu}_0 + \ln P(Y = 0) =$$

$$\frac{4}{5}x_1 + 8x_2 - \frac{82}{5} - \ln 2:$$

Անալոգիայով ստանում ենք՝

$$\delta_1(x) = \frac{16}{5}x_1 + 2x_2 - \frac{37}{5} - \ln 2:$$

$\delta_0(x) = \delta_1(x)$ հավասարումով տրվող ուղիղ գիծը կլինի անջատող հիպերհարթությունը: Այսինքն՝ դասերը բաժանող գծի հավասարումը հետևյալն է՝ $4x_1 - 10x_2 + 15 = 0$ (տե՛ս նկ.8.9):



Նկ. 8.9: Անջատող հիպերհարթություն

- Տրված է ուսուցանող հավաքածու (աղյուսակ (8.10): Բայեսի պարզ դասակարգչի (naive classifier) օգնությամբ հաշվել հետևյալ հավանականությունները՝

$$P(Y = 0|X_1 = 1, X_2 = 1) \quad P(Y = 1|X_1 = 1, X_2 = 1):$$

2. Տրված է ուսուցանող հավաքածու (աղյուսակ (8.10): Գծային տարբերակիչ անալիզի մեթոդով կառուցել յուրաքանչյուր դասի տարբերակիչ ֆունկցիան և գտնել անջատող հիպերհարթության հավասարումը:

x_1	0	0	1	1	0	0	1	1	1	0
x_2	0	1	0	1	1	1	1	1	1	1
y	0	0	0	0	0	1	1	1	1	1

Նկ. 8.10: Ուսուցանող հավաքածու

8.10 Բազմաստիճան դասակարգում (multi-class classification)

Մինչ այժմ մենք դիտարկում էինք բինար դասակարգման խնդիրներ: Մակայն պրակտիկայում կան խնդիրներ, երբ անհրաժեշտ է որոշել, թե օբյեկտներն են K , ($k > 2$) դասերից որին են պատկանում: Այս խնդրի լուծման համար կա լայն տարածում ստացած երկու մոտեցում:

- մեկը բոլորի դեմ (one-vs-all),
- բոլորը բոլորի դեմ (all-vs-all):

Առաջին դեպքում կառուցվում են K ալգորիթմներ, որոնցից ամեն մեկը որոշակի դաս անջատում է մնացած բոլոր դասերից՝ $a_i(x, w_i) = 1 \rightarrow x \in C_i$, $i = 1, 2, \dots, K$:

Օրինակ՝ նկ. 8.11-ում a_1 ալգորիթմը C_1 դասը անջատում է C_2 , C_3 դասերից: a_2 -ը անջատում է C_2 -ը C_1 , C_3 դասերից: Իսկ a_3 -ն անջատում է C_3 -ը C_1 և C_2 դասերից: Տվյալ օրինակում ալգորիթմը որոշվում է հետևյալ առնչությամբ՝

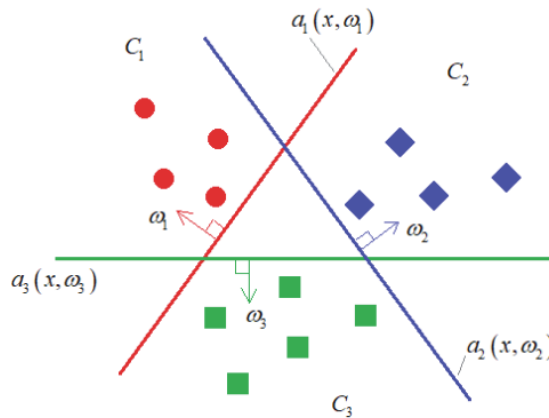
$$a_k(x, w_k) = \text{sign}((x, w_k) - w_{0k}), \quad \|w_k\| = 1, \quad k = 1, 2, 3:$$

Դրանից հետո սահմանում է նոր պրեդիկտոր՝

$$a(x) = \arg \max_{i \in [1:K]} ((x, w_i) - w_{0i}):$$

Այսինքն՝ a պրեդիկտորը առավելություն է տալիս այն ալգորիթմին, որի դեպքում x օբյեկտը առավել է հեռացված անջատող հիպերհարթությունից:

Երկրորդ դեպքում յուրաքանչյուր գույգ (i, j) , $i \neq j$ ինդեքսների դեպքում կառուցվում է C_i և C_j դասերը անջատող $a_{i,j}$ ալգորիթմ: Տրված օբյեկտի համար կազմում ենք $a_{i,j}(x)$ մատրիցը ($a_{i,i}(x) = 0$ պայմանով): Ընտրում ենք այն համարի տողը, որում մեկերի քանակն առավելագույն է: Հենց այդ համարով դասին էլ դասակարգում են x օբյեկտը: Այսինքն՝ դասը ընտրվում է մեծամասնական քվեարկության սկզբունքով:



Նկ. 8.11: Դասերի անջատում հիպերհարթություններով

Բոլոր այս ալգորիթմների դեպքում կարելի է որոշել x օբյեկտի k -րդ դասին պատկանելու հավանականությունը հետևյալ բանաձևով, որը կոչվում է SoftMax՝

$$p(a(x) = y_k) = \frac{\exp((x, w_k) - w_{0k})}{\sum_{j=1}^K \exp((w_j, x) - w_{0j})} :$$

Իսկ w_j , w_{0j} , $j = 1, 2, \dots, K$ պարամետրի արժեքներն ընտրվում են տարբեր կորստի ֆունկցիաների մինիմիզացիայի արդյունքում: Մասնավորապես հենման վեկտորների մեթոդով՝

$$Q(a, X) = \sum_{i=1}^l \sum_{k=1}^K [y_i = k] \ln(1 - y_i((w_k, x_i) - w_{0k})) \rightarrow \min_{w_k, w_{0k}} :$$

Այստեղ $X = \{x_1, x_2, \dots, x_l\}$ բազմությունը ուսուցանող հավաքածուն է:

Քննարկենք նաև առավելագույն հավանականությունների մեթոդի լոգարիթմական կորստի ֆունկցիան նույնպես:

Դիցուք

$$f_i = (w_i, x) + b_i, \quad S_i = \frac{\exp(f_i)}{\sum_{k=1}^K \exp(f_k)}, \quad b_i = -w_{0,i} :$$

$L = -\ln(S_k)$ ֆունկցիան կոչվում է k -րդ դասի լոգարիթմական կորստի ֆունկցիա: Պետք է մինիմիզացնել

$$L = - \sum_{i=1}^l \sum_{k=1}^K \ln(S_k(x_i))$$

Ֆունկցիան ըստ w_k, b_k պարամետրերի: Որպեսզի կատարվի գրադիենտային իջեցում, պետք է հաշվել այդ կորստի ֆունկցիայի ածանցյալները ըստ պարամետրերի: Համաձայն բարդ ֆունկցիայի ածանցման կանոնի՝ ունենք

$$\frac{\partial L}{\partial w_i} = \frac{\partial L}{\partial S_k} \frac{\partial S_k}{\partial f_i} \frac{\partial f_i}{\partial w_i}:$$

Հեշտ է տեսնել, որ

$$\frac{\partial S_k}{\partial f_i} = S_k(1 - S_i), \quad i = k,$$

$$\frac{\partial S_k}{\partial f_i} = -S_i S_k, \quad i \neq k,$$

$$\frac{\partial L}{\partial S_k} = -\frac{1}{S_k},$$

$$\frac{\partial f_i}{\partial w_i} = x:$$

$$\frac{\partial L}{\partial b_i} = S_i - 1, \quad i = k$$

$$\frac{\partial L}{\partial f_i} = S_i, \quad i \neq k:$$

Հետևաբար ստոխաստիկ գրադիենտային իջեցումը այս դեպքում կունենա հետևյալ տեսքը՝

$$w_i^{n+1} = w_i^n - \alpha \frac{\partial L}{\partial w_i}(w_i^n, x),$$

$$b_i^{n+1} = b_i^n - \frac{\partial L}{\partial b_i}(w_i^n, x), \quad i = 1, 2, \dots, k$$

Այստեղ x օբեյեկտը յուրաքանչյուր քայլում պատահականորեն է ընտրվում $X = \{x_1, x_2, \dots, x_l\}$ հավաքածուից: k -ն նույնպես ընտրվում է *random*:

Խնդիր: Ապացուցել, որ

$$l_k(w_1, w_2, \dots, w_l) = -\ln \frac{\exp((w_k, x))}{\sum_{i=1}^l \exp((w_i, x))}$$

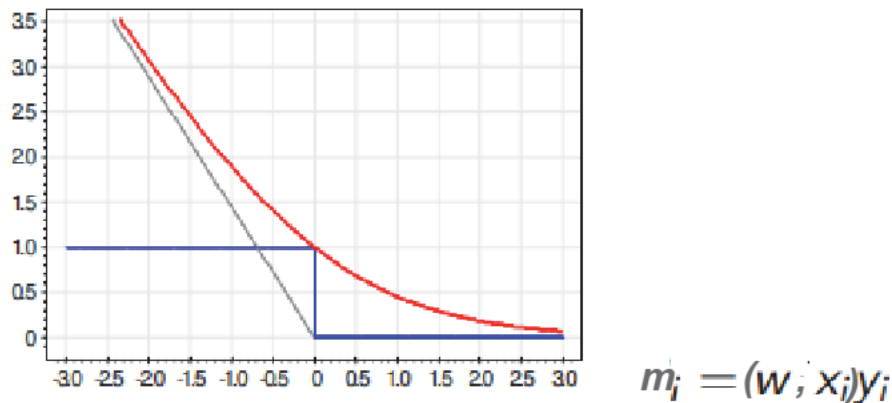
Ֆունկցիան ուռուցիկ է:

8.11 Նյուտոն-Ռաֆսոնի մեթոդը լոգիստիկ ռեգրեսիայում

Դիտարկենք բինար դասակարգման խնդիրը և էմպիրիկ ռիսկի ֆունկցիոնալը՝

$$Q(w) = \sum_{i=1}^l [m_i < 0],$$

որտեղ $m_i = (w, x_i)y_i$, $y_i \in Y = \{-1, 1\}$: Այս ֆունկցիան մոտարկենք վերևից լոգարիթմական կորստի $L(m_i) = \ln(1 + e^{m_i})$ ֆունկցիայով (տե՛ս նկ.8.12):



Նկ. 8.12: Հենքային ֆունկցիայի մոտարկում լոգարիթմական կորստի ֆունկցիայով

Կատարենք

$$\tilde{Q}(w) = \sum_{i=1}^l \ln(1 + e^{m_i})$$

ֆունկցիոնալի մինիմիզացիա ըստ w պարամետրի Նյուտոն-Ռաֆսոնի մեթոդով:

Նյուտոն-Ռաֆսոնի մեթոդը տվյալ դեպքում որոշվում է հետևյալ ռեկուրենտ առնչությամբ՝

$$w_{k+1} = w_k - \alpha_k (\tilde{Q}''(w_k))^{-1} \tilde{Q}'(w_k):$$

Հաշվենք $\tilde{Q}'$ գրադիենտը և $\tilde{Q}''$ հեսիանի կոմպոնենտները:

Դրա համար օգտվենք բարդ ֆունկցիայի ածանցման կանոնից, ինչպես նաև սիգմոյդի ածանցման բանաձևից: Ունենք

$$\frac{\partial \tilde{Q}}{\partial w_j} = - \sum_{i=1}^l (1 - \sigma_i) y_i f_j(x_i),$$

$$\frac{\partial^2 \tilde{Q}}{\partial w_j \partial w_k} = \sum_{i=1}^l (1 - \sigma_i) \sigma_i f_j(x_i) f_k(x_i), \quad j, k = 1, 2, \dots, n,$$

որտեղ

$$\sigma_i = \sigma(y_i(w, x_i)), \quad \sigma(z) = \frac{1}{1 + e^{-z}},$$

իսկ $f_j(x_i)$ -ը x_i վեկտորի j -րդ կոորդինատն է:

Դիցուք

$$F = (f_j(x_i))_{l \times n}, \quad D = \text{diag}((1 - \sigma_i) \sigma_i):$$

Այդ դեպքում

$$(\tilde{Q}'')^{-1}(w_k) \tilde{Q}'(w_k) = -(F^T D F)^{-1} \frac{y_i}{\sigma_i}:$$

Հեշտ է նկատել, որ այս հավասարության աջ մասը հետևյալ գծային ռեգրեսիայի

խնդրի լուծումն է՝

$$G(\omega) = \sum_{i=1}^l (1 - \sigma_i) \sigma_i \left((\omega, x_i) - \frac{y_i}{\sigma_i} \right)^2 \rightarrow \min_{\omega}:$$

Այժմ քայլ առ քայլ նկարագրենք Նյուտոն-Ռաֆսոնի մեթոդի ալգորիթմը տվյալ խնդրի համար:

- $w = (F^T F)^{-1} F^T y$
- $k = 1, 2, 3, \dots$ արժեքների համար կատարել հաշվարկներ՝
- $\sigma_i = \sigma(y_i(w, x_i)), i = 1, 2, \dots, l,$
- $\gamma_i = \sqrt{\sigma_i(1 - \sigma_i)}, i = 1, 2, \dots, l,$
- $\tilde{F} = \text{diag}(\gamma_1, \gamma_2, \dots, \gamma_l) F,$
- $\tilde{y}_i = y_i \sqrt{(1 - \sigma)/\sigma_i}, i = 1, 2, \dots, l,$
- $w = w + \alpha_k (\tilde{F}^T \tilde{F})^{-1} \tilde{F}^T \tilde{y},$
- եթե $\{\sigma_i\}$ արժեքները քիչ են փոխվում, ապա ցիկլից դուրս գալ:

Վերջում նշենք, որ Նյուտոն-Ռաֆսոնի մեթոդի զուգամիտության արագության պայմաններին մանրամասն կարելի է ծանոթանալ [46] աշխատանքում:

9

Ամրապնդմամբ ուսուցում (reinforcement learning)

9.1 Մարկովյան պրոցեսներ

Դիցուք ունենք մի համակարգ, որը ժամանակի t պահին կարող է գտնվել $\{s_1, s_2, s_3, \dots\}$ վիճակներից որևէ մեկում: Ենթադրենք նաև՝ որ հայտնի է մի վիճակից մյուս վիճակին անցնելու հավանականությունը: Ընդ որում այդ հավանականությունը բավարարում է հետևյալ պայմանը՝

$$P[s_{t+1}|s_t] = P[s_{t+1}|s_1, s_2, \dots, s_t]:$$

Սա նշանակում է, որ հավանականությունը, թե համակարգը s_t վիճակից կանցնի s_{t+1} վիճակի կախված չէ այն բանից, թե ինչ վիճակներում է այն եղել ժամանակի $1, 2, \dots, t-1$ պահերին: Այսինքն՝ անցյալի ինֆորմացիան հաշվի չի առնվում:

Դիցուք

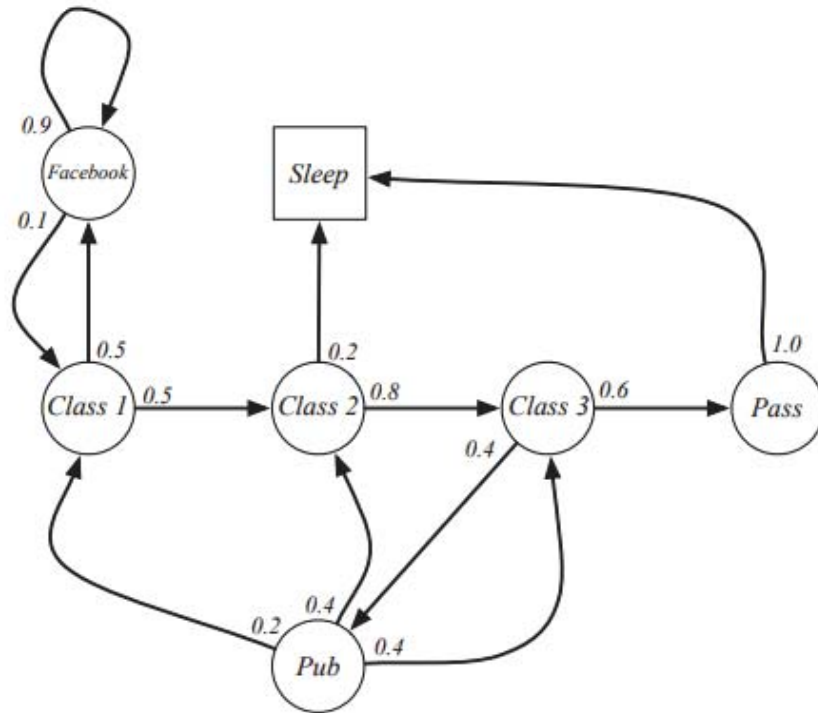
$$P_{ss'} = P[s_{t+1} = s' | s_t = s]:$$

Այս թվերից կազմված մատրիցը կոչվում է անցման մատրից (տե՛ս նկ. 9.1):

Այս մատրիցի կարևոր հատկությունն այն է, որ նրա յուրաքանչյուր տողի էլեմենտների գումարը հավասար է մեկի: Այժմ սահմանենք մարկովյան շղթա:

$$\mathcal{P} = \begin{matrix} & \text{to} \\ \text{from} & \begin{bmatrix} \mathcal{P}_{11} & \dots & \mathcal{P}_{1n} \\ \vdots & & \\ \mathcal{P}_{n1} & \dots & \mathcal{P}_{nn} \end{bmatrix} \end{matrix}$$

Նկ. 9.1: Անցման մատրից



Նկ. 9.2: Մարկովյան շղթայի օրինակ

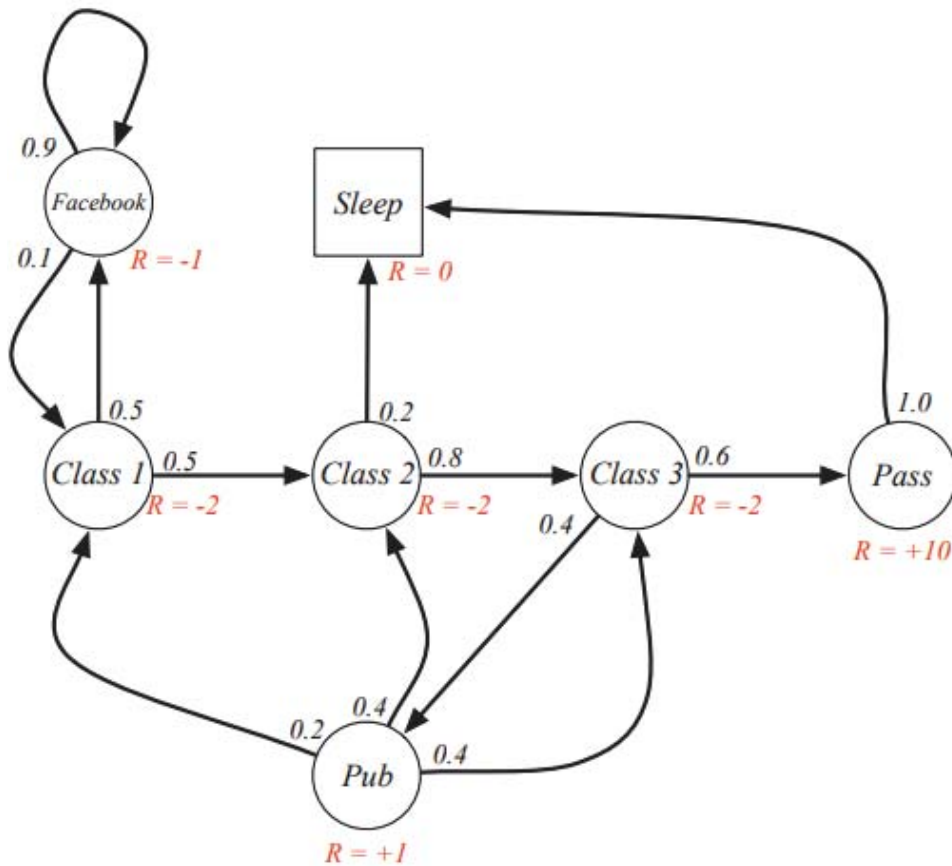
$$\mathcal{P} = \begin{matrix} & \begin{matrix} C1 & C2 & C3 & Pass & Pub & FB & Sleep \end{matrix} \\ \begin{matrix} C1 \\ C2 \\ C3 \\ Pass \\ Pub \\ FB \\ Sleep \end{matrix} & \left[\begin{array}{ccccccc} & & 0.5 & & & 0.5 & \\ & & & 0.8 & & & 0.2 \\ & & & & 0.6 & 0.4 & \\ & & & & & & 1.0 \\ 0.2 & 0.4 & 0.4 & & & & \\ 0.1 & & & & & 0.9 & \\ & & & & & & 1 \end{array} \right] \end{matrix}$$

Նկ. 9.3: Անցման մատրիցի կառուցվածքը

Մարկովյան շղթան որոշվում է (S, P) զույգով, որտեղ S -ը և $s_1, s_2, \dots$ վիճակների վերջավոր հաջորդականություն է, իսկ P -ն անցման մատրիցն է: Նման շղթայի օրինակ բերված է նկ. 9.2-ում, որում անցման P մատրիցն ունի նկար 9.3-ում ներկայացված կառուցվածքը:

Այժմ քննարկենք պարզևատրմամբ մարկովյան պրոցեսներ (Markov reward process): Այն որոշվում է հետևյալ չորս պարամետրերով՝ (S, P, T, γ) , որտեղ

- S -վիճակների վերջավոր բազմություն է,
- P -ն անցման հավանականային մատրիցն է,
- R -ը պարզևատրման ֆունկցիան է՝ $R_s = E[R_{t+1} | s_t = s]$,
- $\gamma \in [0, 1]$ -ն զեղչի գործակիցն է (discount factor):



Նկ. 9.4: Պարզևատրմամբ մարկովյան պրոցես

Այստեղ

$$[R_{t+1}|s_t = s] = \sum_{r \in R} \sum_{s' \in S} P(r, s'|s)r,$$

որտեղ $P(r, s'|s)$ հավանականությունն է այն բանի, որ համակարգը s վիճակից կտեղափոխվի s' վիճակի r պարգևատրմամբ:

Այդպիսի պրոցես պատկերված է նկ. 9.4-ում:

Այժմ յուրաքանչյուր վիճակի համար սահմանենք արժեքային ֆունկցիա: Դիցուք ունենք մարկովյան մի $s_1, s_2, \dots, s_t, s_{t+1}, \dots, s_T$ շղթա: Նշանակենք՝

$$G_t = R_{t+1} + \gamma R_{t+2} + \dots \gamma^{T-2} R_T :$$

Այս գումարը մարկովյան շղթայի ամբողջական գեղջով պարգևատրումն է ժամանակի t պահից հետո: Դիցուք $v(s) = E[G_t|s_t = s]$: Սա, այսպես ասած, ակնկալվող եկամտաբերությունն է՝ սկսած s գազաթից: Օրինակ՝ նկ. 9.5-ում պատկերված մարկովյան պրոցեսի համար հաշվենք եկամտաբերության ֆունկցիայի արժեքը $s_1 = \text{Class1}(C1)$ գազաթի համար $\gamma = 1/2$ գեղջի գործակցի դեպքում: Այդ գազաթից սկսած՝ կան հետևյալ շղթաները, որոնք ավարտվում են *Sleep* տերմինալ գազաթում (տե՛ս նկ. 9.5):

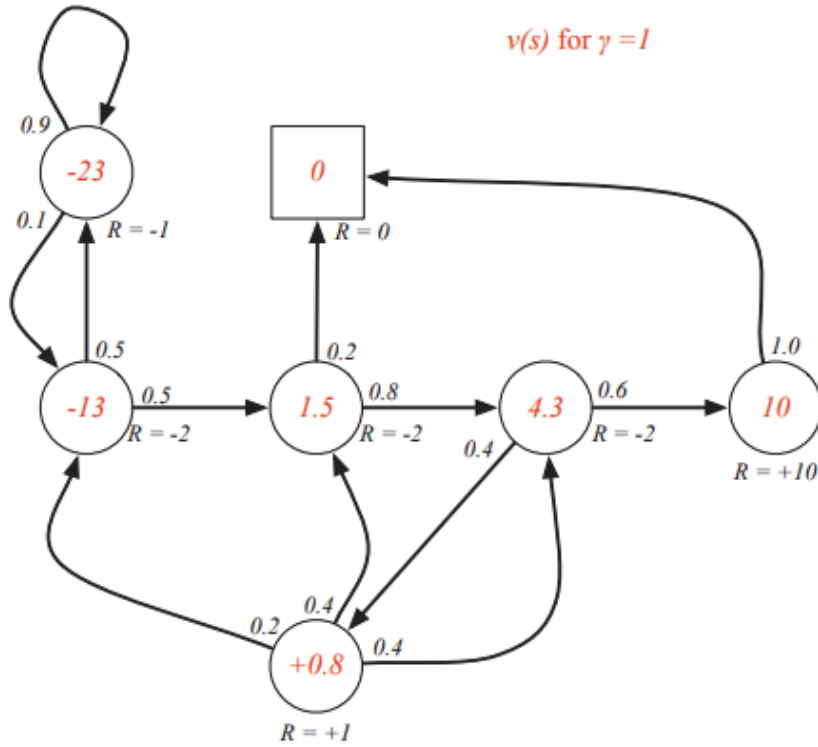
Հետևաբար՝

$$v(s_1) = \frac{-2.25 - 3.125 - 3.41 - 3.20}{4} = -2.99:$$

$$G_1 = R_2 + \gamma R_3 + \dots + \gamma^{T-2} R_T$$

C1 C2 C3 Pass Sleep	$v_1 = -2 - 2 * \frac{1}{2} - 2 * \frac{1}{4} + 10 * \frac{1}{8}$	=	-2.25
C1 FB FB C1 C2 Sleep	$v_1 = -2 - 1 * \frac{1}{2} - 1 * \frac{1}{4} - 2 * \frac{1}{8} - 2 * \frac{1}{16}$	=	-3.125
C1 C2 C3 Pub C2 C3 Pass Sleep	$v_1 = -2 - 2 * \frac{1}{2} - 2 * \frac{1}{4} + 1 * \frac{1}{8} - 2 * \frac{1}{16} \dots$	=	-3.41
C1 FB FB C1 C2 C3 Pub C1 ...	$v_1 = -2 - 1 * \frac{1}{2} - 1 * \frac{1}{4} - 2 * \frac{1}{8} - 2 * \frac{1}{16} \dots$	=	-3.20
FB FB FB C1 C2 C3 Pub C2 Sleep			

Նկ. 9.5: $v(s_1)$ հաշվումը ($\gamma = 1/2$)



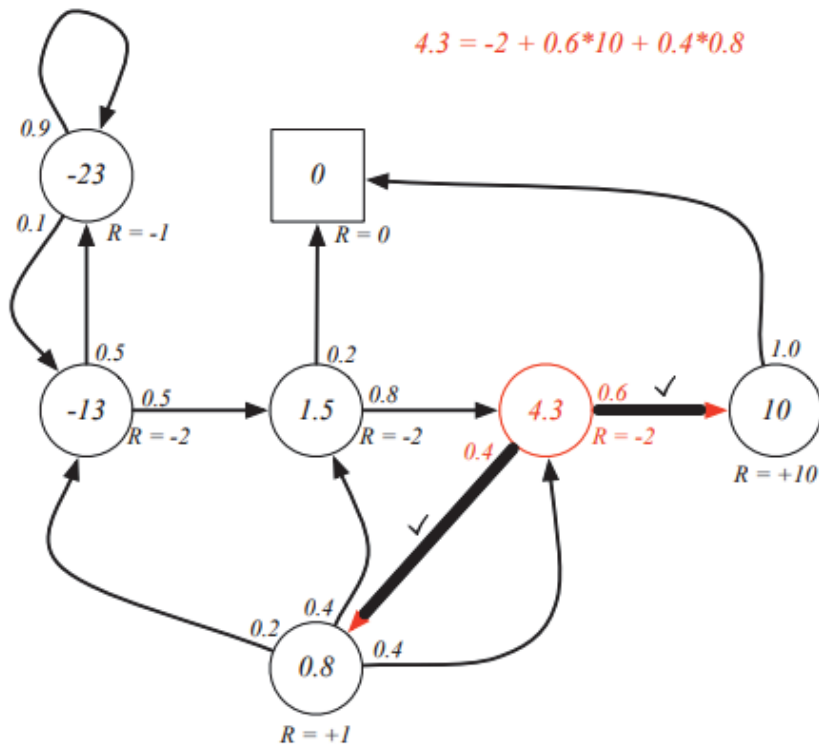
Նկ. 9.6: $v(s)$ հաշվումը ($\gamma = 1$)

Նկ. 9.6-ում տրված են v ֆունկցիայի արժեքները բոլոր գազաթների համար $\gamma = 1$ գեղի գործակցի դեպքում: Իսկ նկ. 9.7-ում տրվում է ֆունկցիայի արժեքների հաշվման սխեման:

Այժմ տեսնենք, թե ինչ հավասարման է բավարարում $v(s)$ ֆունկցիան: Ունենք

$$\begin{aligned}
 v(s) &= E[G_t | s_t = s] = \\
 &= E[R_{t+1} + \gamma T_{t+2} + \gamma^2 R_{t+3} + \dots | s_t = s] = \\
 &= E[R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + \dots | s_t = s)] \\
 &= E[R_{t+1} + \gamma G_{t+1} | s_t = s] \\
 &= E[R_{t+1} + \gamma v(s_{t+1} | s_t = s)] = R_s + \gamma \sum_{s' \in S} p_{ss'} v(s') :
 \end{aligned}$$

Այս հավասարումը մատրիցային տեսքով ունի հետևյալ տեսքը՝ $v = R + \gamma P v$:



Նկ. 9.7: $v(s)$ ֆունկցիայի արժեքների հաշվման սխեման

Այստեղից կստանանք

$$V = (I - \gamma P)^{-1} R:$$

Սակայն այս բանաձևով ֆունկցիայի արժեքները հաշվելը հարմար չէ, որովհետև եթե համակարգի վիճակների քանակը մեծ է, ապա հակադարձ մատրիցի հաշվողական բարդությունը՝ $O(n^3)$, շատ մեծ կլինի: Դրա համար նշված հավասարումը կարելի է լուծել հետևյալ ռեկուրենտ բանաձևով՝

$$v_{k+1} = R + \gamma P v_k:$$

Բայց որպեսզի ցույց տանք, որ այս հաջորդականությունը զուգամիտում է հավասարման լուծմանը, պետք ապացուցել, որ

$$F(v) = R + \gamma P v$$

օպերատորը սեղմող է: Իրոք, ունենք

$$\|F(v) - F(w)\| = \gamma \|P(v) - P(w)\| \leq \|v - w\|:$$

Համաձայն սեղմող արտապատկերումների սկզբունքի՝ v_k հաջորդականությունը երկրաչափական պրոգրեսիայի արագությամբ զուգամիտում է v ֆունկցիային:

9.2 Ամրապնդմամբ ուսուցում

Մարկովյան որոշումների պրոցեսը (Markov decision proces) կամ կարճ՝ MDP մաթեմատիկական մոդել է, որն օգտագործվում է ML-ում որոշումների կայացման խնդիրները մոդելավորելու համար: Այդ պրոցեսները նկարագրվում են հետևյալ հինգ պարամետրերի օգնությամբ (S, A, P, R, γ): Այստեղ՝

- S -ը վիճակների բազմությունն է (տարածությունն է, state space), որում ժամանակի յուրաքանչյուր պահի կարող է գտնվել «միջավայրը»:
- A -ն գործողությունների վերջավոր բազմությունն է (action space):
- P -ն անցման ֆունկցիան է (transition function), որով միջավայրը մի վիճակից անցնում է մի ուրիշ վիճակի $a \in A$ գործողության ընտրության արդյունքում: Այդ անցման ֆունկցիան տրվում է $P(s'|s, a)$ պայմանական հավանականության միջոցով: Սա ցույց է տալիս, թե ինչ հավանականությամբ միջավայրը կանցնի s' վիճակի, եթե այն գտնվում է s վիճակում, և կատարվել է $a \in A$ գործողություն: Այսինքն՝

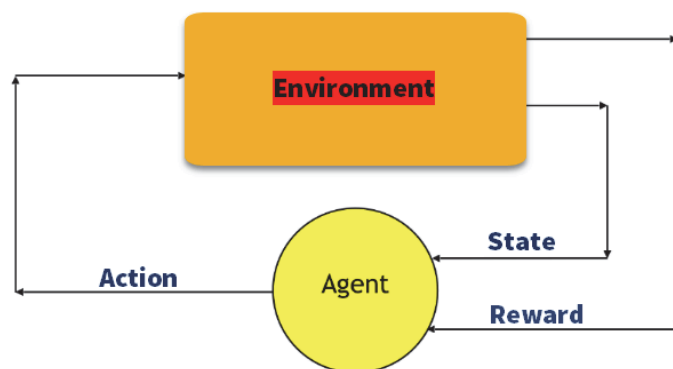
$$P_{ss'}^a = P[s_{t+1} = s' | s_t = s, A_t = a]:$$

- $r : S \times A \rightarrow R$ -ն պարգևատրման ֆունկցիան է (reward function), որի արժեքը կոչվում է ուսուցանող ազդանշան:

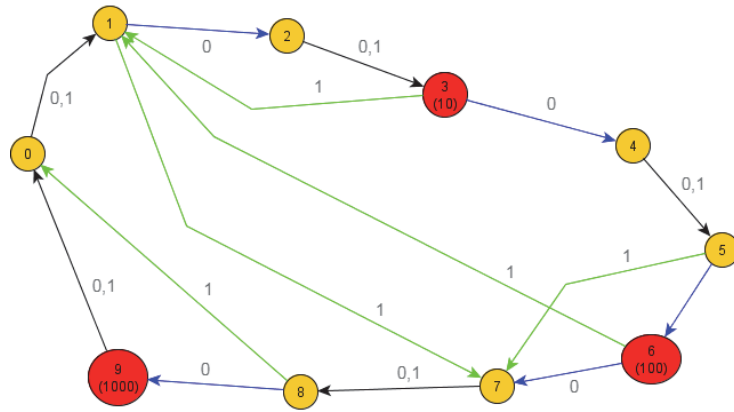
$$R_s^a = E[R_{t+1} | s_t = s, A_t = a]:$$

- $\gamma \in [0, 1]$ գեղչի գործակիցն է:

MDP-ի աշխատանքի սկզբունքը տրված է նկար 9.8-ում: Բերենք պարզ օրինակ, որը սխեմատիկ ներկայացված է նկար 9.9-ում:



Նկ. 9.8: MDP



Նկ. 9.9: Վիճակների անցման գրաֆ

Դիցուք

- $A = \{0, 1\}$,
- $S = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$,
- անցման մատրիցը տրված է գրաֆի տեսքով (տե՛ս նկ. 9.9),
- R_s^a -ն ներկայացված է գրաֆում (վիճակներում փակագծերի մեջ տրված են R -ի արժեքները: Եթե այն նշված չէ, նշանակում է՝ այդտեղ պարգևատրում չկա):

Խնդիրն է գտնել **մաքսիմալ պարգևատրմամբ** շղթան: 9.9 գրաֆում սլաքների վրա գրված թվերը ցույց են տալիս, թե որ գործողության ընթացքում է միջավայրը մի վիճակից տեղափոխվում ուրիշ վիճակի: Ընդունված է ասել, որ այն օբյեկտը, որը ազդում է միջավայրի վրա, կոչվում է գործակալ (agent): Գործակալը գործում է որոշակի կանոններով, որոնք կոչվում են ստրատեգիաներ: Ստրատեգիան սահմանվում է հետևյալ բանաձևով՝

$$\pi(a|s) = P[A_t = a | s_t = s]:$$

Այսինքն՝ սա հավանականությունն է այն բանի, որ գործակալը կընտրի $a \in A$ գործողությունը, եթե միջավայրը s վիճակում է, և այդ հավանականությունը ժամանակից կախված չէ: Այժմ դիտարկենք MDP պրոցեսը ֆիքսած π ստրատեգիայի դեպքում՝

$$(S, P^\pi, R^\pi, \gamma),$$

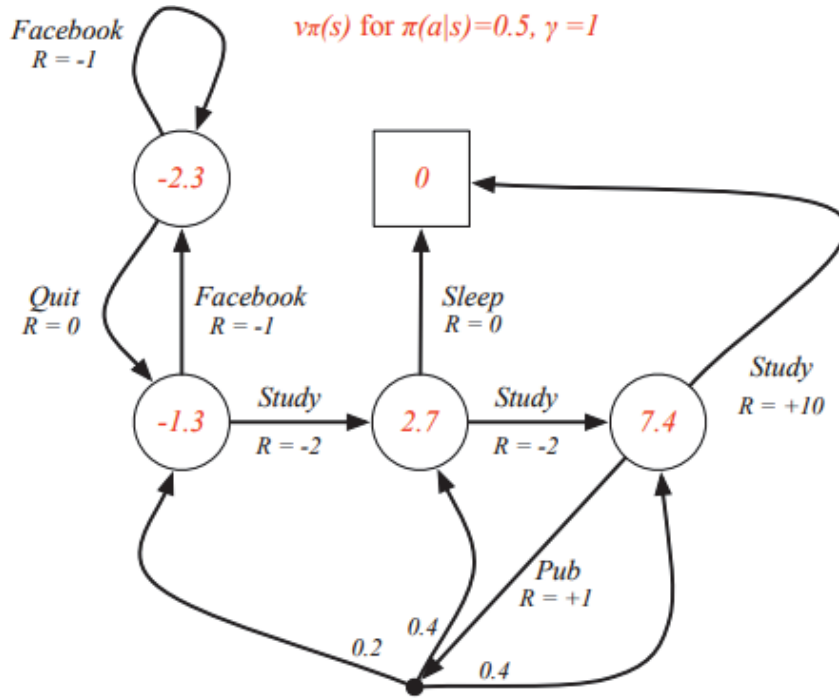
որտեղ

$$P_{ss'}^\pi = \sum_{a \in A} \pi(a|s) P_{ss'}^a,$$

$$R_s^\pi = \sum_{a \in A} \pi(a|s) R_s^a:$$

Սահմանենք նաև վիճակի արժեքի ֆունկցիան (state-value function) v_π ստրատեգիայի դեպքում՝

$$v_\pi(s) = E_\pi[G_t | s_t = s]:$$



Նկ. 9.10: state-value function

Նկար 9.10-ում կատարված է v_π ֆունկցիայի արժեքների հաշվարկ դիտարկվող օրինակի համար:

Անալոգիայով սահմանենք, այսպես ասած, *action – value* ֆունկցիան՝

$$Q_\pi(s, a) = E_\pi[G_t | s_t = s, A_t = a]:$$

Այժմ ներկայացնենք այն հավասարումները, որոնց այս ֆունկցիաները բավարարում են:

Սահմանումներից ակնհայտ բխում է, որ

$$v_\pi(a) = \sum_{a \in A} \pi(a|s) Q_\pi(s, a),$$

$$Q_\pi(s, a) = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_\pi(s'):$$

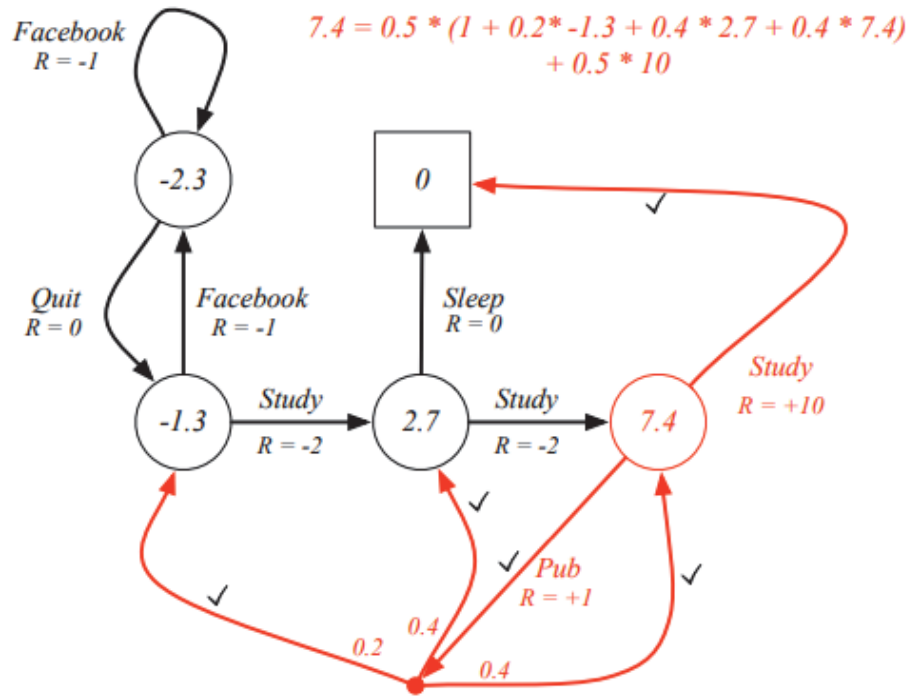
Այսպիսով, այս երկուսից կստանանք

$$v_\pi(s) = \sum_{a \in A} (R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_\pi(s')),$$

$$Q_\pi(s, a) = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a \sum_{a' \in A} \pi(a'|s') Q_\pi(s', a'):$$

Կարելի է ցույց տալ, որ v_π ֆունկցիան արտահայտվում է հետևյալ բանաձևով՝

$$v_\pi = (I - \gamma P^\pi)^{-1} \mathbf{R}^\pi:$$



Նկ. 9.11: $v_\pi(s)$ ֆունկցիայի արժեքների հաշվման սխեման

Նկ. 9.11-ում տրված է $v_\pi(s)$ ֆունկցիայի արժեքների հաշվման սխեման:
 Այժմ սահմանենք օպտիմալ $v_*(s)$ և $Q_*(s, a)$ ֆունկցիաները՝

$$v_*(s) = \max_{\pi} v_\pi(s),$$

$$Q_*(s, a) = \max_{\pi} Q_\pi(s, a):$$

Նկարներ 9.12- 9.13 ներկայացված են այդ ֆունկցիաների արժեքները նախորդ դիտարկվող օրինակի համար:

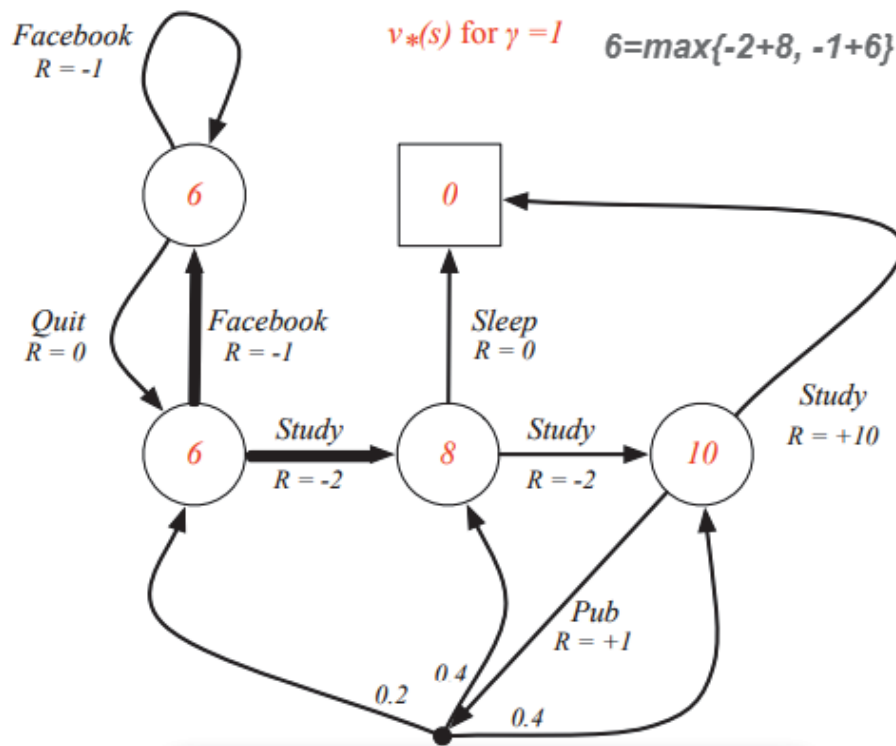
Այժմ սահմանենք օպտիմալ ստրատեգիայի գաղափարը: Կասենք, որ $\pi \geq \pi'$, եթե

$$v_\pi(s) \geq v_{\pi'}(s) \quad \forall s \in S:$$

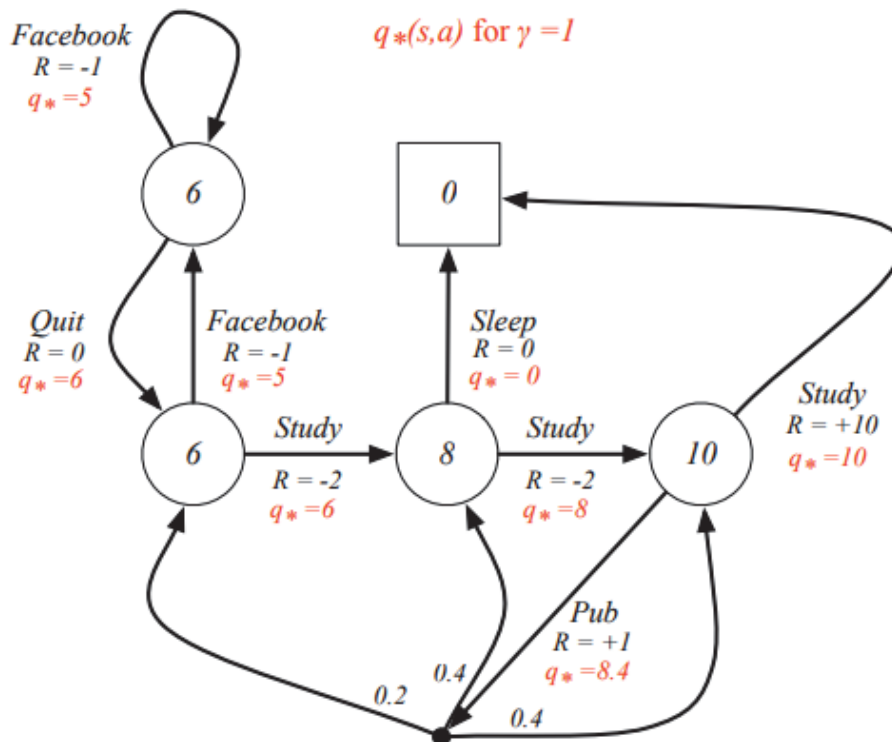
Տեղի ունի օպտիմալ ստրատեգիայի գոյության վերաբերյալ թեորեմը, որը կրում է Մարկովի անունը:

Թեորեմ 9.1. Կամայական մարկովյան MDP պրոցեսի համար գոյություն ունի այնպիսի π_* ստրատեգիա, որ

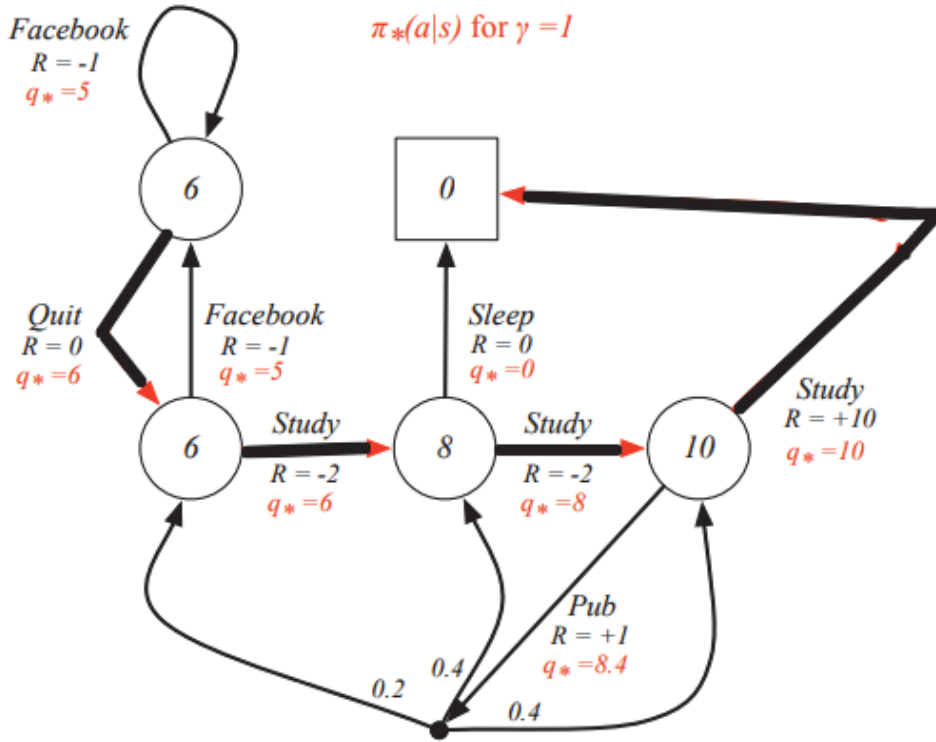
- $\pi_* \geq \pi \quad \forall \pi$,
- $v_{\pi_*}(s) = v_*(s)$,
- $Q_{\pi_*}(s, a) = Q_*(s, a) \quad \therefore$,
- $\pi_*(a, s) = 1$, եթե $a = \arg \max_{b \in A} Q_*(s, b)$, հակառակ դեպքում՝ 0:



Նկ. 9.12: $v_*(s)$ օպտիմալ ֆունկցիայի արժեքները



Նկ. 9.13: $Q_*(s,a)$ օպտիմալ ֆունկցիայի արժեքները



Նկ. 9.14: $\pi_*(s)$ օպտիմալ ստրատեգիա

Կարելի է ցույց տալ, որ այդ օպտիմալ ֆունկցիաները բավարարում են Բելմանի հավասարումները՝

$$v_*(s) = \max_a (R_S^a + \gamma \sum_{s' \in S} P_{ss'}^a v_*(s')),$$

$$Q_*(s, a) = R_S^a + \sum_{s' \in S} P_{ss'}^a \max_{a' \in A} Q_*(s', a):$$

Նկար 9.14-ում տրված է դիտարկվող օրինակին վերաբերվող օպտիմալ π_* ստրատեգիան:

Այսպիսով, ամփոփենք վերը շարադրվածը համառոտ: Ենթադրենք՝ ունենք մի ֆիքսված π ստրատեգիա, որի գործողության ընթացքում ձևավորվել է հետևյալ հաջորդականությունը՝ $\tau = (s_0, a_0, s_1, a_1, \dots)$: Ժամանակի t պահին գործակալը կստանա $r_t = r(s_t, a_t)$ պարգևատրում: Ընդհանուր պարգևատրումը (գեղչով) հավասար կլինի $\sum_{t \geq 0} \gamma^t r_t$, որը պատահական մեծություն է: Խնդիրը այն է, որ պետք է գտնել այնպիսի ստրատեգիա, որ այդ գումարի միջին արժեքը լինի մաքսիմալ: Այսինքն՝

$$E_{\tau \sim \pi} \sum_{t \geq 0} \gamma^t r_t \rightarrow \max_{\pi}:$$

Այս խնդիրը լուծվում է Բելմանի դինամիկ ծրագրավորման սկզբունքով: Նախ սահմանվում է վիճակի գնահատման օպտիմալ ֆունկցիա՝ $Q_*(s, a) = \max_{\pi} E_{[\tau \sim \pi | s_0 = s, a_0 = a]} \sum_{t \geq 0} \gamma^t r_t$, որտեղ $[\tau \sim \pi | s_0 = s, a_0 = a]$ -ն նշանակում է, որ ընտրվել է սկզբում $s_0 = s$ վիճակը և $a_0 = a$ գործողությունը, և հետագայում միջավայրի հետ փոխազդելու արդյունքում գեներացվել է τ հետագիծը (տրայեկտորիան): Այժմ սահմանենք $\pi_*(s)$ ստրատեգիա հետևյալ սկզբունքով՝ $\pi_*(s) = \arg \max_{a \in A} Q_*(s, a)$:

Այսպես կառուցված ստրատեգիան կոչվում է Բելմանի օպտիմալության սկզբունքով կառուցված ստրատեգիա: Այժմ հարց է առաջանում՝ ինչպես կառուցել այդ Q_* ֆունկցիան: Այդ ֆունկցիան բավարարում է Բելմանի հավասարումը՝

$$Q(s, a) = r(s, a) + \gamma E_{s' \sim p(s'|s, a)} \max_{a' \in A} Q(s', a'):$$

Այս ոչ գծային հավասարումը լուծում են պարզ իտերացիայի եղանակով, ինչպես $x = f(x)$ տիպի հավասարումները: Ընտրվում նախնական մի $Q_0(s, a) : S \times A \rightarrow R$ ֆունկցիա և ռեկուրենտ կառուցվում են $Q_k(s, a)$ ֆունկցիաները հետևյալ բանաձևով՝

$$Q_{k+1}(s, a) = r(s, a) + \gamma E_{s' \sim p(s'|s, a)} \max_{a' \in A} Q_k(s', a'):$$

Այս հաջորդականությունը զուգամիտում է $Q_*(s, a)$ ֆունկցիային: Այս պնդումը ապացուցելու համար դիտարկենք հետևյալ օպերատորը՝

$$(HQ)(s, a) = R_s^a + \sum_{s' \in S} P_{ss'}^a \max_{a' \in A} Q(s', a):$$

Սա գործում է $Q : S \times A \rightarrow R$ ֆունկցիաների բազմության վրա: Յույց տանք, որ այն սեղմող օպերատոր է: Իրոք, ունենք

$$\begin{aligned} & \|H(Q_1) - H(Q_2)\| = \\ &= \max_{(s, a)} \left| \gamma \sum_{s' \in S} P_{s, s'}^a \max_{a' \in A} Q_1(s', a) - \gamma \sum_{s' \in S} P_{s, s'}^a \max_{a' \in A} Q_2(s', a) \right| = \\ &= \gamma \max_{(s, a)} \sum_{s' \in S} P_{s, s'}^a \left| \max_{a' \in A} Q_1(s, a') - \max_{a' \in A} Q_2(s, a') \right| \leq \\ &\leq \gamma \max_{s, a} |Q_1(s, a) - Q_2(s, a)| = \\ &= \gamma \|Q_1 - Q_2\|: \end{aligned}$$

Այսպիսով, սեղմող արտապատկերումների սկզբունքի համաձայն՝ վերևի ռեկուրենտ առնչությամբ կառուցված հաջորդականությունը զուգամիտում է Q_* ֆունկցիային:

Սակայն այս իտերացիոն պարզ ալգորիթմի փոխարեն սովորաբար կիրառվում է հետևյալ ալգորիթմը, որը գրականության մեջ կոչվում է Q -learning ալգորիթմ, որում յուրաքանչյուր (s, a) մուտքային տվյալների համար կատարում ենք Q օգտակարության ֆունկցիայի *random* ինիցիալիզացիա (տե՛ս նկ. 9.15):

	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>
<i>A</i>	0	0	0	0	0	0
<i>B</i>	0	0	0	0	0	0
<i>C</i>	0	0	0	0	0	0
<i>D</i>	0	0	0	0	0	0
<i>E</i>	0	0	0	0	0	0
<i>F</i>	0	0	0	0	0	0

Նկ. 9.15: Սկզբնական Q_0 մատրիցը

Այնուհետև ալգորիթն աշխատում է հետևյալ ընթացակարգով.

Q-learning($S, A, R, T, \beta, \gamma$).

```

for  $s \in S$  :           for  $a \in A$  :  $Q(s, a) = rand()$ 
while Q is not converged:       $s = \forall s \in S$ 
  • while  $s$  is not terminal:
     $\pi(s) = \arg \max_a Q(s, a)$ 
     $a = \pi(s)$ 
     $r = R(s, a)$ 
     $s' = T(s, a)$ 
     $Q(s', a) = (1 - \beta)Q(s', a) + \beta(r + \gamma \max_{a'} Q(s', a'))$ 
   $s = s'$            return Q

```

Տեղի ունի գուգամիտության հետևյալ թեորեմը (տե՛ս [31]):

Թեորեմ 9.2. *Դիցուք MDP-ում վիճակների և գործողությունների բազմությունը վերջավոր է: Այդ դեպքում ռեկուրենտ*

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha_t(s_t, a_t)[r_t + \gamma \max_{b \in A} Q_t(s_{t+1}, b) - Q_t(s_t, a_t)]$$

կանոնով կառուցված հաջորդականությունը 1-հավանականությամբ գուգամիտում է Q_ ֆունկցիային: Այստեղ*

$$\gamma \in (0, 1), \alpha_t(s, a) \geq 0, \sum_t \alpha_t(s, a) = +\infty, \sum_t \alpha_t^2(s, a) < +\infty, \forall (s, a) \in S \times A:$$

Այժմ դիտարկենք թվային մի օրինակ, որում Q -ն որոշվում է

$$Q(state, action) = R(state, action) + \gamma \max[Q(state, allaction)]$$

բանաձևով: Ենթադրենք՝ $\gamma = 0.8$, և սկզբնական վիճակը B -ն է, իսկ Q մատրիցի սկզբնական վիճակը տրված է նկ 9.15-ի մատրիցով, որի բոլոր էլեմենտները զրոյական են: Իսկ նկար 9.16-ում տրված է R պարզևատրման մատրիցը:

$$\mathbf{R} = \begin{array}{c|cccccc} state \backslash action & A & B & C & D & E & F \\ \hline A & - & - & - & - & 0 & - \\ B & - & - & - & 0 & - & 100 \\ C & - & - & - & 0 & - & - \\ D & - & 0 & 0 & - & 0 & - \\ E & 0 & - & - & 0 & - & 100 \\ F & - & 0 & - & - & 0 & 100 \end{array}$$

Նկ. 9.16: Պարզեցված R մատրից

$$\mathbf{Q} = \begin{array}{c|cccccc} & A & B & C & D & E & F \\ \hline A & 0 & 0 & 0 & 0 & 0 & 0 \\ B & 0 & 0 & 0 & 0 & 0 & 100 \\ C & 0 & 0 & 0 & 0 & 0 & 0 \\ D & 0 & 0 & 0 & 0 & 0 & 0 \\ E & 0 & 0 & 0 & 0 & 0 & 0 \\ F & 0 & 0 & 0 & 0 & 0 & 0 \end{array}$$

Նկ. 9.17: Q_1 մատրից

Նայելով այս մատրիցի երկրորդ տողին՝ նկատում ենք, որ B վիճակի համար հնարավոր է երկու գործողություն՝ «գնալ D վիճակ կամ F վիճակ»: Ենթադրենք՝ ընտրում ենք երկրորդ գործողությունը: Այժմ գտնվում ենք F վիճակում և նայելով R մատրիցի վեցերորդ տողին տեսնում ենք, որ այս վիճակի համար կա երեք հնարավոր գործողություն՝ գնալ B, E կամ F վիճակների: Հետևաբար՝ համաձայն

$$Q(state, action) = R(state, action) + \gamma \max[Q(state, allaction)]$$

բանաձևի՝ կստանանք՝

$$Q(B, F) = R(B, F) + 0.8 \max\{Q(F, B), Q(F, E), Q(F, F)\} = 100 + 0.8 \times = 100:$$

Հետևաբար նոր Q մատրիցը կունենա հետևյալ տեսքը՝ (տե՛ս նկ. 9.17):

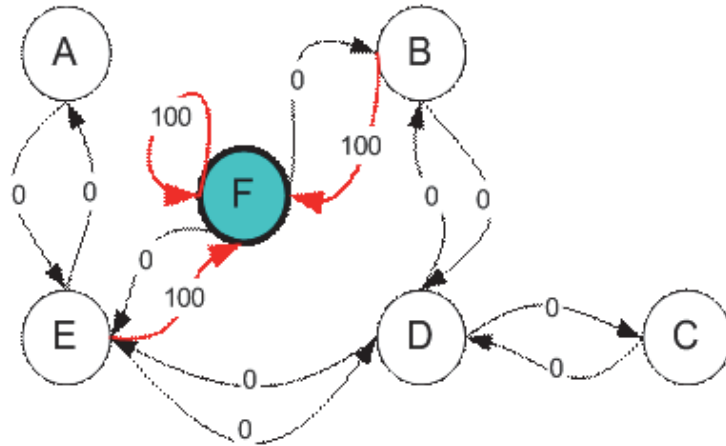
Քանի որ F -ը վերջնական վիճակն էր, ապա մենք վերջացրինք մեկ քայլ (Էպիզոդ): Հաջորդ Էպիզոդում մենք սկսում ենք պատահական ուրիշ վիճակից: Ենթադրենք՝ այժմ որպես սկզբնական (ինիցիալ) վիճակ ընտրված է D -ն: Նոր մատրիցից երևում է, որ այդ վիճակից կարելի է տեղափոխվել B, C կամ E վիճակների: Ենթադրենք գնում ենք B : Այդ դեպքում

$$Q(D, B) = R(D, B) + 0.8 \max\{Q(B, D), Q(B, F)\} = 0 + 0.8 \max\{0, 100\} = 80:$$

Այստեղից՝ Q մատրիցը կունենա նկար 9.18-ում ներկայացված տեսքը:

	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>
<i>A</i>	0	0	0	0	0	0
<i>B</i>	0	0	0	0	0	100
<i>Q = C</i>	0	0	0	0	0	0
<i>D</i>	0	80	0	0	0	0
<i>E</i>	0	0	0	0	0	0
<i>F</i>	0	0	0	0	0	0

Նկ. 9.18: Q_2 մատրից



Նկ. 9.19: Վիճակների և գործողությունների գրաֆ

Այս բոլորը գրաֆի (դիագրամայի) տեսքով պատկերված է նկ. 9.19-ում:

Այսպես շարունակ՝ որոշ էպիզոդներից հետո, Q -ն վերջնականում կունենանք $Q_{finally}$ մատրիցի կառուցվածքը (տես նկ. 9.20): Նորմավորենք այս մատրիցը՝ բոլոր

<i>state \ action</i>	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>
<i>A</i>	-	-	-	-	400	-
<i>B</i>	-	-	-	320	-	500
<i>Q = C</i>	-	-	-	320	-	-
<i>D</i>	-	400	256	-	400	-
<i>E</i>	320	-	-	320	-	500
<i>F</i>	-	400	-	-	400	500

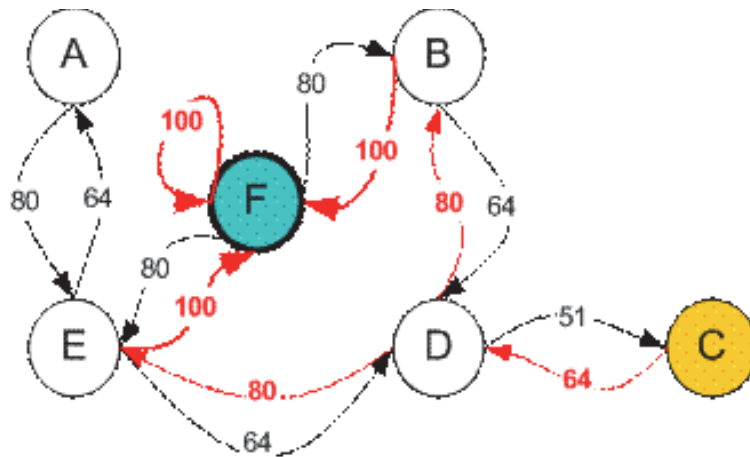
Նկ. 9.20: $Q_{finally}$ մատրիցը

Էլեմենտները բաժանելով հինգի վրա՝ կստանանք Q_{norm} մատրիցը (նկ. 9.21):

Իսկ այդ մատրիցին համապատասխանող գրաֆը պատկերված է նկար 9.22-ում: Այս գրաֆի միջոցով կարող ենք գտնել կամայական սկզբնական վիճակից վերջնական վիճակ գնալու օպտիմալ ճանապարհը: Օրինակ՝ C սկզբնական վիճակից վերջնական F վիճակ գնալու օպտիմալ ճանապարհը կլինի $C \rightarrow D \rightarrow B \rightarrow F$ շղթան:

<i>state \ action</i>	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>	<i>F</i>
<i>A</i>	–	–	–	–	80	–
<i>B</i>	–	–	–	64	–	100
<i>C</i>	–	–	–	64	–	–
<i>D</i>	–	80	51	–	80	–
<i>E</i>	64	–	–	64	–	100
<i>F</i>	–	80	–	–	80	100

Նկ. 9.21: Q_{norm} մատրիցը



Նկ. 9.22: Q_{norm} մատրիցի գրաֆը

10

Օնլայն օպտիմիզացիա: Օնլայն ուսուցման մոդելներ

Դիտարկենք օնլայն ուսուցման հետևյալ խաղը. $t = 1, 2, \dots, T$ պահերին առաջին խաղացողը վերցնում է w_t վեկտորը K բազմությունից: Դրանից հետո նրա հակառակորդը նրան տալիս է $f_t : H \rightarrow R$ կորստի ֆունկցիան F դասից: Առաջին խաղացողի նպատակն է կառուցել այնպիսի w_t , $t = 1, 2, \dots, T$ ստրատեգիաներ, որ

$$\frac{1}{T} \sup_{w \in H} \text{regret}_T(w) \rightarrow 0, \text{ երբ } T \rightarrow \infty,$$

որտեղ

$$\text{regret}_T(w) \equiv \sum_{t=1}^T f_t(w_t) - \sum_{t=1}^T f_t(w):$$

w_t ստրատեգիան կառուցելու ժամանակ առաջին խաղացողը կարող է օգտագործել միայն նախորդ քայլերի ինֆորմացիան: Այսինքն՝ w_t -ի ընտրության ժամանակ խաղացողին հայտնի չեն $f_t, f_{t+1}, \dots, f_T$ ֆունկցիաները:

10.1 Օնլայն (առցանց) օպտիմիզացիա

Դիցուք ուռուցիկ են հիպոթեզների H բազմությունը և կորստի $l : H \times Z \rightarrow R$ ֆունկցիան (ցանկացած ֆիքսած $z \in Z$ պարամետրերի դեպքում): Դիցուք

$$w_t \in H, z_t \in Z, t = 1, 2, \dots, T,$$

$$\text{regret}_T(w) = \sum_{t=1}^T l(w_t, z_t) - \min_{w \in H} \sum_{t=1}^T l(w, z_t):$$

Դիտարկենք հետևյալ ալգորիթմը, որը կոչվում է օնլայն գրադիենտային իջեցում:

- Ֆիքսում ենք կամայական $\eta > 0$ պարամետր:
- $w_1 = 0$ և $t = 1, 2, \dots, T$ պահերին ալգորիթմը ստանում է z_t վեկտորը և հաշվում իր կորուստը՝ $f_t(\cdot) = l(\cdot, z_t)$:

- Կատարվում է w պարանետրի արժեքների «թարմացում»՝

$$w_{t+1/2} = w_t - \eta v_t, \quad v_t \in \partial f_t(w_t), \quad w_{t+1} = \arg \min_{w \in H} \|w - w_{t+1/2}\|:$$

Թեորեմ 10.1. *Տեղի ունի հետևյալ անհավասարությունը՝*

$$\text{regret}_T(w) = \sum_{t=1}^T l(w_t, z_t) - \sum_{t=1}^T l(w, z_t) \leq \frac{\|w\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^T \|v_t\|^2:$$

Ապացույց: Ապացույցը անալոգ է ստոխաստիկ գրադիենտային իջեցմանը:
Ունենք

$$\begin{aligned} & \|w_{t+1} - w\|^2 - \|w_t - w\|^2 = \|w_{t+1} - w\|^2 - \\ & - \|w_{t+1/2} - w\|^2 + \|w_{t+1/2} - w\|^2 - \|w_t - w\|^2 \leq \\ & \leq \|w_{t+1/2} - w\|^2 - \|w_t - w\|^2 = \\ & = \|w_t - \eta v_t - w\|^2 - \|w_t - w\|^2 = \\ & - 2\eta(w_t - w, v_t) + \eta^2 \|v_t\|^2 = \\ & = -2\eta(f_t(w_t) - f_t(w)) + \eta^2 \|v_t\|^2: \end{aligned}$$

Գումարելով այս անհավասարությունները, հաշվի առնելով, որ $w_1 = 0$, կստանանք

$$\begin{aligned} \sum_{t=1}^T (f_t(w_t) - f_t(w)) & \leq \frac{\|w_1 - w\|^2 - \|w_{T+1} - w\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^T \|v_t\|^2 \leq \\ & \leq \frac{\|w\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^T \|v_t\|^2: \end{aligned}$$

Այժմ եթե ենթադրենք, որ f_t ֆունկցիաները ρ -լիպշիցյան են և $\|w\| \leq B$, և վերցնելով $\eta = B/(\rho\sqrt{T})$, կստանանք

$$\sup_{w \in H} \text{regret}_T(w) = \sum_{t=1}^T f_t(w_t) - \inf_w \sum_{t=1}^T f_t(w) \leq B\rho\sqrt{T}:$$

Այստեղից

$$\frac{1}{T} \sup_{w \in H} \text{regret}_T(w) \leq \frac{B\rho\sqrt{T}}{T} \rightarrow 0:$$

10.2 Perceptron օնլայն ալգորիթմը

Perceptron ալգորիթմը բինար դասակարգման դասական ալգորիթմներից է, երբ որպես հիպոթեզների բազմություն ընտրվում է համասեռ կիսատարածությունները և ենթադրվում է, որ հավաքածուն խիստ բաժանելի է հիպերհարթությամբ:

Դիցուք $X = \mathbb{R}^d$, $Y = \{-1, 1\}$: Որպես հիպոթեզների բազմություն վերցնում ենք հետևյալ բազմությունը՝

$$H = \{x \rightarrow \text{sign}(w, x) : w \in \mathbb{R}^d\}:$$

Տրված է որոշակի հավաքածու՝ $S = (x_1, y_1), \dots, (x_i, y_i)$: t -րդ քայլում ալգորիթմը ստանում է x_t վեկտոր: Կատարվում է x_t օբյեկտի նշիչի (մետկայի) կանխատեսում հետևյալ՝ $p_t = \text{sign}(w_t, x_t)$ բանաձևով, որտեղ w_t վեկտորն ընտրվում է ալգորիթմի կողմից: Դրանից հետո ալգորիթմը ստանում է y_t ռեալ նշիչը և հաշվում է կանխատեսման սխալանքը: Եթե $y_t \neq p_t$, ապա սխալանքը հավասար է մեկի, հակառակ դեպքում՝ զրոյի: Այսինքն՝ կորուստը կարտահայտվի հետևյալ ֆունկցիայով՝

$$l(w_t, (x_t, y_t)) = \mathbb{1}_{[y_t(w_t, x_t) \leq 0]}:$$

Այսպիսով, $l(w, z)$ ֆունկցիան ալգորիթմի կորստի ֆունկցիան է: Պետք է գնահատել ընդհանուր սխալանքը T ժամանակահատվածում՝

$$\sum_{i=1}^T l(w_t, z_t):$$

Եթե t -րդ փուլում ալգորիթմը սխալվում է, ապա կորստի այս ֆունկցիան վերևից մոտարկում ենք մեզ հայտնի հենքային $f_t(w) = \max\{0, 1 - y_t(w, x_t)\}$ ֆունկցիայով: Իսկ եթե այս փուլում ալգորիթմը չի սխալվում, ապա սահմանում ենք՝ $f_t(w) = 0$: Հետևաբար՝

- f_t -ն ուռուցիկ ֆունկցիա է:
- $f_t(w) \geq l(w_t, (x_t, y_t)) \quad \forall w$: Մասնավորապես այս անհավասարությունը տեղի ունի w_t վեկտորների համար: Այն փուլերում, երբ ալգորիթմը ճիշտ է կանխատեսում $f_t(w) = 0$: Ակնհայտ է, որ այդ դեպքում f_t -ն ուռուցիկ ֆունկցիա է: Բացի դրանից՝ $f_t(w_t) = l(w_t, (x_t, y_t)) = 0$:

Կիրառելով օնլայն գրադիենտային իջեցման վերը նկարագրած ալգորիթմը, w_t ստրատեգիաների կառուցման համար կունենանք հետևյալ ընթացակարգը՝

- $w_1 = 0$:
- $w_{t+1} = w_t - \eta v_t$, որտեղ $v_t \in \partial f_t(w_t)$:

Մեր դեպքում՝ եթե $y_t(w_t, x_t) > 1$, ապա $f_t = 0$, և հետևաբար՝ $v_t = 0$: Հակառակ դեպքում հեշտ է նկատել, որ $v_t = -y_t x_t \in \partial f_t(w_t)$: Հետևաբար կստանանք հետևյալ ռեկուրենտ առնչությունը՝

$$w_{t+1} = \begin{cases} w_t, & \text{եթե } y_t(w_t, x_t) > 1, \\ w_t + \eta y_t x_t, & \text{հակառակ դեպքում:} \end{cases}$$

M -ով նշանակենք այն փուլերի բազմությունը, որոնցում $\text{sign}(w_t, x_t) \neq y_t$: Քանի որ ալգորիթմի կանխատեսումը t -րդ փուլում արտահայտվում է

$$p_t = \text{sign}(w_t, x_t) = \text{sign}\left(\eta \sum_{i \in M: i < t} y_t(x_i, x_t)\right)$$

քանաձևով, ապա կանխատեսման արդյունքը, ինչպես նաև M բազմությունը կախված չեն $\eta > 0$ թվից: Այնպես որ, Պերսեպտրոնի ալգորիթմը ունի հետևյալ տեսքը՝

- $w_1 = 0$,
- $p_t = \text{sign}(w_t, x_t)$ (ալգորիթմի կանխատեսումը t պահին),
- եթե $y_t(w_t, x_t) \leq 1$, ապա

$$w_{t+1} = w_t + y_t x_t,$$

հակառակ դեպքում $w_{t+1} = w_t$:

Սա արդեն Ռոզենբլատի քննարկած նմանակն է: Կիրառելով թեորեմ 10.1-ը՝ կստանանք՝

$$\begin{aligned} \sum_{t=1}^T f_t(w_t) - \sum_{t=1}^T f_t(w) &\leq \\ &\leq \frac{\|w\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^T \|v_t\|^2: \end{aligned}$$

Եթե ենթադրենք, որ $\|x_t\| \leq L$, ապա

$$\|v_t\| \leq \|x_t\| \leq L_T:$$

$|M|$ -ով նշանակենք ալգորիթմի սխալվելու քանակը T քայլերից (փուլերից) հետո: Ակնհայտ է, որ

$$|M| \leq \sum_{t=1}^T f_t(w_t);$$

Հետևաբար՝

$$|M| - \sum_{t=1}^T f_t(w) \leq \frac{\|w\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^T M_T L_T^2:$$

Տեղադրելով այս անհավասարության մեջ

$$\eta = \frac{\|w\|}{L\sqrt{|M|}},$$

կստանանք՝

$$|M| - L\|w\|\sqrt{|M|} - \sum_{t=1}^T f_t(w) \leq 0:$$

Լուծելով այս քառակուսի անհավասարությունը M_T -ի նկատմամբ՝ կստանանք

$$|M| \leq \sum_{t=1}^T f_t(w) + L_T \|w\| \sqrt{\sum_{t=1}^T f_t(w) + L_T^2 \|w\|^2}:$$

Մասանավորապես, եթե գոյություն ունի այնպիսի w վեկտոր, որ $y_t(w, x_t) \geq 1$, $t = 1, 2, \dots, T$, ապա

$$|M| \leq L^2 \|w\|^2:$$

Հենման վեկտորների մեթոդից արդեն գիտենք, որ $y_t(w, x_t) \geq 1$, $t = 1, 2, \dots, T$ պայմանը նշանակում է հիպերհարթությամբ անջատման պայման, իսկ $\rho = \frac{2}{\|w\|}$ թիվը արտահայտում է մաքսիմալ անջատող շերտի հաստության մեծությունը: Այսպիսով վերջնականում կստանանք

$$|M| \leq \frac{L_T^2}{2\rho^2}:$$

Հիշենք, որ այս գնահատականը ստացվել էր Ռոզենբլատի ալգորիթմի քննարկման ժամանակ: Այս անհավասարությունից հետևում է, որ որքան մեծ է անջատող շերտի հաստությունը, այնքան փոքր է ալգորիթմի սխալների քանակը:

10.3 Բայես-Լապլասի կանխատեսման մոդելը

Դիցուք ունենք զրոներից և մեկերից բաղկացած հետևյալ հաջորդականությունը

$$S^n = (x_1, x_2, \dots, x_n):$$

Սա կարելի է մեկնաբանել որպես n անկախ փորձերի արդյունք: Ընդ որում ենթադրվում է, որ մեկ փորձի ընթացքում մեկ թվի հանդես գալու հավանականությունը հավասար է θ -ի, իսկ զրոյինը՝ $1 - \theta$: Խնդիր է դրվում ինչպես կանխատեսել $n + 1$ փորձի ելքը՝ իմանալով նախորդ փորձերի արդյունքները: Կան կանխատեսման տարբեր մոդելներ: Քննարկենք մի եղանակ, որը կրում է Բայես-Լապլասի անունը: X պատահական մեծությունն ընդունում է երկու արժեք՝ $X \in \{0, 1\}$: Այժմ հաշվենք, որ θ պարամետրը պատահական մեծություն է: Ենթադրվում է, որ հայտնի է հետևյալ պայմանական հավանականությունը՝

$$P(X = x|\theta) = \theta^x (1 - \theta)^{1-x}:$$

Գտնենք հետևյալ պայմանական հավանականությունը՝

$$P(X = x|S^n) = \sum_{\theta} P(X = x|\theta, S^n)P(\theta|S^n) = \sum_{\theta} P(X = x|\theta)P(\theta|S^n):$$

Երկրորդ հավասարությունը հետևում է այն փաստից, որ մենք այստեղ $X = x$ և S^n պատահարները համարում ենք անկախ θ պայմանի առկայության պարագայում:

Ըստ Բայեսի բանաձևի՝ ունենք

$$P(\theta|S^n) = \frac{P(S^n|\theta)P(\theta)}{P(S^n)} = \frac{1}{P(S^n)} \prod_{i=1}^n P(X = x_i|\theta)P(\theta):$$

Տեղադրելով սա վերևի հավասարության մեջ՝ կստանանք

$$P(X = x|S^n) = \frac{1}{P(S^n)} \sum_{\theta} P(X = x_i|\theta) \prod_{i=1}^n P(X = x_i|\theta)P(\theta):$$

Եթե θ պարամետրն անընդհատ պատահական մեծություն է և ունի հավանականային $P(\theta)$ խտություն, ապա

$$P(X = x|S^n) = \frac{1}{P(S^n)} \int_0^1 \theta^{x+\sum_i x_i} (1-\theta)^{1-x+\sum_i (1-x_i)} P(\theta) d\theta:$$

Դժվար չէ նկատել, որ եթե θ -ն հավասարաչափ է բաշխված ($P(\theta) = 1$), ապա

$$P(X = x|S^n) = \frac{1}{P(S^n)} \int_0^1 \theta^{x+\sum_i x_i} (1-\theta)^{1-x+\sum_i (1-x_i)} d\theta:$$

Այստեղից՝

$$P[(S^n, x)] = P(X = x|S^n)P(S^n) = \int_0^1 \theta^{x+\sum_i x_i} (1-\theta)^{1-x+\sum_i (1-x_i)} d\theta:$$

Դիցուք S^n հաջորդականության մեջ կա n_1 հատ մեկեր և n_2 հատ զրոներ: Նշանակենք՝

$$P(S^n) \equiv P_L(S^n) = \int_0^1 \theta^{n_1} (1-\theta)^{n_2} d\theta, \quad n_1 + n_2 = n:$$

Ինդուկցիայով հաշվենք այս ինտեգրալը: Ունենք՝

$$n_1 = n, \quad \int_0^1 \theta^n d\theta = \frac{1}{n+1}:$$

Ենթադրենք՝

$$\int_0^1 \theta^{n_1+1} (1-\theta)^{n_2-1} d\theta = \frac{1}{(n+1)(C_n^{n_1+1})}:$$

Կատարելով մասերով ինտեգրում՝ կստանանք

$$\int_0^1 \theta^{n_1} (1-\theta)^{n_2} d\theta = \frac{n-n_1}{n_1+1} \int_0^1 \theta^{n_1+1} (1-\theta)^{n_2-1} d\theta = \frac{1}{(n+1)C_n^{n_1}}:$$

Հետևաբար՝

$$\begin{aligned} P_L(x_{n+1} = 1|S^n) &= \frac{P_L(S^n, 1)}{P_L(S^n)} = \\ &= \frac{\frac{1}{(n+2)C_{n+1}^{n_1+1}}}{\frac{1}{(n+1)C_n^{n_1}}} = \frac{n_1 + 1}{n + 2}, \\ P_L(x_{n+1} = 0|S^n) &= \frac{n_2 + 1}{n + 2}: \end{aligned}$$

Այս մեթոդով կանխատեսումը կրում է Լապլասի անունը: Նման կանխատեսման որակական գնահատականը կարելի է տալ որևէ կորստի ֆունկցիայի միջոցով: Համեմատենք այս կանխատեսումը մաքսիմալ ճշմարտանմանության մոդելով կանխատեսման հետ: Որպես կորստի ֆունկցիա այս մեդելում կարելի է վերցնել լոգարիթմական կորստի ֆունկցիան՝

$$L_p(S^n) = -\ln(\theta^{n_1}(1-\theta)^{n_2}):$$

Ինֆորմացիայի տեսությունից հայտնի է, որ այս թիվը համընկնում է այն երկուական բիտերի քանակին, որոնք անհրաժեշտ են S^n հաջորդականությունը կոդավորելու համար, որում կլինեն n_1 հատ 1 և n_2 զրոներ, իսկ մեկերի հանդես գալու հավանականությունը հավասար է θ -ի: Դժվար չէ նկատել, որ

$$\sup_{0 \leq \theta \leq 1} \theta^{n_1}(1-\theta)^{n_2} = \left(\frac{n_1}{n}\right)^{n_1} \left(\frac{n_2}{n}\right)^{n_2}:$$

Բայես-Լապլասի կանոնով կանխատեսման դեպքում որպես կորստի ֆունկցիա վերցնենք արտահայտությունը՝

$$L(S^n) = -\ln \int_0^1 \theta^{n_1}(1-\theta)^{n_2} d\theta:$$

Հետևաբար՝

$$\begin{aligned} L(w^n) - \inf_p L_\theta(w^n) &\leq \\ &\leq \ln \frac{\sup_{0 \leq \theta \leq 1} \theta^{n_1}(1-\theta)^{n_2}}{\int_0^1 \theta^{n_1}(1-\theta)^{n_2} d\theta} = \\ &= \ln \frac{\left(\frac{n_1}{n}\right)^{n_1} \left(\frac{n_2}{n}\right)^{n_2}}{\frac{1}{(n+1)C_n^{n_1}}} \leq \ln(n+1): \end{aligned}$$

Այսպիսով, կիրառելով Լապլասի կանոնը, մենք օգտագործում ենք լրացուցիչ $\ln(n+1)$ բիտեր օպտիմալ կոդի համեմատ: Բայց մյուս կողմից կարող ենք եզրակացնել, որ

$$\frac{L(S^n) - \inf_p L(S^n)}{n} \leq \frac{\ln(n+1)}{n} \rightarrow 0, \text{ եթե } n \rightarrow \infty:$$

Սի պարզ օրինակի վրա կատարենք կանխատեսումներ մաքսիմալ ճշմարտանմանության և Բայես-Լապլասի մոդելներով:

Ենթադրենք՝ հարցում է կատարվում դպրոցում, թե ով գիտի հետևյալ

ֆորմուլան CH_3OH (էթիլային սպիրտ): Ենթադրենք կատարված չորսհարյուր հարցումներից հարյուրասնյոթը ճիշտ են պատասխանել: Բնական է ենթադրել, որ հաջորդ հարցաքննվողի ճիշտ պատասխանի հավանականությունը հավասար է $117/400$: Այժմ ցույց տանք, որ այդ հավանականությունը մաքսիմալ հավանական է: Կասենք, որ X պատահական մեծությունը ունի Բեռնուլի բաշխում, եթե այն ընդունում է երկու արժեք՝ 1 կամ 0 p_0 և $1 - p_0$ հավանականություններով: Այդ դեպքում բաշխման ֆունկցիան ունի հետևյալ տեսքը՝

$$p(x|p_0) = p_0^x(1 - p_0)^{1-x}, \quad x \in \{0, 1\}:$$

Քանի որ կատարված է չորսհարյուր փորձ, ապա դրանց ելքերը հետևյալ վեկտորն է՝ $x = (x_1, x_2, \dots, x_{400})$: հավանականությունը այն բանի, որ կլինի 117 հատ մեկ չորսհարյուր փորձերից հավասար է

$$P(x, p_0) = \prod_{i=1}^{400} p_0^{x_i} (1 - p_0)^{1-x_i} = p_0^{117} (1 - p_0)^{283}:$$

Այժմ մաքսիմիզացնենք այս հավանականությունը՝ ընտրելով պարամետրը այնպես, որ այդ հավանականությունը լինի մաքսիմալ: Դա նույն է ինչ-որ մաքսիմիզացնել լոգարիթմական ֆունկցիան՝

$$\ln(p(x, p_0)) = 117 \ln p_0 + 283 \ln(1 - p_0):$$

Այս ֆունկցիան գոգավոր է, հետևաբար այն կետերը, որոնք նրա ածանցյալին զրո են դարձնում, դրանք մաքսիմումի կետեր են: Ածանցյալը հավասարեցնելով զրոյի՝ կստանանք $p_0 = 117/400$:

Իսկ եթե կանխատեսում անենք Բայես-Լապլասի մոդելով, ապա կստանանք $p_0 = 118/402$:

10.4 Weighted-Majority ալգորիթմը

Դիցուք ունենք N փորձագետներ, որոնք յուրաքանչյուր t քայլում տալիս են կանխատեսում $p_t^i \in \{0, 1\}$, $i = 1, 2, \dots, n$: Ենթադրենք՝ հնարավոր ելքերը նույնպես քինար հաջորդականություն են $\{w_1, w_2, \dots, w_t, \dots\}$: Ընդունենք մի փորձագետ ճիշտ է կանխատեսում բոլոր հնարավոր ելքերը, այսինքն՝

$$p_t^i = w_t, \quad t = 1, 2, \dots:$$

Դիտարկենք, այսպես ասած «մեծամասնության ալգորիթմը»: Այն յուրաքանչյուր t քայլում որոշում է այն փորձագետներին, որոնք մինչև այդ ոչ մի սխալ կանխատեսում չեն արել, այսինքն՝

$$B_t = \{i/p_j^i = w_j, \quad j = 1, 2, \dots, t - 1\}:$$

Ալգորիթմը t պահին գեներացնում է 1, եթե B_t բազմության կեսից ավելին այդ ժամանակ կանխատեսել են 1, հակառակ դեպքում տալիս է 0: Այսինքն՝

$$p_t = \begin{cases} 1, & \text{եթե } |i : i \in B_t, p_t^i = 1| \geq |B_t|/2, \\ 0, & \text{հակառակ դեպքում:} \end{cases}$$

Թեորեմ 10.2. Ենթադենք՝ կա մի փորձագետ, որը բոլոր քայլերում կատարում է ճիշտ կանխատեսում: Այսինքն՝ գոյություն ունի $i \in \{1, 2, \dots, N\}$, այնպիսին, որ

$$p_t^i = w_t, \quad t = 1, 2, \dots:$$

Այդ դեպքում «մեծամասնության» ալգորիթմը կատարում է ոչ ավել՝ քան $\lceil \log_2 N \rceil$ սխալներ:

Ապացույց: Եթե ալգորիթմը որևէ t պահի սխալ կանխագուշակում է անում, ապա

$$|B_{t+1}| \leq \lceil |B_t|/2 \rceil:$$

Ըստ ենթադրության՝ ունենք

$$|B_1| = N, \quad |B_t| \geq 1 \quad \forall t:$$

Հետևաբար

$$1 \leq B_{t+1} \leq N2^{-M},$$

որտեղ M -ը ալգորիթմի սխալվելու քանակն է $1, 2, \dots, t$ ժամանակահատվածում: Լոգարիթմելով այս անհավասարությունը՝ կստանանք

$$M \leq \log_2 N:$$

Այժմ դիտարկենք այն դեպքը, որ չկա փորձագետ, որը ճշգրիտ գուշակի ապագա ելքերը: Այս դեպքում կա «կշիռներով մեծամասնություն» ալգորիթմը, որը հետևյալն է:

- FOR $t=1, 2, \dots, T$ $w_0^1 = \dots w_0^N = 1$: i —րդ փորձագետը տալիս է կանխատեսում

$$p_t^i \in \{0, 1\}:$$

- Ալգորիթմը տալիս է իր կանխատեսումը՝ p_t :

IF

$$\sum_{i:p_t^i=0} w_t^i > \sum_{i:p_t^i=1} w_t^i,$$

THEN $p_t = 0$

ELSE $p_t = 1$

END IF

- «Բնությունը» տալիս է իր ելքը՝ $w_t \in \{0, 1\}$: Այնուհետև ալգորիթմը կատարում է փորձագետների կշիռների վերահաշվարկ: Դիցուք

$$E_t = \{i : p_t^i \neq w_t\}:$$

Սրանք այն փորձագետներ են, որոնք կատարել են սխալ կանխատեսում: Կատարվում է կշիռների փոփոխություն հետևյալ բանաձևով՝

$$w_{t+1}^i = \begin{cases} (1 - \epsilon)w_t^i, & \text{երբ } i \in E_t \\ w_t^i, & \text{հակառակ դեպքում:} \end{cases}$$

END FOR

Նշենք, որ մինչև տվյալ պահի ինֆորմացիան հասանելի է բոլոր փորձագետներին, այսինքն՝ «խաղը», այսպես ասած, «լրիվ ինֆորմացիայով խաղ է»: Դիցուք

$$0 < \epsilon < 1,$$

$$L_T^i = \sum_{t=1}^T |p_t^i - w_t^i|$$

(i -րդ փորձագետի բոլոր սխալ կանխատեսումների քանակն է T քայլերից հետո),

$$L_T = \sum_{t=1}^T |p_t - w_t|$$

(ալգորիթմի սխալվելու քանակը T ժամանակահատվածում):

Տեղի ունի հետևյալ թեորեմը՝

Թեորեմ 10.3. Կամայական i համար տեղի ունի

$$L_T \leq \left(\frac{2}{1 - \epsilon}\right)L_T^i + \frac{2}{\epsilon} \ln N$$

անհավասարությունը:

Ապացույց: Դիցուք

$$W_t = \sum_{i=1}^N w_t^i, \quad m = \min_{0 \leq i \leq N} L_T^i:$$

(m -ը «լավագույն» փորձագետի կատարած սխալների քանակն է): Դիցուք այդ մինիմումը հասանելի է դառնում i -րդ փորձագետի դեպքում: Այդ պարագայում ակնհայտ է, որ

$$W_t > w_t^i \geq (1 - \epsilon)^m \quad \forall t:$$

Մյուս կողմից, երբ ալգորիթմը կատարում է սխալ կանխատեսում t -րդ քայլում, ապա

$$\sum_{i \in E_t} w_t^i \geq \frac{W_t}{2}:$$

Հետևաբար՝

$$\begin{aligned} W_{t+1} &= \sum_{i \in E_t} (1 - \epsilon) w_t^i + \sum_{i \notin E_t} w_t^i = \\ &= \sum_{i=1}^N w_t^i - \epsilon \sum_{i \in E_t} w_t^i \leq \\ &\leq W_t (1 - \epsilon/2): \end{aligned}$$

Ակնհայտ է նաև (ըստ սահմանման)՝

$$W_{t+1} \leq W_t \quad \forall t:$$

Այստեղից կատանանք

$$\frac{W_T}{W_0} = \prod_{t=0}^{T-1} \frac{W_{t+1}}{W_t} \leq (1 - \epsilon/2)^M,$$

որտեղ $M = L_T$, այսինքն ալգորիթմի կատարած սխալների քանակն է T փուլերից հետո: Քանի որ

$$W_0 = \sum_{i=1}^N w_0^i = N,$$

ապա, հաշվի առնելով վերևի անհավասարությունները՝ կունենանք

$$\frac{(1 - \epsilon)^m}{N} < \frac{W_T}{W_0} \leq (1 - \epsilon/2)^{L_T}:$$

Լոգարիթմելով այս անհավասարության երկու կողմերը, կունենանք՝

$$m \ln(1 - \epsilon) - \ln N < L_T \ln(1 - \epsilon/2) \Rightarrow$$

$$m \ln(1 - \epsilon) - \ln N < -\frac{\epsilon}{2} L_T \Rightarrow$$

$$m \ln \frac{1}{1 - \epsilon} + \ln N > \frac{\epsilon}{2} L_T \Rightarrow$$

$$m \left(\frac{2}{\epsilon} \right) \ln \frac{1}{1 - \epsilon} + \left(\frac{2}{\epsilon} \right) \ln N > L_T \Rightarrow$$

$$\left(\frac{2}{1 - \epsilon} \right) m + \frac{2}{\epsilon} \ln N > L_T:$$

Վերջին անհավասարությունը ստացվել է կիրառելով

$$\frac{1}{y} \ln \left(\frac{1}{1 - y} \right) \leq \frac{1}{1 - y}$$

անհավասարությունը, որը $\ln(1 + x) \leq x$ հայտնի անհավասարության հետևանքն է, երբ

$$x = \frac{y}{1 - y}:$$

10.5 Առցանց կորուստների բաշխման օպտիմալ ալգորիթ

Այժմ դիտարկենք կանխատեսման ևս մեկ մոդել: Դիցուք ունենք N փորձագետներ, որոնք կանխատեսում են եղանակը: Յուրաքանչյուր t պահի փորձագետներից ամեն մեկը հաղորդում է իր կորստի մասին՝ $l_t^i \in [0, 1], i = 1, 2, \dots, n$: Սա նշանակում է, որ i -րդ փորձագետը t պահին l_t^i չափով սխալվել է իր կանխատեսման մեջ:

Սկզբնական պահին, , բոլոր փորձագետներին տրվում է հավասար կշիռներ՝

$$w_0^i = 1/N:$$

Ալգորիթը ստանալով $\bar{l}_t = (l_t^1, \dots, l_t^N)$ փորձագետների կորստի ֆունկցիան, կազմում է

$$\bar{p}_t = \frac{\bar{w}_t}{\sum_{i=1}^N w_t^i}$$

վեկտորը և այնուհետև կատարում է իր որոշումը՝

$$(\bar{p}_t, \bar{l}_t) = \sum_{i=1}^n p_t^i l_t^i,$$

որը մեկնաբանվում է որպես ալգորիթի կորուստ կամ սխալվելու չափանիշ: Դրանից հետո ալգորիթը կատարում է փորձագետների կշիռների վերահաշվարկ հետևյալ կերպ՝

$$w_{t+1}^i = w_t^i \beta^{l_t^i},$$

որտեղ

$$\beta = (1 + \sqrt{2/T})^{-1}:$$

,
Նշանակենք՝

$$L_T^i = \sum_{t=1}^T l_t^i, \quad L_T = \sum_{t=1}^T (\bar{p}_t, \bar{l}_t):$$

Ճշմարիտ է հետևյալ գնահատականը.

Թեորեմ 10.4.

$$L_T \leq \min_i L_T^i + \sqrt{2T} + \ln N:$$

Ապացույց: Քանի որ β^r ֆունկցիան ուռուցիկ է ($r, \beta \in (0, 1)$), ապա

$$b^r \leq 1 - (1 - \beta)r,$$

և հետևաբար կունենանք

$$\begin{aligned}
\sum_{i=1}^N w_{t+1}^i &= \sum_{i=1}^N w_t^i \beta^{l_t^i} \leq \\
&\leq \sum_{i=1}^N w_t^i (1 - (1 - \beta) l_t^i) = \\
&= \left(\sum_{i=1}^N w_t^i \right) (1 - (1 - \beta) (\overline{p}_t, \overline{l}_t)) :
\end{aligned}$$

Հաջորդաբար կիրառելով այս անհավասարությունները $t = 1, 2, \dots, T$ արժեքների համար՝ կստանանք

$$\begin{aligned}
\sum_{i=1}^N w_{T+1}^i &\leq \prod_{t=1}^T (1 - (1 - \beta) (\overline{p}_t, \overline{l}_t)) \leq \\
&\leq \exp \left(- (1 - \beta) \sum_{t=1}^T (\overline{p}_t, \overline{l}_t) \right) :
\end{aligned}$$

Այստեղից, քանի որ $\sum_{i=1}^N w_1^i = 1$, ապա

$$\ln \left(\sum_{i=1}^N w_{T+1}^i \right) \leq - (1 - \beta) L_T :$$

Հետևաբար՝

$$L_T \leq \frac{-\ln(\sum_{i=1}^N w_{T+1}^i)}{1 - \beta} :$$

Ունենք նաև

$$w_{T+1}^i = w_1^i \prod_{t=1}^T \beta^{l_t^i} = w_1^i \beta^{L_T^i} :$$

Ուստի

$$L_T \leq \frac{\ln(1/\beta)}{1 - \beta} \min_i L_T^i + \frac{\ln N}{1 - \beta} :$$

Լեմմա 10.1. Ենթադրենք՝

$$0 \leq L \leq \tilde{L}, \quad 0 \leq R \leq \tilde{R}, \quad \beta = \left(1 + \sqrt{2\tilde{R}/\tilde{L}} \right)^{-1} :$$

Այդ դեպքում

$$-\frac{\ln \beta}{1 - \beta} L + \frac{1}{1 - \beta} R \leq L + \sqrt{2\tilde{R}\tilde{L}} + R :$$

Ապացույց: Ճշմարիտ է հետևյալ անհավասարությունը, որից մենք կօգտվենք ապացույցի ընթացքում՝

$$-\ln \beta \leq \frac{1 - \beta^2}{2\beta}, \quad \beta \in (0, 1] :$$

Հետևաբար՝

$$L \frac{-\ln \beta}{1 - \beta} + \frac{1}{1 - \beta} R \leq L \frac{1 + \beta}{2\beta} + \frac{1}{1 - \beta} R =$$

$$\begin{aligned}
&= \frac{1}{2}L(1 + \frac{1}{\beta}) + \frac{1}{1-\beta}R = \\
&= L + \frac{1}{2}L\sqrt{2\tilde{R}/\tilde{L}} + \frac{1}{1 - \frac{1}{1+\sqrt{2\tilde{R}/\tilde{L}}}}R \leq \\
&\leq L + \sqrt{1/(2)\tilde{R}\tilde{L}} + R + R\sqrt{\tilde{L}/(2\tilde{R})} \leq \\
&L + \sqrt{2\tilde{L}\tilde{R}} + R:
\end{aligned}$$

Լեմման ապացուցվեց:

Այժմ քանի որ $L_T^i \leq T$, ապա տեղադրելով լեմմայի անհավասարության մեջ $\tilde{L} = T$, $\tilde{R} = \ln N$, կստանանք՝

$$L_T \leq \min_i L_T^i + \sqrt{2T \ln N} + \ln N:$$

Այս անհավասարությունից հետևում է, որ

$$\frac{L_T - \min_i L_T^i}{T} \rightarrow 0, \text{ երբ } T \rightarrow \infty:$$

Այսինքն՝ ալգորիթմի և լավագույն փորձագետի կատարած սխալների միջինը T ժամանակահատվածում ձգտում է զրոյի, երբ $T \rightarrow \infty$:

Վերջում նշենք, որ այս ալգորիթմի թերությունը այն է, որ β -ն կախված է T ժամանակահատվածից:

Այժմ բերենք մի օրինակ երկու փորձագետների դեպքում, որը ցույց կտա, որ պարզապես ընդունել որոշում հետևելով այսպես ասած լիդերի որոշմանը, կարող է մեծ կորուստներ պատճառել: Հիշեցնենք, որ լիդերը t պահին այն փորձագետն է, որի գումարային սխալանքը մինչև այդ պահը ամենաքիչն է: Դիցուք $t = 0, 1, 2, \dots, 6$, և առաջին ու երկրորդ փորձագետների կորուստներն են՝

$$l^1 = (1/2, 0, 1, 0, 1, 0, 1), \quad l^2 = (0, 1, 0, 1, 0, 1, 0):$$

$t = 0$ պահին լիդեր է երկրորդ փորձագետը: Դրանից հետո ալգորիթմի կորուստը կլինի 1: Այնուհետև լիդեր է առաջին փորձագետը: Հետևաբար ալգորիթմի կորուստը կավելանա ես մեկով: Եվ այսպես շարունակ՝ յուրաքանչյուր պահի ալգորիթմի կորուստը առնվազն երկու անգամ մեծ կլինի յուրաքանչյուր փորձագետի կորստից:

Այժմ նկարագրենք այս թերեմի ևս մի տարբերակ, որում տրվում է ավելի ճշգրիտ գնահատական:

Դիցուք

- N -ը փորձագետների քանակն է, իսկ T ՝ փուլերի (ռաունդների):
- Կատարվում է սկզբնական արժեքների վերագրում (ինիցիալիզացիա), և բոլոր փորձագետներին տրվում է միևնույն գնահատականը՝ $\hat{w}^1 = (1, \dots, 1)$:
- $w^t = \hat{w}^t / Z_t$, $Z_t = \sum_i \hat{w}_i^t$:

- t պահին ալգորիթմը ստանում է փորձագետների $v_t \in [0, 1]^N$ սխալանքի վեկտորը և այնուհետև ձևավորում է իր սխալանքի չափը՝ (w^t, v_t) :
- Այնուհետև ալգորիթմը կատարում է փորձագետներին տրված կշիռների թարմացում հետևյալ կանոնով՝

$$\hat{w}_i^{t+1} = \hat{w}_i^t e^{-\eta v_{t,i}}, \quad \eta = \sqrt{\frac{2 \ln N}{T}}:$$

Ճշմարիտ է սույն պնդումը.

Թեորեմ 10.5. *Տեղի ունի հետևյալ գնահատականը՝*

$$\sum_{t=1}^T (w^t, v_t) - \min_{i \in [N]} \sum_{t=1}^T v_{t,i} \leq \sqrt{2T \ln N}:$$

Ապացույց: Ունենք

$$\begin{aligned} \ln \frac{Z_{t+1}}{Z_t} &= \ln \sum_i \frac{\hat{w}_i^t}{Z_t} e^{-\eta v_{t,i}} = \\ &= \ln \sum_i w_i^t e^{-\eta v_{t,i}}: \end{aligned}$$

Քանի որ $e^{-a} \leq 1 - a + \frac{a^2}{2}$, $a \in (0, 1)$ և $\sum_i w_i^t = 1$, ապա

$$\begin{aligned} \ln \frac{Z_{t+1}}{Z_t} &\leq \ln \sum_i w_i^t (1 - \eta v_{t,i} + \eta^2 v_{t,i}^2 / 2) = \\ &= \ln \left(1 - \sum_i (\eta v_{t,i} - \eta^2 v_{t,i}^2 / 2) \right): \end{aligned}$$

Եկատենք, որ

$$b \equiv \sum_i w_i^t (\eta v_{t,i} - \eta^2 v_{t,i}^2 / 2) \in (0, 1):$$

Այդ դեպքում $\ln(1 - b) \leq -b$: Հետևաբար վերևի անհավասարությունից կստանանք

$$\begin{aligned} \ln \frac{Z_{t+1}}{Z_t} &\leq - \sum_i w_i^t (\eta v_{t,i} - \eta^2 v_{t,i}^2 / 2) = \\ &= -\eta (w^t, v_t) + \eta^2 \sum_i w_i^t v_{t,i}^2 / 2 \leq \\ &= -\eta (w^t, v_t) + \eta^2 / 2: \end{aligned}$$

Գումարելով այս անհավասարությունները ըստ t -ի՝ կստանանք

$$\ln Z_{T+1} - \ln Z_1 \leq -\eta \sum_{t=1}^T (w^t, v_t) + \frac{T\eta^2}{2}: \quad (10.1)$$

Քանի որ կամայական i -ի համար $\hat{w}_i^{T+1} = e^{-\eta \sum_t v_{t,i}}$, ապա

$$\ln Z_{T+1} = \ln \left(\sum_i e^{-\eta \sum_t v_{t,i}} \right) \geq \ln \left(\max_i e^{-\eta \sum_t v_{t,i}} \right) = -\eta \min_i \sum_t v_{t,i}:$$

Այստեղից, հաշվի առնելով (10.1) անհավասարությունը, ինչպես նաև այն փաստը, որ $\ln Z_1 = \ln N$, կստանանք

$$\sum_{t=1}^T (w^t, v_t) - \min_i \sum_{t=1}^T v_{t,i} \leq \frac{\ln N}{\eta} + \frac{\eta T}{2}:$$

Այստեղ տեղադրելով η -ի արժեքը՝ կստանանք պահանջվող անհավասարությունը:
Թեորեմն ապացուցվեց:

Այս թեորեմը ունի հետևյալ կարևոր մեկնաբանությունը: Դիցուք ունենք հիպոթեզների վերջավոր $H = \{h_1, h_2, \dots, h_N\}$ բազմությունը: Յուրաքանչյուր h_i հիպոթեզ դիտարկենք որպես մի փորձագետ, որը t -րդ փուլում կանխատեսում է x_t օբյեկտի նշիչը: Այդ կանխատեսման և իրական նշիչի տարբերությունը հիպոթեզի սխալանքն է՝

$$v_{t,i} = |h_i(x_t) - y_t|, \quad y_t \in \{0, 1\}:$$

Դրանից հետո ալգորիթմը կատարում է իր կանխատեսումը՝

$$p_t = \sum_i w_i^t h_i(x_t) \in [0, 1]:$$

Այդ դեպքում ալգորիթմի կանխատեսման սխալանքը կլինի

$$|p_t - y_t| = \left| \sum_{i=1}^N w_i^t h_i(x_t) - y_t \right| = \left| \sum_{i=1}^N w_i^t (h_i(x_t) - y_t) \right|:$$

Այժմ, եթե $y_t = 1$, ապա $h_i(x_t) - y_t \leq 0$, $\forall i$: Հետևաբար վերևի արտահայտությունը հավասար կլինի՝

$$\sum_i w_i^t |h_i(x_t) - y_t|:$$

Եթե $y_t = 0$, ապա $h_i(x_t) - y_t \geq 0$, $\forall i$, և հետևաբար վերևի արտահայտությունը նույնպես կհավասարվի

$$\sum_i w_i^t |h_i(x_t) - y_t|$$

արտահայտությանը: Ուրեմն ամեն դեպքում

$$|p_t - y_t| = \sum_{i=1}^N w_i^t |h_i(x_t) - y_t| = (w^t, v_t):$$

Այնուհետև կիրառելով թեորեմ 10.5-ը՝ կստանանք հետևյալ պնդումը:

Թեորեմ 10.6. Դիցուք H -ը հիպոթեզների վերջավոր բազմություն է: Այդ դեպքում գոյություն ունի օնլայն դասակարգման այնպիսի ալգորիթմ՝ յու-

ընտրվող $p_t \in [0, 1]$ կանխատեսումով, որ տեղի ունի հետևյալ անհավասարությունը՝

$$\sum_{t=1}^T |p_t - y_t| - \min_{h \in H} \sum_{t=1}^T |h(x_t) - y_t| \leq \sqrt{2T \ln(|H|)}:$$

Այս թեորեմից հետևում է, որ գոյություն ունի օնլայն բինար դասակարգման այնպիսի ալգորիթմ A , որ այդ ալգորիթմի համար

$$\frac{1}{T} \sup_{h \in H} \text{regret}_T(A) \rightarrow 0, \text{ երբ } T \rightarrow \infty,$$

որտեղ

$$\text{regret}_T(A) \equiv \sum_{t=1}^T |p_t - y_t| - \sum_{t=1}^T |h(x_t) - y_t|:$$

10.6 Օնլայն կանխատեսման ընդհանուր ալգորիթմ

Այժմ դիտարկենք կանխատեսման հետևյալ ընդհանուր մոդելը: Դիցուք ունենք երկու տեղանոց հետևյալ ֆունկցիան՝ $\lambda(\omega, \gamma)$, որտեղ $\omega \in \Omega$, $\gamma \in \Gamma$: Այս ֆունկցիան մեկնաբանենք որպես կորուստի ֆունկցիա, որը արժեքներ է ընդունում $[0, 1]$ հատվածից: Ենթադրենք ունենք N փորձագետներ, որոնք կատարում են կանխատեսումներ: Նկարագրենք հետևյալ ընթացակարգը:

$t = 1, 2, \dots$ փուլերում

- յուրաքանչյուր i -րդ փորձագետ կատարում է իր կանխատեսումը՝ $\xi_t^i \in \Gamma$:
- Այնուհետև «ալգորիթմը» ընդունում է իր որոշումը՝ $\gamma_t \in \Gamma$, ելնելով փորձագետների կատարած կանխատեսումներից:
- Այնուհետև «բնությունը» հաղորդում է ճշգրիտ ելքը՝ $\omega_t \in \Omega$:
- i -րդ փորձագետը հաշվում է իր կորուստը՝

$$L_t^i = L_{t-1}^i + \lambda(\omega_t, \xi_t^i):$$

- « Ալգորիթմը » նույնպես հաշվում է իր գումարային կորուստը t քայլերի ընթացքում՝

$$L_t = L_{t-1} + \lambda(\omega_t, \gamma_t),$$

որտեղ $L_0 = 0$:

Խնդիրն այն է, որ ալգորիթմը պետք ընտրի այնպիսի որոշումներ՝ $\gamma_1, \gamma_2, \dots$, որ

$$\lim_{T \rightarrow \infty} \frac{1}{T} (L_T - \min_i L_T^i) \leq 0:$$

Այժմ ենթադրենք, որ Γ -ն ուռուցիկ ենթաբազմություն է $\mathbb{R}^n$ -ից, իսկ λ -ն ուռուցիկ ֆունկցիա է ըստ γ պարամետրի: Ենթադրենք նաև, որ ալգորիթմը t -րդ փուլում իր կանխատեսումը կատարում է հետևյալ կերպ՝

$$\gamma_t = \frac{\sum_{i=1}^N w_{t-1}^i \xi_t^i}{\sum_{i=1}^N w_{t-1}^i} = \sum_{i=1}^N \hat{w}_{t-1}^i \xi_t^i, \quad (10.2)$$

որտեղ

$$\hat{w}_{t-1}^i = \frac{w_{t-1}^i}{\sum_{i=1}^N w_{t-1}^i}, \quad w_{t-1}^i = e^{-\eta L_{t-1}^i}, \quad \gamma > 0: \quad (10.3)$$

Նշենք, որ այստեղ L_{t-1}^i -ն i -րդ փորձագետի ընդհանուր կորուստն է $t-1$ փուլերի ընթացքում:

Թեորեմ 10.7. Վերը նշված պայմանների դեպքում տեղի ունի հետևյալ անհավասարությունը՝

$$L_T - \min_i L_T^i \leq \frac{\ln N}{\eta} + \frac{T\eta}{8}, \quad (10.4)$$

որտեղ

$$L_T = \sum_{t=1}^T \lambda(\omega_t, \gamma_t), \quad L_T^i = \sum_{t=1}^T \lambda(\omega_t, \xi_t^i):$$

Ապացույց: Նշանակենք՝

$$W_t \equiv \sum_{i=1}^N w_t^i = \sum_{i=1}^N e^{-\eta L_t^i}, \quad W_0 = N:$$

Քանի որ $\omega_0^i = 1$, $i = 1, 2, \dots, N$, ապա

$$\ln \frac{W_T}{W_0} = \ln \left(\sum_{i=1}^N e^{-\eta L_T^i} \right) - \ln N \geq$$

$$\ln \left(\max_i e^{-\eta L_T^i} \right) - \ln N = -\eta \min_i L_T^i - \ln N: \quad (10.5)$$

Մյուս կողմից

$$\ln \frac{W_t}{W_{t-1}} = \ln \frac{\sum_{i=1}^N e^{-\eta \lambda(\omega_t, \xi_t^i)} e^{-\eta L_{t-1}^i}}{\sum_{i=1}^N e^{-\eta L_{t-1}^i}} = \ln E[e^{-\eta \lambda(\omega_t, \xi_t^i)}], \quad (10.6)$$

որտեղ մաթեմատիկական սպասումը դիտարկվում է համաձայն

$$\hat{w}_{t-1}^i = \frac{w_{t-1}^i}{\sum_{i=1}^N w_{t-1}^i}, \quad i = 1, 2, \dots, N$$

հավանականային բաշխման: Այժմ կիրառենք Հոֆդինգի անհավասարությունը՝

$$\ln E[e^{sX}] \leq sE[X] + \frac{s^2(b-a)^2}{8}:$$

$X = \lambda(\omega_t, \xi_t^i)$ պատահական մեծության նկատմամբ, որի հանդես գալու

հավանականությունը հավասար է $\hat{\omega}_{t-1}^i$, իսկ $a = 0$, $b = 1$: Հաշվի առնելով նաև, որ λ ֆունկցիան ուռուցիկ է ըստ երկրորդ արգումենտի (10.6) անհավասարությունից՝ կունենանք՝

$$\begin{aligned}\ln \frac{W_t}{W_{t-1}} &\leq -\eta \frac{\sum_{i=1}^N w_{t-1}^i \lambda(\omega_t, \xi_t^i)}{\sum_{i=1}^N w_{t-1}^i} + \frac{\eta^2}{8} \leq \\ &\leq -\eta \lambda\left(\omega_t, \frac{\sum_{i=1}^N w_{t-1}^i \xi_t^i}{\sum_{i=1}^N w_{t-1}^i}\right) + \frac{\eta^2}{8} = \\ &= -\eta \lambda(\omega_t, \gamma_t) + \frac{\eta^2}{8},\end{aligned}$$

որտեղ γ_t -ն ալգորիթմի կանխատեսումն է ըստ (10.2)-ի բանաձևի: Գումարելով այս անհավասարությունները ըստ $t = 1, 2, \dots, T$ -ի, կստանանք՝

$$\ln \frac{W_T}{W_0} = \sum_{t=1}^T \ln \frac{W_t}{W_{t-1}} \leq -\eta L_T + \frac{\eta^2}{8} T : \quad (10.7)$$

Այստեղից, օգտագործելով նաև (10.5) անհավասարությունը՝ կստանանք

$$L_T \leq \min_i L_T^i + \frac{\ln N}{\eta} + \frac{\eta}{8} T :$$

Եթե վերցնենք $\eta = \sqrt{8 \ln N / T}$, ապա անհավասարության աջ մասը հավասար կլինի $\sqrt{(1/2) T \ln N}$: Թեորեմն ապացուցվեց:

10.7 Ֆոն Նեյմանի թեորեմը մինիմաքսի մասին

Այս պարագրաֆում տրվում է խաղերի տեսության հիմնական թեորեմի (Նեյմանի թեորեմը մինիմաքսի մասին) ապացույցը՝ ելնելով նախորդ պարագրաֆում քննարկված օնլայն կանխատեսման ընդհանուր ալգորիթմում ստացված գնահատականից:

Դիցուք X -ը և Y -ը կամայական բազմություններ են: Դիտարկենք երկու հոգու անտագոնիստական խաղ: Առաջին խաղացողը ընտրում $x \in X$ ստրատեգիա, իսկ երկրորդ խաղացողը վերցնում է y ստրատեգիա Y բազմությունից: Ենթադրվում է, որ խաղացողները այդ ընտրությունները կատարում են իրարից անկախ: Դիցուք $f(x, y)$ -ը առաջին խաղացողի շահույթի և միաժամանակ երկրորդի՝ կորստի ֆունկցիան է: Երբ $f(x, y) < 0$, ապա առաջինի շահույթը բացասական է, և հետևաբար $f(x, y)$ -ը տվյալ դեպքում նրա կորստի չափն է: Առաջին խաղացողի նպատակն է մաքսիմիզացնել իր շահույթը, իսկ երկրորդ խաղացողինը՝ մինիմիզացնել իր կորուստը:

Եթե առաջին խաղացողը ընտրել է $x \in X$ ստրատեգիա, ապա նրա շահույթը կլինի ոչ պակաս, քան $\inf_{y \in Y} f(x, y)$ անկախ այն բանից, թե ինչ կընտրի երկրորդ խաղացողը: Այս մեծությունը կոչվում է առաջինի երաշխավորված եկամուտ, որը

հետևյալ թիվն է՝

$$\underline{v} = \sup_{x \in X} \inf_{y \in Y} f(x, y):$$

$x_0 \in X$ վեկտորը կոչվում է «maxmin» ստրատեգիա, եթե

$$\underline{v} = \inf_{y \in Y} f(x_0, y):$$

Անալոգիայով, երկրորդ խաղացողի երաշխավորված արդյունք կոչվում է

$$\bar{v} = \inf_{y \in Y} \sup_{x \in X} f(x, y)$$

մեծությունը:

$y_0 \in Y$ վեկտորը կոչվում է «minmax» ստրատեգիա, եթե

$$\bar{v} = \sup_{x \in X} f(x, y_0):$$

Տեղի ունի հետևյալ պնդումը:

Լեմմա 10.2. Կամայական անտագոնիստական խաղում $\underline{v} \leq \bar{v}$, այսինքն՝

$$\sup_{x \in X} \inf_{y \in Y} f(x, y) \leq \inf_{y \in Y} \sup_{x \in X} f(x, y):$$

Ապացույց: Կամայական $x \in X, y \in Y$ -ի համար ունենք

$$\inf_{y \in Y} f(x, y) \leq f(x, y) \leq \sup_{x \in X} f(x, y):$$

Այստեղից՝

$$\inf_{y \in Y} f(x, y) \leq \sup_{x \in X} f(x, y):$$

Քանի որ այս անհավասարության ձախ մասը կախված է x -ից, իսկ աջ մասը՝ ոչ, ապա

$$\sup_{x \in X} \inf_{y \in Y} f(x, y) \leq \sup_{x \in X} f(x, y), \quad \forall y \in Y:$$

Հետևաբար՝

$$\underline{v} = \sup_{x \in X} \inf_{y \in Y} f(x, y) \leq \inf_{y \in Y} \sup_{x \in X} f(x, y) = \bar{v}:$$

$(x_0, y_0) \in X \times Y$ կետը կոչվում է f -ի թամբի կետ, եթե

$$f(x, y_0) \leq f(x_0, y_0) \leq f(x_0, y) \quad \forall x \in X, y \in Y:$$

Թամբի կետ երկու խաղացողի համար ձեռնտու է: Դա կայուն վիճակ է, որովհետև սահմանումից հետևում է, որ եթե մի խաղացող փորձի փոխել իր ստրատեգիան, ապա դրանից ինքը չի շահի: Հետևաբար (x_0, y_0) թամբի կետը կոչվում է խաղի լուծում, և անտագոնիստական խաղի հիմնական հարցը նման լուծման գոյության և որոնման խնդիրն է: Տեղի ունի թամբի կետի գոյության հետևյալ անհրաժեշտ և

բավարար պայմանը:

Թեորեմ 10.8. Որպեսզի $f(x, y)$ ֆունկցիան ունենա թամբի կետ, անհրաժեշտ է և բավարար, որ

$$\max_{x \in X} \inf_{y \in Y} f(x, y) = \min_{y \in Y} \sup_{x \in X} f(x, y) : \quad (10.8)$$

Բացի դրանից՝ x_0 -ն «maxmin» կետ է, իսկ y_0 -ն՝ «minmax»:

Ապացույց: Ենթադրենք՝ (x_0, y_0) -ն թամբի կետ է: Այդ դեպքում

$$\bar{v} \leq \sup_{x \in X} f(x, y_0) \leq f(x_0, y_0) \equiv v = \inf_{y \in Y} f(x_0, y) \leq \underline{v}:$$

Այստեղից, համաձայն լեմմա 10.2-ի, կունենանք՝

$$\underline{v} = \bar{v}:$$

Այժմ ենթադրենք, որ տեղի ունի (10.8) հավասարությունը: Դիցուք x_0 -ն «maxmin» ստրատեգիա է, իսկ y_0 -ն՝ «minmax»: Ցույց տանք, որ դրանց զույգը թամբի կետ է: Իրոք, ունենք

$$f(x_0, y_0) \geq \inf_{y \in Y} f(x_0, y) = \underline{v} = \bar{v} = \sup_{x \in X} f(x, y_0) \geq f(x_0, y_0):$$

Այստեղից հետևում է, որ բոլոր անհավասարությունները կարելի է փոխել հավասարության նշանով, և հետևաբար (x_0, y_0) -ն թամբի կետ է:

Այժմ բերենք խաղի օրինակ, որտեղ չկա թամբի կետ: Դիցուք

$$X = Y = \{1, 2\}, \quad f(i, j) = a_{i,j}, \quad i = 1, 2, \quad j = 1, 2,$$

իսկ շահույթի $(a_{i,j})$ մատրիցը ունի հետևյալ տեսքը՝

$$\begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}:$$

Տվյալ դեպքում առաջին խաղացողի լավագույն երաշխավորված եկամուտը կլինի՝

$$\underline{v} = \max_i \min_j = -1:$$

Իսկ երկրորդ խաղացողի լավագույն երաշխավորված արդյունքը կլինի՝

$$\bar{v} = \min_j \max_i a_{i,j} = 1:$$

Հետևաբար խաղը թամբի կետ չունի: Այժմ քննարկենք, թե ինչպիսի բավարար պայմանների դեպքում խաղը կունենա թամբի կետ: Տեղի ունի հետևյալ թեորեմը, որը գրականության մեջ հայտնի է որպես Նեյմանի թեորեմ: Այն համարվում է հակամարտ (անտագոնիստական) խաղերի հիմնական թեորեմ:

Թեորեմ 10.9. Դիցուք $X \subset \mathbb{R}^n$, $Y \subset \mathbb{R}^m$ ուռուցիկ կոմպակտ բազմություններ

են: $f(x, y)$ անընդհատ ֆունկցիան ուռուցիկ է ըստ x -ի և գոգավոր ըստ y -ի: Այդ դեպքում գոյություն ունի թամբի կետ:

Ապացույց: Y բազմության կոմպակտությունից հետևում է, որ դրանում գոյություն ունի ϵ -ցանց: Այսինքն՝ գոյություն ունի այնպիսի վերջավոր $\{y^1, y^2, \dots, y^N\} \subset Y$ բազմություն, որ Y -ի յուրաքանչյուր կետ գտնվում է այդ բազմության ϵ -շրջակայքում: Ռեկուրսիվ կառուցենք կետերի $y_1, y_2, \dots, y_n \in Y$ և $x_1, x_2, \dots, x_n \in X$ հաջորդականություններ հետևյալ կերպ: Դիցուք $x_0 \in X$: Կամայական $t = 1, 2, \dots, n$ ինդեքսների համար սահմանենք

$$y_t = \frac{\sum_{i=1}^N y^i \exp(-\eta \sum_{s=0}^{t-1} f(x_s, y^i))}{\sum_{i=1}^N \exp(-\eta \sum_{s=0}^{t-1} f(x_s, y^i))},$$

$$\eta = \sqrt{(8 \ln N)/n}:$$

Իսկ x_t -ն ընտրվում այնպես, որ

$$f(x_t, y_t) = \max_{x \in X} f(x, y_t):$$

Այժմ կիրառենք նախորդ պարագրաֆի թեորեմ 10.7-ի արդյունք $\lambda(x, y) = f(x, y)$ կորստի ֆունկցիայի դեպքում: Այս պարագայում y^i -ին i -րդ փորձագետի կանխատեսումն է, $i = 1, 2, \dots, N$, x_t -ն ռեալ էլքերն են, $t = 1, 2, \dots, n$, y_t -ն ալգորիթմի կանխատեսումներն են: Համաձայն վերոհիշյալ թեորեմի կունենանք

$$\sum_{t=1}^n f(x_t, y_t) \leq \min_{i=1,2,\dots,N} \sum_{t=1}^n f(x_t, y^i) + \sqrt{(1/2)n \ln N}:$$

Այստեղից

$$\frac{1}{n} \sum_{t=1}^n f(x_t, y_t) \leq \frac{1}{n} \min_{i=1,2,\dots,N} \sum_{t=1}^n f(x_t, y_i) + \sqrt{(\ln N)/(2n)}: \quad (10.9)$$

Այստեղից՝ հաշվի առնելով նաև, որ $f(x, y)$ -ը ուռուցիկ է ըստ x -ի և գոգավոր ըստ y -ի, կստանանք՝

$$\begin{aligned} \min_{y \in Y} \max_{x \in X} f(x, y) &\leq \\ &\leq \max_{x \in X} f(x, \frac{1}{n} \sum_{t=1}^n y_t) \leq \\ &\leq \max_{x \in X} \frac{1}{n} \sum_{t=1}^n f(x, y_t) \leq \\ &\leq \frac{1}{n} \sum_{t=1}^n \max_{x \in X} f(x, y_t) \leq \\ &\leq \frac{1}{n} \sum_{t=1}^n f(x_t, y_t) \leq \\ &\leq \min_{i=1,2,\dots,N} \frac{1}{n} \sum_{t=1}^n f(x_t, y_i) + \sqrt{\ln N/(2n)} \leq \end{aligned}$$

$$\begin{aligned} &\leq \min_{i=1,\dots,N} f\left(\frac{1}{n} \sum_{t=1}^n x_t, y^i\right) + \sqrt{\ln N/(2n)} \leq \\ &\leq \max_{x \in X} \min_{i=1,\dots,N} f(x, y^i) + \sqrt{\ln N/(2n)}: \end{aligned}$$

Այստեղ անցնելով սահմանի, երբ $n \rightarrow \infty$, կստանանք

$$\min_{y \in Y} \max_{x \in X} f(x, y) \leq \max_{x \in X} \min_{i=1,\dots,N} f(x, y^i):$$

Այնուհետև ϵ -ը ձգտեցնելով զրոյի՝ վերջնականում կստանանք՝

$$\min_{y \in Y} \max_{x \in X} f(x, y) \leq \max_{x \in X} \min_{y \in Y} f(x, y):$$

Թեորեմն ապացուցվեց:

10.8 Մեքենայական ուսուցման ֆունդամենտալ թեորեմը օնլայն ուսուցման համակարգում

Այս պարագրաֆում քննարկվում են օպտիմալ պրեդիկտորների գոյության պրոբլեմները օնլայն ուսուցման ուռուցիկ խնդիրներում:

Դիցուք

- A -ն օնլայն ուսուցման ալգորիթմ է, $H \subset \mathbb{R}^d$ -ը՝ հիպոթեզների ուռուցիկ բազմություն, իսկ $\delta \in (0, 1)$ պարամետր է:
- Սկզբնական փուլում ընտրվում է $h_1 \in H$ հիպոթեզ:
- t -րդ փուլում վերցվում է (x_t, y_t) օբյեկտը հավանականային D բաշխումից և տրվում կորստի ուռուցիկ l ֆունկցիայի մուտքին՝ $f_t(h) \equiv l(h(x_t), y_t)$:
- A ալգորիթմը աշխատում է ֆունկցիաների $(f_1, \dots, f_t)$ հաջորդականության վրա, ելքում տալով նոր հիպոթեզ՝ $h_{t+1} = A(f_1, \dots, f_t)$:
- Այս ցիկլը կատարվում է $t = 1, 2, \dots, T$ փուլերի համար:
- Վերջում ալգորիթմը վերադարձնում է $\bar{h} = (1/T) \sum_{t=1}^T h_t$ հիպոթեզը:

Ցույց տանք, որ $\bar{h}$ հիպոթեզը բավարարում է հետևյալ կարևոր պայմանը:

Թեորեմ 10.10. Ցանկացած $\delta \in (0, 1)$ թվի համար տեղի ունի հետևյալ անհավասարությունը՝

$$P\left(\text{error}(\bar{h}) \leq \min_{h \in H} \text{error}(h) + \frac{\sup_{h \in H} \text{regret}_T(A)}{T} + \sqrt{\frac{8 \ln(2/\delta)}{T}}\right) > 1 - \delta,$$

որտեղ $\text{error}(h) = E_{(x,y) \sim D} l(h(x), y)$:

Ապացույց: Ապացույցը հենվում է հետևյալ լեմմայի վրա:

Լեմմա (Ազդումա): Դիցուք ունենք $\{X_i\}_{i=1}^T$ պատահական մեծությունների հաջորդականությունը, որը բավարարում է հետևյալ պայմանները՝

- $E[X_{t+1}|X_t, X_{t-1}, \dots, X_1] = X_t$:
- $|X_{t+1} - X_t| \leq 1, t = 1, 2, \dots, T-1$:

Այդ դեպքում

$$P[|X_t - X_0| > c] \leq 2e^{\frac{-c^2}{2T}}:$$

Այս անհավասարությունից հետևում է, որ

$$P[X_T - X_0 > c] \leq e^{\frac{-c^2}{2T}},$$

$$P[X_0 - X_T > c] \leq e^{\frac{-c^2}{2T}}:$$

Պարզության համար ենթադրենք, որ $l(\hat{y}, y) \in [0, 1], y, \hat{y} \in Y$:
Կատարենք նշանակումներ՝

$$Z_t \equiv \text{error}(h_t) - l(h(x_t), y_t), X_t = \sum_{i=1}^t Z_t:$$

Նկատենք, որ

$$E_{(x,y) \sim D}[Z_t|X_{t-1}] = \text{error}(h_t) - E_{(x,y) \sim D}[l(h_t(x), y)] = 0:$$

Հետևաբար՝

$$E[X_{t+1}|X_t, X_{t-1}, \dots, X_1] = E[Z_{t+1}|X_t] + X_t = X_t:$$

Ակնհայտ է նաև, որ

$$|X_t - X_{t-1}| = |Z_t| \leq 1:$$

Ուստի կատարվում են Ազդումայի լեմմայի պայմանները:
Հետևաբար՝

$$P(X_T > c) \leq e^{-c^2/2T} : \quad (10.10)$$

Տեղադրելով այս անհավասարության մեջ $c = \sqrt{2T \ln(2/\delta)}$ ՝ կստանանք

$$P\left[\frac{1}{T} \sum_{t=1}^T \text{error}(h_t) - \frac{1}{T} \sum_{t=1}^T l(h_t(x_t), y_t) > \sqrt{\frac{2 \ln(2/\delta)}{T}}\right] \leq \frac{\delta}{2} : \quad (10.11)$$

Նման անհավասարություն ճշմարիտ է նաև $h^* = \arg \min_{h \in H} \text{error}(h)$ հիպոթեզի նկատմամբ: Այսինքն՝

$$P\left[\frac{1}{T} \sum_{t=1}^T \text{error}(h^*) - \frac{1}{T} \sum_{t=1}^T l(h^*(x_t), y_t) < -\sqrt{\frac{2 \ln(2/\delta)}{T}}\right] \leq \frac{\delta}{2} : \quad (10.12)$$

Հարմարության համար նշանակենք

$$G_1 = \frac{1}{T} \sum_{t=1}^T \text{error}(h_t) - \frac{1}{T} \sum_{t=1}^T l(h_t(x_t), y_t),$$

$$G_2 = \frac{1}{T} \sum_{t=1}^T \text{error}(h^*) - \frac{1}{T} \sum_{t=1}^T l(h^*(x_t), y_t):$$

Այս նշանակումներից հետո կունենանք՝

$$\begin{aligned} & \frac{1}{T} \sum_{t=1}^T \text{error}(h_t) - \text{error}(h^*) = \\ & G_1 - G_2 + \frac{1}{T} \sum_{t=1}^T l(h_t(x_t), y_t) - \frac{1}{T} \sum_{t=1}^T l(h^*(x_t), y_t) \leq \\ & \leq \frac{\sup_{h \in H} \text{regret}_T(A)}{T} + G_1 - G_2: \end{aligned}$$

Այստեղից և (10.11) – (10.12) անհավասարություններից կստանանք՝

$$\begin{aligned} & P\left[\frac{1}{T} \sum_{t=1}^T \text{error}(h_t) - \text{error}(h^*) > \frac{\text{regret}_T(A)}{T} + 2\sqrt{\frac{2 \ln(2/\delta)}{T}}\right] \\ & \leq P\left[G_1 - G_2 > 2\sqrt{\frac{2 \ln(2/\delta)}{T}}\right] \\ & \leq P\left[G_1 > \sqrt{\frac{2 \ln(2/\delta)}{T}}\right] + P\left[G_2 \leq -\sqrt{\frac{2 \ln(2/\delta)}{T}}\right] \leq \delta: \end{aligned}$$

Ըստ Յենսենի անհավասարության՝ ունենք

$$\text{error}(\bar{h}) \leq \frac{1}{T} \sum_{t=1}^T \text{error}(h_t):$$

Հետևաբար առնվազն $1 - \delta$ հավանականությամբ տեղի ունի

$$\begin{aligned} & \text{error}(\bar{h}) \leq \frac{1}{T} \sum_{t=1}^T \text{error}(h_t) \leq \\ & \leq \text{error}(h^*) + \frac{\sup_{h \in H} \text{regret}_T(A)}{T} + \sqrt{\frac{8 \ln(2/\delta)}{T}} \end{aligned}$$

անհավասարությունը:

Թեորեմն ապացուցվեց:

Դիտողություն: Վերը դիտարկված ալգորիթմները այնպիսին էին, որ $\text{regret}_T(A)/T \rightarrow 0$, և հետևաբար ցանկացած դրական ϵ թվի համար կարող ենք ընտրել այնպիսի $m(\epsilon, \delta)$ բնական թիվ, որ առնվազն $1 - \delta$ հավանականությամբ

տեղի ունի

$$\text{error}(\bar{h}) \leq \min_{h \in H} \text{error}(h) + \epsilon$$

անհավասարությունը $T \geq m(\epsilon, \delta)$ փուլերի համար: Այսինքն՝ $\bar{h}$ -ը ϵ -օպտիմալ պրեդիկտոր է:

11

Զափողականության կրճատում

Զափողականության կրճատումը մեծ չափողականության մեջ գտնվող տրվյալների արտապատկերումն է ավելի փոքր չափողականության տարածության վրա: Դրա անհրաժեշտությունը որոշ դեպքերում կարևոր է: Օրինակ՝ մոտիկ հարևանների այգորիթմում տարածության չափողականության մեծացման գույքընթաց էքսպոնենցյալ աճում է այգորիթմի բարդությունը (տես 12.3 պարագրաֆը): Եվ այսպես, տված են տվյալներ $\mathbb{R}^d$ տարածությունում: Անհրաժեշտ է այդ տվյալները $W^{n,d}$ գծային ձևափոխության միջոցով ($x \rightarrow Wx$) ներդնել $\mathbb{R}^n$, $n < d$ տարածությունում այնպես, որ ողջամիտ ճշգրտությամբ հնարավոր լինի վերականգնել x վեկտորը Wx վեկտորի միջոցով: Հեշտ է տեսնել, որ ճշգրիտ վերականգնել x վեկտորը Wx վեկտորի օգնությամբ հնարավոր չէ: Իրոք, եթե $A^{n,d}$, $n \leq d - 1$ մատրից է, ապա գոյություն ունեն այնպիսի $u, v \in \mathbb{R}^n$, $u \neq v$ վեկտորներ, որ $Au = Av$:

11.1 PCA (Principal component analysis) ալգորիթմը

Գլխավոր էլեմենտների մեթոդը այն հիմնական գործիքներից է, որն օգտագործվում է տվյալները տեսողական (vizual) ներկայացնելու և պահվող տեղեկատվության ծավալը սեղմելու համար: Մոդելի ծավալների կրճատումը հաջողության հասնելու անհրաժեշտ պայման է: Դիցուք տրված է $x_1, x_2, \dots, x_m$ վեկտորների համախումբ $\mathbb{R}^d$ տարածությունում: Պետք է ընտրել այնպիսի «սեղմող» $W^{n,d}$, $n < d$ և «վերականգնող» $U^{d,n}$ մատրիցներ, որ

$$\arg \min_{W^{n,d}, U^{d,n}} \sum_{i=1}^m \|x_i - UWx_i\|^2 \rightarrow \min : \quad (11.1)$$

Թեորեմ 11.1. Դիցուք W և U մատրիցները (11.1) խնդրի լուծումներն են: Այդ դեպքում

$$U^T U = I, \quad W = U:$$

Ապացույց: Ֆիքսենք կամայական U, W մատրիցներ և դիտարկենք $x \rightarrow UWx$ գծային ձևափոխությունը: Դրա արժեքների բազմությունը՝ $Q = \{UWx, x \in \mathbb{R}^d\}$, n -չափանի գծային ենթատարածություն է $\mathbb{R}^d$ -ում: Դիցուք $V^{d,n}$ մատրից է, որի սյուները Q ենթատարածության օրթոնորմալ բազիսային վեկտորներն են՝ $V^T V = I$: Այսպիսով յուրաքանչյուր վեկտոր Q -ից կարող ենք ներկայացնել Vy , $y \in \mathbb{R}^n$ տեսքով: Հետևաբար ցանկացած $x \in \mathbb{R}^d$ և $y \in \mathbb{R}^n$ վեկտորների համար ունենք

$$\|x - Vy\|^2 = \|x\|^2 + (Vy, Vy) - 2(x, Vy) = \|x\|^2 + \|y\|^2 - 2(y, V^T x):$$

Մինիմիզացնենք այս արտահայտությունը ըստ y -ի: Ածանցելով այն և հավասարեցնելով զրոյի՝ կստանանք $y = V^T x$: Այսպիսով

$$VV^T x = \arg \min_{y \in \mathbb{R}^n} \|x - Vy\|^2:$$

Հետևաբար՝

$$\sum_{i=1}^m \|x_i - UWx_i\|^2 \leq \sum_{i=1}^m \|x_i - VV^T x_i\|^2:$$

Թեորեմն ապացուցվեց:

Այսպիսով կարող ենք եզրակացնել, որ (10.1) խնդիրը կարող ենք վերաձևակերպել հետևյալ կերպ՝

$$\arg \min_{U \in \mathbb{R}^{d,n}, U^T U = I} \sum_{i=1}^m \|x_i - UU^T x_i\|^2 \rightarrow \min:$$

Պարզեցնենք այս խնդիրը: Կամայական $x \in \mathbb{R}^d$ վեկտորի և $U^{d,n}$, ($U^T U = I$) մատրիցի համար ունենք

$$\|x - UU^T x\|^2 = \|x\|^2 - 2(x, UU^T x) + (UU^T x, UU^T x) =$$

$$\|x\|^2 - (x, UU^T x) = \|x\|^2 - \text{trace}(U^T x x^T U):$$

Այստեղ մատրիցի հետք (trace) կոչվում է նրա գլխավոր անկյունագծի էլեմենտների գումարը: Այսպիսով (10.1) խնդիրը համարժեք է հետևյալին՝

$$\arg \max_{U^{d,n}: U^T U = I} \text{trace}\left(U^T \sum_{i=1}^m x_i x_i^T U\right):$$

Նշանակենք $A = \sum_{i=1}^m x_i x_i^T$: Այս մատրիցը սիմետրիկ է, հետևաբար նա ունի սպեկտրիալ վերլուծություն՝ $A = V D V^T$, որտեղ D -ն անկյունագծային մատրից է, որի անկյունագծում A մատրիցի սեփական արժեքներն են, իսկ V մատրիցի սյուները համապատասխան սեփական վեկտորն են ($V^T V = V V^T = I$): Քանի որ A մատրիցը դրական կիսաորոշյալ է, ապա (ընդհանրությունը չիսախտելով) ենթադրենք, որ

$$D_{1,1} \geq D_{2,2} \geq \dots \geq D_{d,d} \geq 0:$$

Թեորեմ 11.2. *PCA (11.1) խնդրի լուծումը U , $W = U^T$ մատրիցներն են, որտեղ U մատրիցի սյուները A մատրիցի n ամենամեծ սեփական արժեքներին համապատասխանող սեփական վեկտորներն են:*

Ապացույց: Դիցուք $A = VDV^T$, և $U^{d,n}$ -ն մատրից է օրթոգոնալ սյուներով: Նշանակենք $B = V^T U$: Այդ դեպքում $VB = VV^T U = U$: Հետևաբար՝

$$U^T AU = B^T V^T V D V^T V B = B^T D B:$$

Այստեղից կունենանք՝

$$\text{trace}(U^T AU) = \sum_{j=1}^d D_{j,j} \sum_{i=1}^n B_{j,i}^2:$$

Նկատենք նաև, որ

$$B^T B = U^T V V^T U = U^T U = I:$$

Հետևաբար B մատրիցի սյուները նույնպես օրթոգոնալ են: Ուստի՝

$$\sum_{j=1}^d \sum_{i=1}^n B_{j,i}^2 = n:$$

Նշանակենք $\tilde{B}$ -ով $(d \times d)$ -չափանի մատրից, որի առաջին n սյուները B -ի սյուներն են և $\tilde{B}^T \tilde{B} = I$: Այդ դեպքում

$$\sum_{i=1}^d \tilde{B}_{j,i}^2 = 1, \forall j:$$

Հետևաբար՝ $\sum_{i=1}^n B_{j,i}^2 \leq 1$: Ուրեմն՝

$$\text{trace}(U^T AU) = \max_{\beta \in [0,1]^d: \|\beta\|_1 \leq n} \sum_{j=1}^d D_{j,j} \beta_j:$$

Այստեղից հետևում է, որ $\text{trace}(U^T AU) \leq \sum_{j=1}^n D_{j,j}$: Իսկ հավասարություն տեղի կունենա այն դեպքում, երբ U մատրիցի սյուները A մատրիցի առաջին n սեփական արժեքներին համապատասխանող սեփական վեկտորներն են:

Այժմ ցույց տանք, որ վերևում քննարկված *PCA* ալգորիթմը ունի նաև հետևյալ համարժեք մեկնաբանություններ: Ըստ էության որոնվում է մի հարթություն, որի վրա պրոյկտվում են տվյալները՝ ապահովելով տվյալների նվազագույն «աղավաղված» պատկեր:

Քննարկենք այդ պրոյեկցիոն հարթության կառուցման սկզբունքները:

Դիցուք ունենք որևէ $S = \{x_1, x_2, \dots, x_m\}$ հավաքածու: Դիտարկենք հետևյալ երկու խնդիրները:

- Գտնել $n < d$ -չափանի մի M_n գծային բազմաձևություն, որ բոլոր L_2 մեծությունների հեռավորությունների քառակուսիների գումարը այդ բազմաձևությունից

լինի մինիմալ: Այսինքն՝

$$Q(M_n) = \sum_{i=1}^m d(x_i, M_n)^2 \rightarrow \min:$$

- Գտնել $n < d$ -չափանի գծային բազմաձևություն այնպես, որ կետերի պրոյեկցիաների ընդհանուր դիպերսիան լինի մաքսիմալ:

Այժմ ցույց տանք, որ այս երկու խնդիրների լուծումները նույն են և հանգում են կովարիացիոն A մատրիցի սեփական արժեքների որոնման խնդրին:

Դիտարկենք առաջին խնդիրը: n -չափանի գծային բազմազանության էլեմենտները ունեն հետևյալ տեսքը՝

$$u = s_0 + \sum_{i=1}^n \alpha_i s_i,$$

որտեղ s_i , $i = 1, 2, \dots, n$ օրթոնորմավորված վեկտորներ են: Եթե $d(x, M_n) = \|x - u\|$ և $u \in M_n$, ապա $x - u \perp s_i$, $i = 1, 2, \dots, n$: Այստեղից, հաշվի առնելով, որ $\{s_i\}$ համակարգը օրթոնորմալ է, կստանանք

$$0 = (s_j, x - u) = (s_j, x - s_0) - \alpha_j \quad j = 1, 2, \dots, n:$$

Այստեղից հետևում է, որ

$$\alpha_j = (s_j, x - s_0), \quad j = 1, 2, \dots, n,$$

և հետևաբար՝

$$d^2(x, M_n) = \|x - s_0\|^2 - \sum_{j=1}^n (s_j, x - s_0)^2:$$

Այսպիսով, հիմնական խնդիրը կարելի է ձևակերպել որպես քառակուսային ծրագրավորման խնդիր. պետք է գտնել $\{s_i\}_{i=1}^n$ վեկտորների այնպիսի համախումբ, որ

$$Q(s_0, s_1, \dots, s_n) = \sum_{i=1}^m \left(\|x_i - s_0\|^2 - \sum_{j=1}^n (s_j, x_i - s_0)^2 \right) \rightarrow \min,$$

$$\|s_j\| = 1, \quad j = 1, 2, \dots, n, \quad (s_i, s_j) = 0, \quad i \neq j, \quad i, j = 1, 2, \dots, n:$$

Քանի որ նպատակային ֆունկցիոնալի s_0 փոփոխականի վրա սահմանափակում չկա, ապա դրա գրադիենտը մինիմումի կետում պետք է լինի զրո: Այսինքն՝

$$Q'_{s_0} = -2 \sum_{i=1}^m (x_i - s_0) + 2 \sum_{j=1}^n \sum_{i=1}^m (x_i - s_0) s_j = 0:$$

Այստեղից կստանանք՝

$$\sum_{i=1}^m (x_i - s_0) = \sum_{j=1}^n s_j (s_j, \sum_{i=1}^m (x_i - s_0)):$$

Այս հավասարությունը տեղի ունի, եթե օրինակ՝

$$\sum_{i=1}^m (x_i - s_0) = 0:$$

Հետևաբար՝

$$s_0 = \frac{1}{m} \sum_{i=1}^m x_i:$$

Նշանակենք

$$v_i = x_i - s_0, \quad i = 1, 2, \dots, m:$$

Այդ դեպքում հիմնական խնդիրը կունենա հետևյալ տեսքը՝

$$\tilde{Q}(s_1, \dots, s_n) = \sum_{j=1}^n \|V s_j\|^2 \rightarrow \max,$$

$$\|s_j\| = 1, \quad (s_i, s_j) = 0, \quad i \neq j, \quad i, j = 1, 2, \dots, n,$$

որտեղ

$$V = (v_1, v_2, \dots, v_m)^T,$$

$v_i, i = 1, \dots, m$ - վեկտոր սյուներ են:

Խնդիրը լուծենք Լագրանժի անորոշ գործակիցների մեթոդով: Կազմենք Լագրանժի ֆունկցիան՝

$$L(s_1, \dots, s_n, \lambda_1, \dots, \lambda_n) = \tilde{Q}(s_1, \dots, s_n) - \sum_{j=1}^n \lambda_j (\|s_j\|^2 - 1):$$

Հաշվելով Լագրանժի ֆունկցիայի գրադիենտը ըստ $s_j, \quad j = 1, \dots, n$ փոփոխականների՝ կստանանք

$$L'_{s_j} = 2 \sum_{i=1}^m (s_j, v_i) v_i - 2\lambda_j s_j = 2(V^T V s_j - \lambda_j s_j) = 0, \quad j = 1, 2, \dots, n:$$

Այստեղից հետևում է, որ $\{s_j\}_{j=1}^n$ համակարգը կորելացիոն $V^T V$ մատրիցի սեփական վեկտորներն են: Հայտնի է, որ կորելացիոն մատրիցը սիմետրիկ է և դրական կիսաորոշյալ: Հետևաբար $\tilde{Q}$ ֆունկցիոնալը կունենա հետևյալ տեսքը՝

$$\begin{aligned} \tilde{Q}(s_1, s_2, \dots, s_n) &= \sum_{j=1}^n \|V s_j\|^2 = \sum_{j=1}^n (V s_j)^T V s_j = \sum_{j=1}^n ((V^T V s_j)^T s_j = \\ &= \sum_{i=1}^n \lambda_j s_j^T s_j = \sum_{i=1}^n \lambda_j: \end{aligned}$$

Հետևաբար Q ֆունկցիան կընդունի մաքսիմալ արժեք $V^T V$ մատրիցի առաջին n ամենամեծ սեփական արժեքներին համապատասխանող սեփական վեկտորների $\{s_j\}_{j=1}^n, \quad j = 1, 2, \dots, n$ համակարգի վրա:

Այսպիսով կարող ենք ձևակերպել հետևյալ թեորեմը.

Թեորեմ 11.3. Կամայական $S = \{x_1, x_2, \dots, x_m\}$ համակարգի համար օպտիմալ գծային բազմաձևության էլեմենտները ունեն հետևյալ տեսքը՝

$$u = s_0 + \sum_{i=1}^m \alpha_i x_i,$$

որտեղ

$$s_0 = \frac{1}{m} \sum_{i=1}^m x_i,$$

իսկ $s_j, j = 1, 2, \dots, n$ վեկտորները՝ $V^T V$ կորելացիոն մատրիցի առաջին n մաքսիմալ սեփական արժեքներին համապատասխանող սեփական վեկտորներն են:

Այժմ ցույց տանք, որ եթե համակարգը կենտրոնացված է, այսինքն՝ երբ նրա միջինը հավասար է զրոյի ($\frac{1}{m} \sum_{i=1}^m x_i = 0$), ապա վերը նշված գծային բազմաձևությունը այն ենթատարածությունն է, որի վրա կետերի պրոյեկցիաների ընդհանուր դիսպերսիան (ցրվածությունը) պետք է լինի մաքսիմալ: Այսինքն՝ պետք է գտնել

$$M_n(s_1, s_2, \dots, s_n), \quad s_i \perp s_j, \quad i \neq j, \quad \|s_j\| = 1,$$

ենթատարածություն այնպիսին, որ

$$\sum_{i=1}^m \|\Pi_{M_n} x_i\|^2 = \sum_{i=1}^m \left\| \sum_{j=1}^n (x_j, s_j) s_j \right\|^2 \rightarrow \max:$$

Այս խնդիրը համարժեք է առաջինին, քանի որ Պյութագորասի թեորեմի համաձայն ունենք

$$\sum_{j=1}^n \|(x_i, s_j) s_j\|^2 = \|\Pi_{M_n} x_i\|^2 = \|x_i\|^2 - \|\text{Ort}_{M_n} x_i\|^2 = \|x_i\|^2 - \|x_i - \sum_{j=1}^n (x_i, s_j) s_j\|^2:$$

Քանի որ աջ մասի առաջին գումարելին կախված չէ s_j վեկտորից, ապա

$$\sum_{i=1}^m \|\Pi_{M_n} x_i\|^2 \rightarrow \max$$

խնդիրը համարժեք է $Q(M_n) \rightarrow \min$ սկզբնական խնդրին: Այսպիսով խնդիրը հանգում է կովարիացիոն A մատրիցի սեփական արժեքների որոնմանը:

Մեկնաբանենք այս թեորեմը օրինակի միջոցով: Դիցուք ունենք երկու դաս՝

$$A = \{(x_1 = \sqrt{3/2}, 0), x_2 = (0, 0), x_3 = (1, 1)\}, \quad y_1 = 1\},$$

$$B = \{x_4 = (-\sqrt{3/2}, 0), x_5 = (-1, -1), y = -1\} \quad (\text{տե՛ս նկ. 11.1}):$$

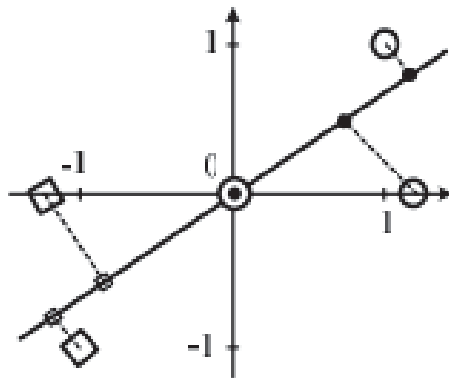
Կովարիացիոն $A = V^T V$ մատրիցը կունենա հետևյալ տեսքը՝

$$\begin{pmatrix} 5 & 2 \\ 2 & 5 \end{pmatrix}:$$

Հետևաբար դրա մաքսիմալ սեփական արժեքը՝ $\lambda = 6$: Ուստի դրան համապատասխանող սեփական վեկտորը կլինի $e_1 = (2, 1)$ -ը: Այստեղից հետևում է, որ օպտիմալ M_1 գիծը կլինի $y = x/2$ ուղիղը: Ուստի կունենանք

$$x'_1 = \frac{\sqrt{30}}{5}, x'_2 = 0, x'_3 = 3\sqrt{20}/10, x'_4 = -\sqrt{30}/5, x'_5 = -3\sqrt{20}/10:$$

(տե՛ս նկ. 11.1): Այստեղ x'_i , $i \in [1 : 5]$ վեկտորները x_i , $i \in [1 : 5]$ օբյեկտների «սեղմված» պատկերներն են միաչափ «գլխավոր» կոմպոնենտի վրա: Այս դեպքում այդ պատկերները թվեր են:



Նկ. 11.1: Գլխավոր էլեմենտների մեթոդը

11.2 Ֆիշերի տարբերակիչը (դիսկրիմինատը)

Գլխավոր էլեմենտների մեթոդով որոնվող գծային բազմաձևության վրա հնարավոր է, որ դասերը իրարից խիստ չանջատվեն, այլ իրար վրա ընկնեն ինչ-որ մասով: Բերենք այդպիսի մի օրինակ: Դիցուք

$$A = \{(0, 1), (0, 2), (1, 0)\}, B = \{(-1, 0), (0, -3)\}:$$

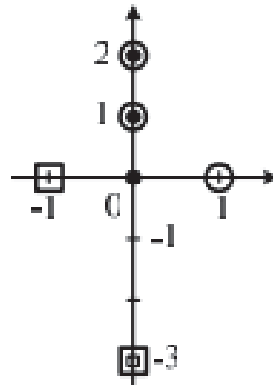
Հաշվենք համապատասխան սեփական վեկտորները.

$$s_0 = \frac{1}{5} \sum_{i=1}^5 5x_i = 0:$$

Գտնենք $V^T V$ կորելացիոն մատրիցի սեփական արժեքները: Ունենք՝

$$\begin{pmatrix} 2 & 0 \\ 0 & 12 \end{pmatrix}:$$

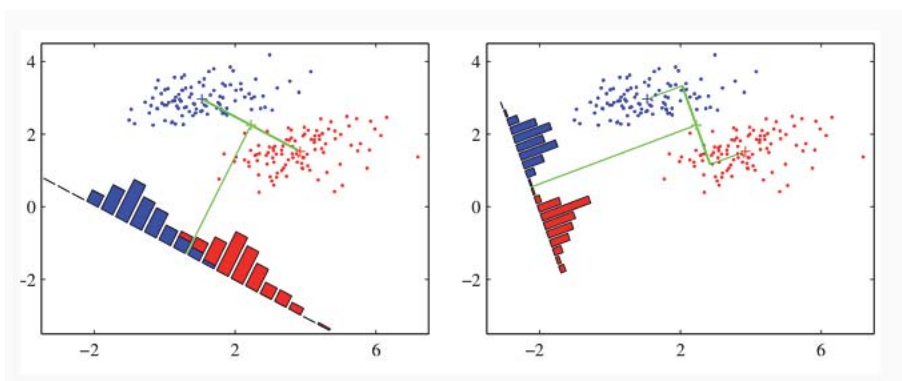
Այսպիսով, մաքսիմալ սեփական արժեքը հավասար է 12-ի: Դրան համապատասխանող սեփական վեկտորն է $(0, 1)$ -ը: Այստեղից հետևում է, որ օպտիմալ M_1 գծային բազմաձեռությունը կլինի oy առանցքը: Դասերի պրոյեկցիաները կունենան ընդհանուր 0 կետը (նկ. 11.2):



Նկ. 11.2: Ֆիշերի տարբերակիչը A և B դասերի համար

Այս օրինակում հեշտ է նկատել, որ պրոյեկցիաների դիսպերսիան մեծ է յուրաքանչյուր դասում, և դասերի միջինները համընկնում են: Այժմ քննարկենք մի հայտանիշ (կրիտերիա), որում հաշվի են առնվում նաև դասերի դիսպերսիաները: Փնտրում ենք մի ուղիղ (դիցուք w ($\|w\| = 1$)-ը նրա ուղղորդ վեկտորն է), որի վրա պրոյեկտվում են հավաքածուի վեկտորները: Այդ ուղիղը պետք է լինի այնպիսին, որ դասերի պրոյեկցիաների միջինների հեռավորությունը լինի մաքսիմալ, իսկ պրոյեկցիաների ընդհանուր ցրվածությունը՝ մինիմալ: Այդ պայմաններին բավարարող ուղիղը կոչվում է Ֆիշերի տարբերակիչ: Եթե x վեկտորը պրոյեկտվում է w վեկտորի ուղղության վրա, ապա պրոյեկցիան հաշվում են $x' = w^T x = (w, x)$ բանաձևով:

Նկար 11.3-ում դասերը պրոյեկտվում են երկու տարբեր ուղիղների վրա: Ինչպես երևում է երկրորդ նկարում, դասերի պրոյեկցիաները ավելի «լավ» են անջատվում, քան առաջինում:



Նկ. 11.3: Կետերի պրոյեկցիան տարբեր ուղիղների վրա

Քննարկենք երկու դասի դեպքը: Օպտիմալ ուղղի որոնման համար Ֆիշերը առաջարկեց որպես կրիտերիա օգտագործել հետևյալ ֆունկցիան՝

$$f(w) = \frac{\|m'_1 - m'_2\|^2}{s_1'^2 + s_2'^2},$$

որտեղ

$$m'_j = \frac{1}{|A_j|} \sum_{x \in A_j} x',$$

$$s_j'^2 = \sum_{x \in A_j} (x' - m'_j)^2, \quad j = 1, 2:$$

Այստեղ m'_j , $j = 1, 2$ պրոյեկտված դասերի միջիններն են, իսկ s'_j , $j = 1, 2$ ՝ դիսպերսիաները ուղղի վրա: Ձևափոխենք այս հայտանիշը: Միջին արժեքների համար ունենք՝

$$m'_j = \frac{1}{|A_j|} \sum_{x' \in A_j} x' = \frac{1}{|A_j|} \sum_{x \in A_j} w^T x = w^T m_j, \quad j = 1, 2:$$

Իսկ պրոյեկցիաների միջինների հեռավորությունը հաշվում են հետևյալ կերպ՝

$$\|m'_1 - m'_2\|^2 = |w^T(m_1 - m_2)|^2 = w^T(m_1 - m_2)(m_1 - m_2)^T w = w^T S_m w,$$

որտեղ

$$S_m = (m_1 - m_2)(m_1 - m_2)^T:$$

Իսկ պրոյեկտված դասերի դիսպերսիաների համար կունենանք՝

$$s_j'^2 = \sum_{x \in A_j} (x' - m'_j)^2 = \sum_{x \in A_j} (w^T(x - m_j)(x - m_j)^T w) = w^T S_j w,$$

որտեղ

$$S_j = \sum_{x \in A_j} (x - m_j)(x - m_j)^T, \quad j = 1, 2:$$

Դիցուք $S = S_1 + S_2$: Այդ դեպքում

$$(s'_1)^2 + (s'_2)^2 = w^T S w,$$

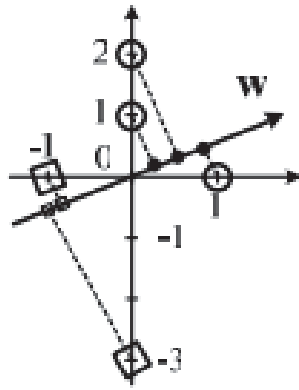
իսկ նպատակային ֆունկցիան կրնդունի հետևյալ տեսքը՝

$$f(w) = \frac{w^T S_m w}{w^T S w}:$$

Կարելի է ցույց տալ, որ այս ֆունկցիոնալը հասնում է իր մաքսիմալ արժեքին այնպիսի w կետում, որը բավարարում է հետևյալ հավասարմանը՝

$$S_m w = \lambda S w : \quad (11.2)$$

Այս հավասարման լուծումը $S^{-1} S_m$ մատրիցի սեփական վեկտորն է, որովհետև $S_m =$



Նկ. 11.4: Ֆիշերի տարբերակիչ

$k(m_1 - m_2)$, որտեղ $k = (m_1 - m_2)^T w$: Հետևաբար որպես (11.2) հավասարման լուծում կարելի է վերցնել

$$w = S^{-1}(m_1 - m_2)$$

վեկտորը: Օրինակի վրա մեկնաբանենք այս պնդումները: Դիցուք

$$A_1 = \{x_1 = (0, 2), x_2 = (0, 1), x_3 = (1, 0)\}, A_2 = \{x_4 = (-1, 0), x_5 = (0, -3)\}:$$

Այդ դեպքում

$$m_1 = (1/3, 1), m_2 = (-1/2, -3/2),$$

$$S_1 = \frac{1}{3} \begin{pmatrix} 2 & -3 \\ -3 & 6 \end{pmatrix},$$

$$S_2 = \frac{1}{2} \begin{pmatrix} 1 & -3 \\ -3 & 9 \end{pmatrix}:$$

Հետևաբար

$$S = S_1 + S_2 = \frac{1}{6} \begin{pmatrix} 7 & -15 \\ -15 & 39 \end{pmatrix}:$$

Այսպիսով կստանանք

$$w = S^{-1}(m_1 - m_2) = \frac{5}{4}(7, 3):$$

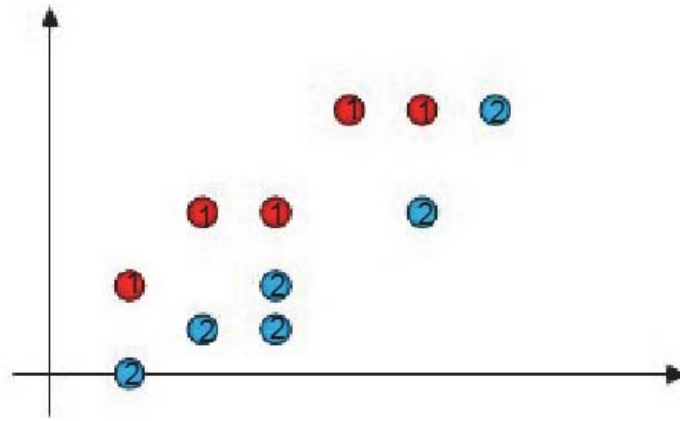
Ուստի՝

$$x'_1 = 15/2, x'_2 = 15/4, x'_3 = 35/4, x'_4 = -35/4, x'_5 = -45/4:$$

Նկար 11.4-ից երևում է, որ դասերի էլեմենտների պրոյեկցիաները ավելի լավ են անջատված և կենտրոնացված:

Այժմ դիտարկենք ևս մեկ օրինակ: Ունենք երկու դասեր

$$A = \{(1, 2), (2, 3), (3, 3), (4, 5), (5, 5)\}, B = \{(1, 0), (2, 1), (3, 1), (5, 3), (6, 5)\}:$$



Նկ. 11.5: Տվյալների երկու դասեր

Գտնենք դասերի կենտրոնները՝

$$m_1 = (3, 3.6), \quad m_2 = (3, 3.2):$$

Հաշվելով ցրվածության մատրիցները յուրաքանչյուր դասի համար՝ կստանանք՝

$$S_1 = \begin{pmatrix} 10 & 8 \\ 8 & 7.2 \end{pmatrix}$$

$$S_2 = \begin{pmatrix} 17.7 & 16 \\ 16 & 16 \end{pmatrix}:$$

Այստեղից կստանանք

$$S = S_1 + S_2 = \begin{pmatrix} 27.3 & 14 \\ 24 & 23.2 \end{pmatrix}:$$

Այսպիսով

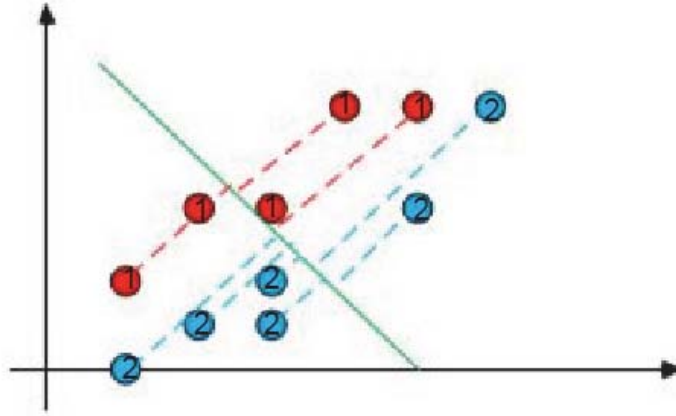
$$w = S^{-1}(m_1 - m_2) = (-0.79, 0.89),$$

իսկ դասերի պրոյեկցիաները այդ ուղղության վրա կլինեն

$$A' = wA^T = (-0.67, 0.73) \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 3 & 3 & 5 & 5 \end{pmatrix} = (0.81, \dots, 0.4),$$

$$B' = wB^T = (-0.65, \dots, -0.25):$$

Այստեղ A' և B' վեկտորների կոորդինատները A և B դասերի համապատասխան օբյեկտների պրոյեկցիաներն են w ուղղության վրա: Նկար 11.5-ում պատկերված են այս օբիեկտների օբյեկտները, իսկ նկար 11.6-ում՝ այն ուղիղը, որի վրա կատարվում է պրոյեկտում:



Նկ. 11.6: Ֆիշերի տարբերակիչով կառուցված ուղիղը A և B դասերի համար

11.3 Չափողականության կրճատում պատահական գծային ձևափոխությամբ

Այս պարագրաֆում կատարվում է վեկտորների (օբյեկտների) չափողականության կրճատում մատրիցների օգնությամբ, որոնց էլեմենտները միևնույն պարամետրով նորմալ բաշխված պատահական մեծություններ են: Պարզվում է, որ այս դեպքերում ինֆորմացիոն կորուստ համարյա չի կատարվում:

Դիցուք ունենք x_1, x_2 վեկտորներ: Կասենք, որ W մատրիցը քիչ է «աղճատում» $\|x_1 - x_2\|$ հեռավորությունը, եթե

$$\frac{\|Wx_1 - Wx_2\|}{\|x_1 - x_2\|}$$

հարաբերությունը «մոտիկ» է մեկին: Սա նշանակում է՝ W մատրիցը շատ չի փոխում այդ վեկտորների էվկլիդեսյան հեռավորությունը:

Լեմմա 11.1. Դիցուք $x \in \mathbb{R}^d$, $W^{n,d}$ մատրիցի է, և $w_{i,j}$ էլեմենտները նորմալ բաշխված անկախ պատահական մեծություններ են: Այդ դեպքում կամայական $\epsilon \in (0, 3)$ թվի համար տեղի ունի

$$P\left[\left|\frac{\|1/\sqrt{n}Wx\|^2}{\|x\|^2} - 1\right| > \epsilon\right] \leq 2e^{-\epsilon^2 n/6}$$

անհավասարությունը:

Նախքան ապացույցը մեզ անհրաժեշտ է հետևյալ պնդումը: Դիցուք $X_1, X_2, \dots, X_k$ նորմալ բաշխված անկախ պատահական մեծություններ են: Այսինքն՝ $X_i \sim N(0, 1)$: Նշանակենք $Z = X_1^2 + \dots + X_k^2$: Ակնհայտ է, որ $E[X_i^2] = 1$, $E[Z] = k$:

Լեմմա 11.2. Ցանկացած $\epsilon \in (0, 3)$ թվի համար տեղի ունի հետևյալ անհավասարությունը՝

$$P\left[(1 - \epsilon)k \leq Z \leq (1 + \epsilon)k\right] \geq 1 - 2e^{-\epsilon^2 k/6}:$$

Ապացույց: Ակնհայտ է, որ $a \geq 0$ -ի դեպքում տեղի ունի

$$e^{-a} \leq 1 - a + \frac{a^2}{2}$$

անհավասարությունը: Հետևաբար $\lambda > 0$ դեպքում

$$E[e^{-\lambda X_1^2}] \leq 1 - \lambda E[X_1^2] + \frac{\lambda^2}{2} E[X_1^4]:$$

Հայտնի է, որ $E[X_1^2] = 1$, $E[X_1^4] = 3$, ուստի՝

$$E[e^{-\lambda X_1^2}] \leq 1 - \lambda + \left(\frac{3}{2}\right)\lambda^2 \leq e^{-\lambda + (3/2)\lambda^2}:$$

Այստեղից, կիրառելով Չեբիշևի անհավասարությունը, կստանանք՝

$$\begin{aligned} P[-Z \geq -(1 - \epsilon)k] &= P[e^{-\lambda Z} \geq e^{-(1-\epsilon)\lambda k}] \\ &\leq e^{(1-\epsilon)k\lambda} E[e^{-\lambda Z}] = e^{(1-\epsilon)k\lambda} \left(E[e^{-\lambda X_1^2}]\right)^k \\ &\leq e^{(1-\epsilon)k\lambda} e^{-\lambda k + (3/2)k\lambda^2} = e^{-\epsilon k\lambda + (3/2)\lambda^2 k}: \end{aligned}$$

Այս անհավասարության մեջ վերցնելով $\lambda = \epsilon/3$ ՝ կստանանք՝

$$P[Z \leq (1 - \epsilon)k] \leq e^{-\epsilon^2 k/6}:$$

Անհավասարության երկրորդ մասը ապացուցելու համար օգտվենք հետևյալ հայտնի անհավասարությունից՝

$$E[e^{\lambda Z^2}] = (1 - 2\lambda)^{-k/2}, \quad \lambda < 1/2:$$

Հաշվի առնելով այս անհավասարությունը՝ կունենանք՝

$$\begin{aligned} P[Z \geq (1 + \epsilon)k] &= P[e^{\lambda Z} \geq e^{(1+\epsilon)k\lambda}] \\ &\leq e^{-(1+\epsilon)k\lambda} E[e^{\lambda Z}] = e^{-(1+\epsilon)k\lambda} (1 - 2\lambda)^{k/2} \\ &\leq e^{-(1+\epsilon)k\lambda} e^{k\lambda} = e^{-\epsilon k\lambda}: \end{aligned}$$

Տեղադրելով այստեղ $\lambda = \epsilon/6$ ՝ կստանանք՝

$$P[Z \geq (1 + \epsilon)k] \leq e^{-\epsilon^2 k/6}:$$

Լեմման ապացուցվեց:

Լեմմա 11.3. Դիցուք ունենք $X_1, X_2, \dots, X_n$ փոխադարձաբար անկախ, նորմալ

բաշխված պատահական մեծություններ՝ $X_i \sim N(\theta_i, \sigma_i^2)$, $i = 1, 2, \dots, n$: Այդ դեպքում

$$X = \sum_{i=1}^n X_i \sim N\left(\sum_{i=1}^n \theta_i, \sum_{i=1}^n \sigma_i^2\right):$$

Ապացույց: Ապացույցը կատարենք $n = 2$, $X = X_1 + X_2$ -ի դեպքում: Հայտնի է, որ

$$f(x) = \int_{-\infty}^{+\infty} f_2(t)f_1(x-t)dt,$$

որտեղ f_1, f_2 -ը համապատասխանաբար X_1 և X_2 պատահական մեծությունների խտության ֆունկցիաներն են, իսկ f -ը՝ դրանց գումարի խտության: Հետևաբար՝

$$f(x) = \frac{1}{2\pi\sigma_1\sigma_2} \int_{-\infty}^{+\infty} \exp\left\{-\frac{1}{2}\left(\frac{t-\theta_1}{\sigma_1}\right)^2 - \frac{1}{2}\left(\frac{x-t-\theta_2}{\sigma_2}\right)^2\right\} dt:$$

Դժվար չէ համոզվել, որ

$$f(x) = \frac{1}{2\pi(\sigma_1^2 + \sigma_2^2)} \exp\left\{-\frac{(x-\theta_1-\theta_2)^2}{2(\sigma_1^2 + \sigma_2^2)}\right\} \quad (\text{տես [29], էջ 95)}:$$

Այժմ անցնենք լեմմա 11.1-ի ապացույցին: Ընդհանրությունը չխախտելով՝ ենթադրենք, որ $\|x\| = 1$: Հետևաբար լեմմայի անհավասարությունը կարող ենք գրել հետևյալ կերպ՝

$$P\left[(1-\epsilon)n \leq \|Wx\|^2 \leq (1+\epsilon)n\right] \geq 1 - 2e^{-\epsilon^2 n/6}:$$

Դիցուք w_i -ին W մատրիցի i -րդ տողն է: $(w_i, x) = w_i^1 x_1 + \dots + w_i^d x_d$ պատահական մեծությունը նորմալ բաշխված պատահական մեծությունների գումար է: Համաձայն նախորդ լեմմայի՝ դրա մաթեմատիկական սպասումը հավասար է զրոյի, իսկ դիսպերսիան՝ $\sum_i x_i^2 = \|x\|^2 = 1$: Օգտագործելով լեմմա 11.2-ը՝ կստանանք մեր պնդման եզրակացությունը:

Թեորեմ 11.4. (Չոնսոն): Դիցուք Q -ն կետերի վերջավոր բազմություն է $\mathbb{R}^d$ տարածությունից: Ենթադրենք նաև, որ $\delta \in (0, 1)$, n թվերը այնպիսին են, որ

$$\epsilon \equiv \sqrt{\frac{6 \ln(2|Q|/\delta)}{n}} \leq 3:$$

Ենթադրենք նաև $W \in \mathbb{R}^{n,d}$ մատրիցի էլեմենտները նորմալ բաշխում ունեն՝ զրոյական մաթեմատիկական սպասումով և $1/n$ դիսպերսիայով: Այդ դեպքում առնվազն $1 - \delta$ հավանականությամբ տեղի ունի

$$\sup_{x \in Q} \left| \frac{\|Wx\|^2}{\|x\|^2} - 1 \right| < \epsilon$$

անհավասարությունը:

Ապացույց: Լեմմա 11.1-ի համաձայն՝ $\epsilon \in (0, 3)$ -ի դեպքում ունենք

$$P\left[\sup_{x \in Q} \left| \frac{\|Wx\|^2}{\|x\|^2} - 1 \right| > \epsilon\right] \leq 2|Q|e^{-\epsilon^2 n/6}:$$

Անհավասարության աջ մասը նշանակելով δ -ով՝ կստանանք՝

$$\epsilon = \sqrt{\frac{6 \ln(2|Q|/\delta)}{n}}:$$

Նկատենք, որ այս գնահատականի մեջ x վեկտորի d չափողականությունը չի մասնակցում:

11.4 SNE և UMAP ալգորիթմների մասին

Դիցուք խնդիր է դրվում մեծ չափողականություն ունեցող տվյալների $\{x_i | x_i \in X\}$ բազմությունը ներդնել ավելի փոքր չափողականությամբ տարածության մեջ: Դիցուք այդ կետերի պատկերները այդ նոր տարածությունում հետևյալներն են՝ $\{y_i | y_i \in Y\}$: Այդ արտապատկերումը կատարվում է հետևյալ սկզբունքներով: Յուրաքանչյուր x_i կետ $P(i, j)$ հավանականությամբ որպես «հարևան» է ընտրում x_j օբյեկտը: Այդ հավանականությունը տրվում է համապատասխան գաուսյան բաշխմամբ՝

$$p(j, i) = \frac{\exp(-\frac{\|x_i - x_j\|^2}{2\sigma_i^2})}{\sum_{k \neq i} \exp(-\frac{\|x_i - x_k\|^2}{2\sigma_i^2})}:$$

Այս բանաձևից հետևում է, որ որքան հեռու է x_j օբյեկտը x_i -ից, այնքան $P(i, j)$ հավանականությունը փոքր է: Նման ձևով սահմանվում է «հավանականային իմաստով նմանություն» y_i օբյեկտների միջև՝

$$q(j, i) = \frac{\exp(-\|y_i - y_j\|^2)}{\sum_{k \neq i} \exp(-\|y_i - y_k\|^2)}:$$

Այս բանաձևը նախորդից տարբերվում է նրանով, որ այստեղ բոլոր դիսպերսիաները վերցված են հաստատուն՝ $\sigma_i = \frac{1}{\sqrt{2}}$: Որպեսզի կատարվի լավ ներդրում, այս հավանականությունների բաշխման տարբերությունը պետք է լինի մինիմալ: Այդ տարբերության չափանիշը սովորաբար արտահայտվում է Քուլբեր-Լեյբլի դիվերգենցիայի միջոցով՝

$$D_{KL}(P||Q) = \sum_j p_j \ln \frac{p_j}{q_j}:$$

Բերենք այս ֆունկցիայի որոշ հատկություններ: Յույց տանք, որ այն սիմետրիկ ֆունկցիա չէ իր արգումենտների նկատմամբ: Այսինքն՝ $D_{KL}(P||Q) \neq D_{KL}(Q||P)$: Այս փաստը հաստատող թվային օրինակը ներկայացված է ներքևի աղյուսակում:

Հավանականային բաշխումների աղյուսակ			
X	0	1	2
P բաշխում	9/25	12/25	4/25
Q բաշխում	1/3	1/3	1/3

Իրոք, հաշվենք $D_{KL}(P||Q)$ և $D_{KL}(Q||P)$ արժեքները և համեմատենք իրար հետ:

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) \ln \frac{P(x)}{Q(x)} =$$

$$= \frac{9}{25} \ln\left(\frac{9/25}{1/3}\right) + \frac{12}{25} \ln\left(\frac{12/25}{1/3}\right) + \frac{4}{25} \ln\left(\frac{4/25}{1/3}\right) = 0.122:$$

Անալոգիայով ստանում ենք՝

$$D_{KL}(Q||P) = 0.139:$$

Ինչպես տեսնում ենք, այս ֆունկցիան սիմետրիկ չէ իր արգումենտների նկատմամբ:

Այժմ ցույց տանք, որ կամայական $P = \{p_1, p_2, \dots, p_n\}$ և $Q = \{q_1, q_2, \dots, q_n\}$ հավանականային բաշխումների համար $D_{KL}(P||Q) \geq 0$: Իրոք, քանի որ $\ln x < 1 - x$, $\forall x > 0$, ապա

$$-\sum_{i=1}^n p_i \ln \frac{q_i}{p_i} \geq$$

$$-\sum_{i=1}^n p_i \left(\frac{q_i}{p_i} - 1\right) = -\sum_{i=1}^n q_i + \sum_{i=1}^n p_i = 0:$$

Ակնհայտ է նաև, որ $D_{KL}(P||P) = 0$:

Այսպիսով, կարելի է եզրակացնել, որ $D_{KL}(P||Q)$ ֆունկցիան, ընդհանրապես ասած, մետրիկա չէ:

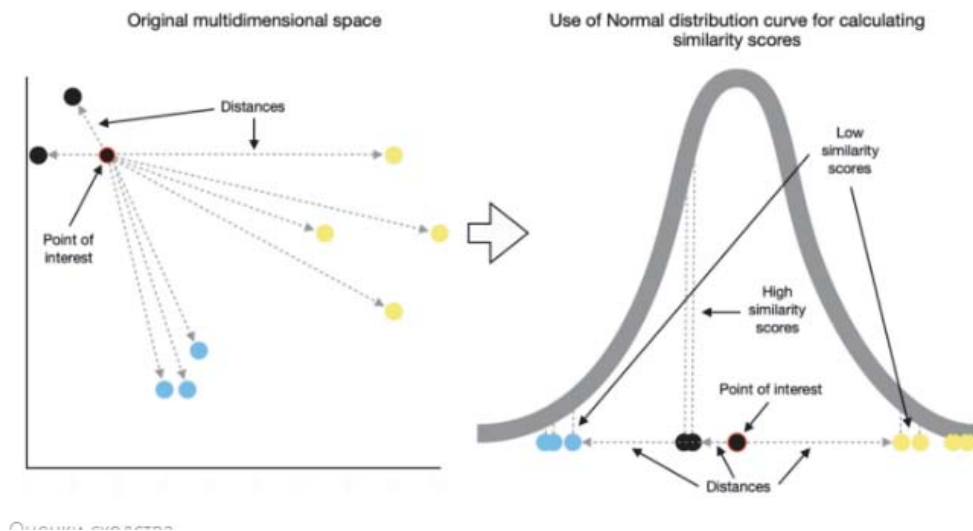
$D_{KL}(P||Q)$ ֆունկցիան կարելի է ներկայացնել հետևյալ տեսքով՝

$$D_{KL}(P||Q) = \sum_j p_j \ln p_j - \sum_j p_j \ln q_j = H(P) - H(P, Q),$$

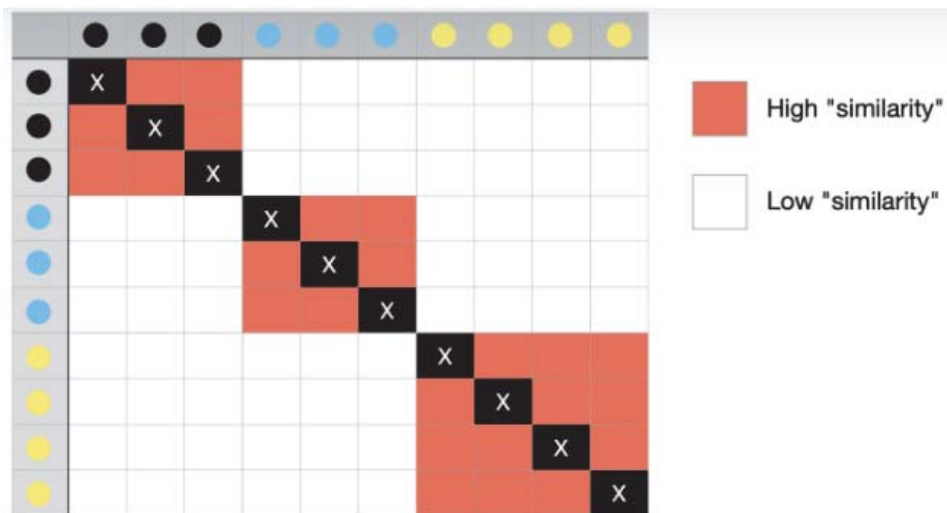
որտեղ $H(P) = \sum_j p_j \ln p_j$ Էնտրոպիայի բանաձևն է, իսկ $H(P, Q) = \sum_j p_j \ln q_j$ արտահայտությունը կոչվում է պայմանական կամ խաչաձև Էնտրոպիա:

Տվյալ դեպքում մենք ունենք $|X|$ քանակով բաշխումներ: Մենք պետք է միմյանիցացնենք բոլոր դիվերգենցիաների գումարը՝

$$C(y_i | y_i \in Y) = \sum_i \sum_j P(j, i) \ln \frac{P(j, i)}{Q(j, i)}:$$



Նկ. 11.7: Օբյեկտների նմանության գնահատում



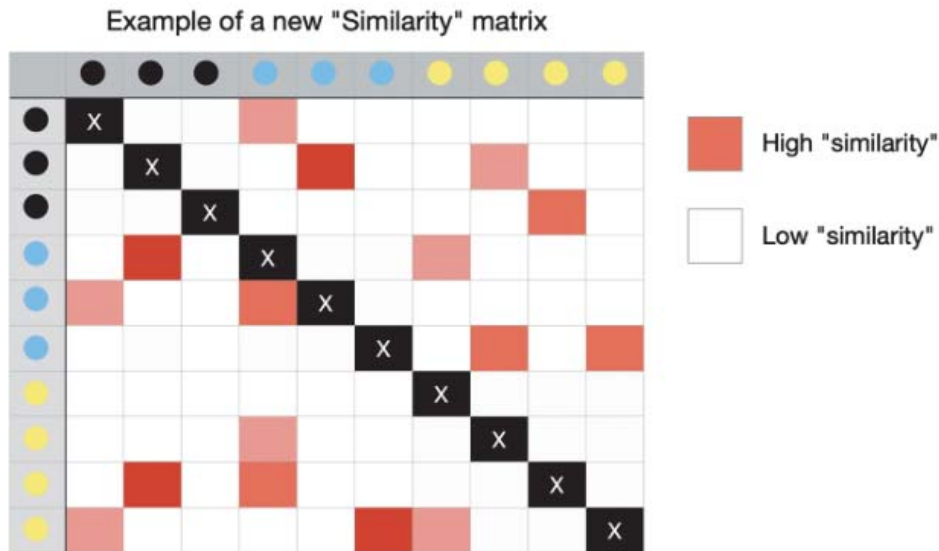
Նկ. 11.8: Նմանության մատրից

Կատարվում է այս ֆունկցիայի մինիմիզացիա գրադիենտային իջեցման եղանակով: Ֆունկցիայի գրադիենտն ունի հետևյալ տեսքը՝

$$\frac{\partial C}{\partial y_i} = 2 \sum_j (p(j, i) - q(j, i) + p(i, j) - q(i, j))(y_i - y_j):$$

Որպես y_i վեկտորի սկզբնական մոտարկում վերցվում է գրոյական մաթեմատիկական սպասումով և փոքր դիսպերսիայով գաուսյան բաշխում ունեցող պատահական մեծության որևէ արժեք:

Դիտարկենք մի օրինակ: Նկ. 11.7-ում պատկերված են օբյեկտներ: Կատարված են յուրաքանչյուր օբյեկտի նմանության գործակիցների հաշվարկ և այն ներկայացված է մատրիցի տեսքով (նկ. 11.8):



Նկ. 11.9: Նմանության նոր մատրից

Այնուհետև պատահականորեն կատարվում է այդ օբյեկտների արտապատկերում փոքր չափողականությունն ունեցող տարածությունում և այդ պատկերների համար կառուցվում է նմանության նոր մատրից (նկ.11.9): Կիրառելով գրադիենտային իջեցում այս նմանության մատրիցը քայլ առ քայլ ձևափոխվում է և նմանվում է սկզբնական մատրիցին: Գրադիենտային իջեցում կատարվում է հետևյալ բանաձևով՝

$$y_i^{t+1} = y_i^t + \eta \frac{\partial C}{\partial y_i} \Big|_{y_i^t}, \quad y_i \in Y:$$

Գոյություն ունի *SNE* ալգորիթի մի քանի մոդիֆիկացիաներ, որոնցից մեկը կոչվում է *Symmetric SNE*: Այս դեպքում

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} \ln \frac{p_{ij}}{q_{ij}},$$

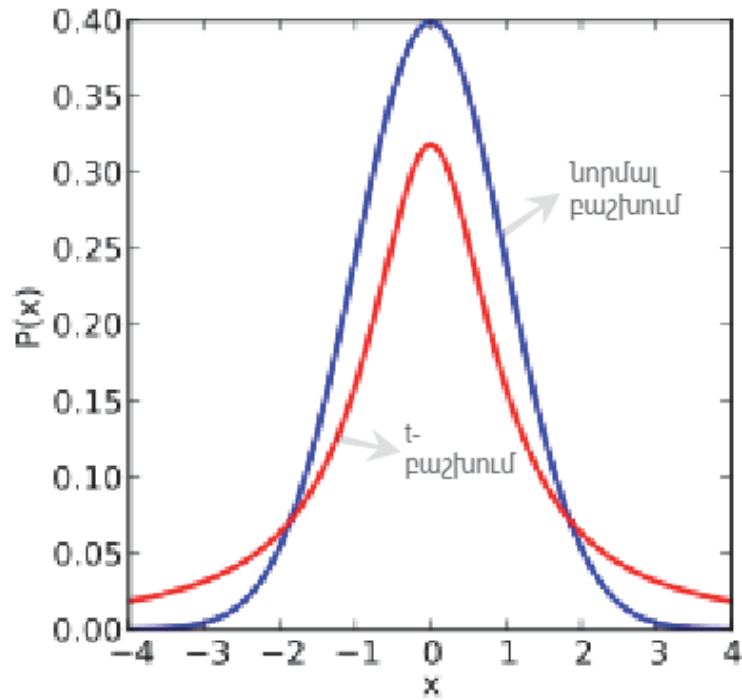
որտեղ

$$q_{ij} = \frac{\exp(-||y_i - y_j||^2)}{\sum_{k \neq l} \exp(-||y_k - y_l||^2)},$$

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2|X|}:$$

Ակնհայտ է, որ $p_{ij} = p_{ji}$, $q_{ij} = q_{ji}$: Իսկ C ֆունկցիայի գրադիենտը տվյալ դեպքում կունենա հետևյալ տեսքը՝

$$\frac{\partial C}{\partial y_i} = 2 \sum_j (p_{ij} - q_{ij})(y_i - y_j):$$



Նկ. 11.10: Նորմալ և t - բաշխումներ

Երկրորդ մոդիֆիկացիան կոչվում է $t - SNE$: Այս դեպքում փոքր չափանի տարածությունում նորմալ բաշխման փոխարեն տրվում է օբյեկտների բաշխման նոր բանաձև, որը կոչվում է t -բաշխում: Այսինքն՝

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}} :$$

Նկ.11.10-ում ներկայացված են նորմալ և t -բաշխումների գրաֆիկները:

Այժմ ներկայացնենք $UMAP$ ալգորիթմի աշխատանքի ընդհանուր սխեման: Սկզբից հաշվարկվում է կետերի հեռավորությունները մեծ չափանի տարածությունում: Այնուհետև կատարվում է պրոյեկտում փոքր չափանի տարածության վրա և հաշվարկվում է այդ պրոյեկցիաների հեռավորությունը: Սահմանվում է նմանության գործակիցներ ինչպես մեծ չափանի տարածությունում գտնվող օբյեկտների, այնպես էլ նրանց պատկերների միջև: Որպես կորստի ֆունկցիոնալ վերցնում են $cross\ entropy$ ֆունկցիան: Հետո օգտագործելով ստոխաստիկ գրադիենտային իջեցում՝ մինիմիզացվում է այդ ֆունկցիոնալը:

Կարճ նկարագրենք $UMAP$ ալգորիթմի սխեման: Դիցուք $X = \{x_1, x_2, \dots, x_N\}$ տվյալների հավաքու է, որի վրա կա $d : X \times X \rightarrow [0, +\infty)$ մետրիկա: $UMAP$ -ը սկզբից գտնում է x_i օբյեկտի k -մոտիկ N_i հարևանների: Այնուհետև այն հաշվում է երկու պարամետր՝ ρ_i և σ_i ՝ հետևյալ բանաձևերով՝

$$\rho_i = \min_{j \in N_i} \{d(x_i, x_j) | d(x_i, x_j) > 0\},$$

$$\sum_{j \in N_i} \exp\left(\frac{-\max(0, d(x_i, x_j)) - \rho_i}{\sigma_i}\right) = \log_2(k):$$

Այնուհետև ալգորիթը, օգտագործելով այդ պարամետրերը, հաշվում է x_i օբյեկտի նմանության գործակիցը (չափանիշը) x_j օբյեկտի նկամամբ հետևյալ բանաձևով՝

$$v_{j|i} = \exp\left(\frac{-\max(0, d(x_i, x_j)) - \rho_i}{\sigma_i}\right):$$

Օգտագործելով $v_{j|i}$ և $v_{i|j}$ արժեքները՝ *UMAP*-ը ձևավորում է $v_{ij} = v_{j|i} + v_{i|j} - v_{j|i}v_{i|j}$ էլեմենտներից բաղկացած սիմետրիկ մատրիցը: Անալոգիայով փոքր չափանի տարածության մեջ y_i և y_j օբյեկտների միջև նմանության միավորը (գործակիցը, similarity score) սահմանվում է հետևյալ բանաձևով՝

$$w_{ij} = (1 + \alpha \|y_i - y_j\|_2^{2\beta})^{-1},$$

որտեղ α -ն և β -ն դրական հաստատուններ են: Ի տարբերություն *SNE* ալգորիթի, որը մինիմիզացնում է Քուլբեր-Լեյբլի ֆունկցիան, այստեղ որպես կորստի ֆունկցիա վերցվում է *cross entropy*-ը v_{ij} և w_{ij} տվյալների միջև: Այսինքն՝

$$C_{UMAP} = \sum_{i \neq j} \left(v_{ij} \ln\left(\frac{v_{ij}}{w_{ij}}\right) - (1 - v_{ij}) \ln\left(\frac{1 - v_{ij}}{1 - w_{ij}}\right) \right):$$

cross entropy-ի այս բանաձևը կարելի է ներկայացնել հետևյալ տեսքով՝

$$\begin{aligned} C_{UMAP} &= \sum_{i \neq j} (v_{ij} \ln v_{ij} + (1 - v_{ij}) \ln(1 - v_{ij})) \\ &\quad - \sum_{i \neq j} (v_{ij} \ln w_{ij} + (1 - v_{ij}) \ln(1 - w_{ij})): \end{aligned}$$

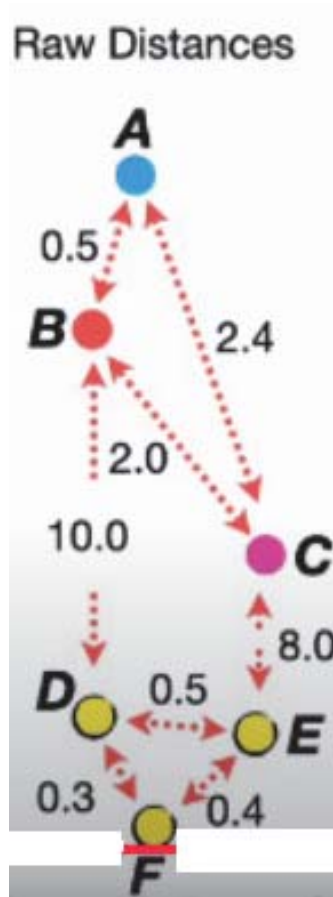
Քանի որ առաջին գումարելին կախված չէ $\{y_i, i = 1, 2, \dots, N\}$ պարամետրերից, ապա մինիմիզացվում է երկրորդ գումարելին՝

$$\begin{aligned} & - \sum_{i \neq j} (v_{ij} \ln w_{ij} + (1 - v_{ij}) \ln(1 - w_{ij})) = \\ & = \sum_{i \neq j} \left(v_{ij} \ln \frac{1}{w_{ij}} + (1 - v_{ij}) \ln \frac{1}{1 - w_{ij}} \right): \end{aligned}$$

Այժմ բերենք մի թվային օրինակ՝ *UNAP*-ի աշխատանքը մեկնաբանելու համար (տե՛ս նկ. 11.11): Նշված գրաֆում, եթե որոշ կետերի հեռավորությունը չի նշված նշանակում է՝ դրանց հեռավորությունը անվերջ է: Օրինակ՝ $d(A, D) = +\infty$:

Ենթադրենք $k = 3$: Այդ դեպքում օրինակի համար C օբյեկտի երեք մոտիկ հարևաններն են A, B, E օբյեկտները: Քանի որ $k = 3$, ապա

$$\log_2(\text{num.neighbors}) = \log_2 3 = 1.6:$$



Նկ. 11.11: Տվյալների հավաքածու և հեռավորություններ

Այժմ հաշվենք յուրաքանչյուր օբյեկտի նմանության չափանիշները իր հարե-վանների հետ: (օրինակ՝ A օբյեկտի համար ρ_A և σ_A պարամետրերի արժեքները): Քանի որ A օբյեկտը ունի երկու մոտիկ հարևան՝ B -ն և C -ն, հետևաբար $\rho_A = 0.5$: Այժմ պետք է գտնել σ_A -ն: Այն հետևյալ հավասարման արմատն է՝

$$\exp\left(-\frac{0.5 - 0.5}{\sigma}\right) + \exp\left(-\frac{(2.4 - 0.5)}{\sigma}\right) = \log_2 3 = 1.6:$$

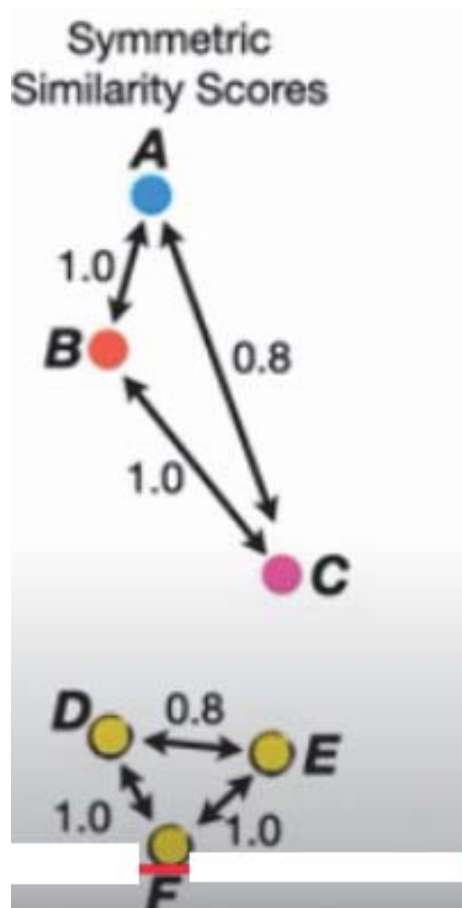
Այստեղից հետևում է, որ $\sigma = 3.7$:

Այս օրինակում (նկ. 11.11) հարմարության համար x օբյեկտի նմանության գործակիցը y օբյեկտի նկատմամբ նշանակենք $v(x|y)$ սիմվոլով:

Այսպիսով, $v(B|A) = 1$, $v(C|A) = 0.6$: Այսպես հաշվարկելով՝ կստանանք նմանության հետևյալ միավորները՝

$$v(A|B) = 1, v(C|B) = 0.6, v(A|C) = 0.6, v(B|C) = 1:$$

Ելնելով այս միավորներից՝ կառուցվում են սիմետրիկ միավորները (simital score), որոնց հաշվարկը կատարվում է վերը նշված բանաձևերով:



Նկ. 11.12: Սիմետրիկ միավորներով գրաֆ

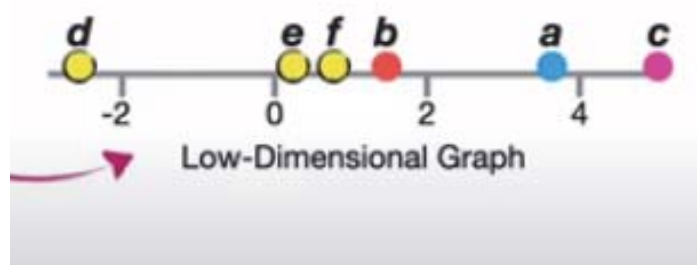
$$v(A, B) = v(B, A) = v(A|B) + v(B|A) - v(A|B)v(B|A) = 1 + 1 - 1 = 1,$$

$$v(B, C) = v(C, B) = (1 + 0.6) - 0.6 = 1, \quad v(A, C) = v(C, A) = 0.6 + 0.6 - 0.6 \times 0.36 = 0.8:$$

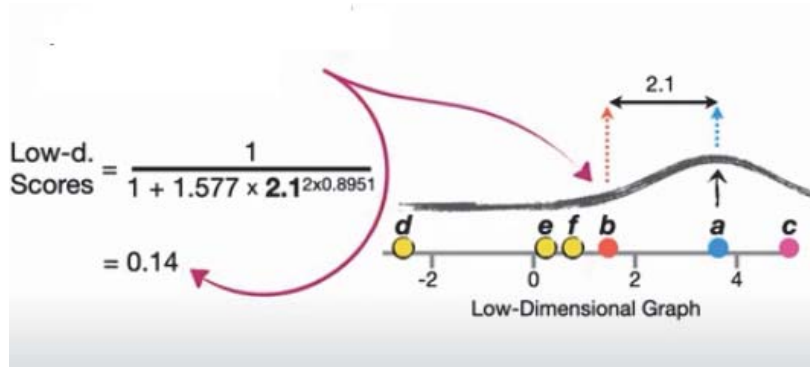
$$v(D, E) = v(E, D) = 0.8, \quad v(D, F) = v(F, D) = 1, \quad v(E, F) = v(F, E) = 1:$$

Այս հաշվարկներից հետո նկ. 11.12-ում տրված է սիմետրիկ միավորներով նոր գրաֆ:

Այժմ ենթադրենք, որ առանցքի վրա օբյեկտների պատկերները դասավորված են ինչպես նկ. 11.13-ում : Այդ նկարում օբյեկտները դասավորված են առանցքի վրա: Այժմ հաշվենք նմանության միավորները օբյեկտների պատկերների միջև



Նկ. 11.13: Օբյեկտների պատկերները առանցքի վրա



Նկ. 11.14: Օբյեկտների պատկերների նմանության միավորները

փոքր չափի տարածությունում: Այդ հաշվարկը կատարվում է

$$q_{ij} = \frac{1}{1 + \alpha d^{2\beta}}, \quad d = \|y_i - y_j\|$$

բանաձևով, որտեղ α և β դրական պարամետրեր են: Նկ. 11.14-ում հաշված է նմանության միավորը a b օբյեկտների միջև $\alpha = 1.577$ և $\beta = 0.8951$ արժեքների դեպքում: Նկ. 11.15-ում ցույց է տրված, որ b օբյեկտի նմանության գործակիցը հարևան a -ի նկատմամբ ավելի փոքր է, քան e օբյեկտի նմանմամբ: B -ի նմանության միավորը A -ի նկատմամբ հավասար է մեկի, իսկ E -ի նկատմամբ՝ զրոյի: Այսպիսով կարող ենք եզրակացնել, որ b օբյեկտը իր տեղում չէ: Այդ օբյեկտի դիրքը e և a օբյեկտների նկատմամբ որոշվում է

$$Cost = \ln\left(\frac{1}{neighbor}\right) + \ln\left(\frac{1}{1 - notneighbor}\right)$$

Ֆունկցիայի արժեքով: Եթե b օբյեկտը դասավորված է ինչպես նկար 11.16-ում, ապա

$$Cost = \ln\left(\frac{1}{0.14}\right) + \ln\left(\frac{1}{1 - 0.23}\right) = 2.23:$$

Եթե այժմ b օբյեկտը շարժենք a -ի ուղղությամբ, ապա $Cost$ -ի արժեքը նվազում է (տես նկ. 11.16): Այս դեպքում

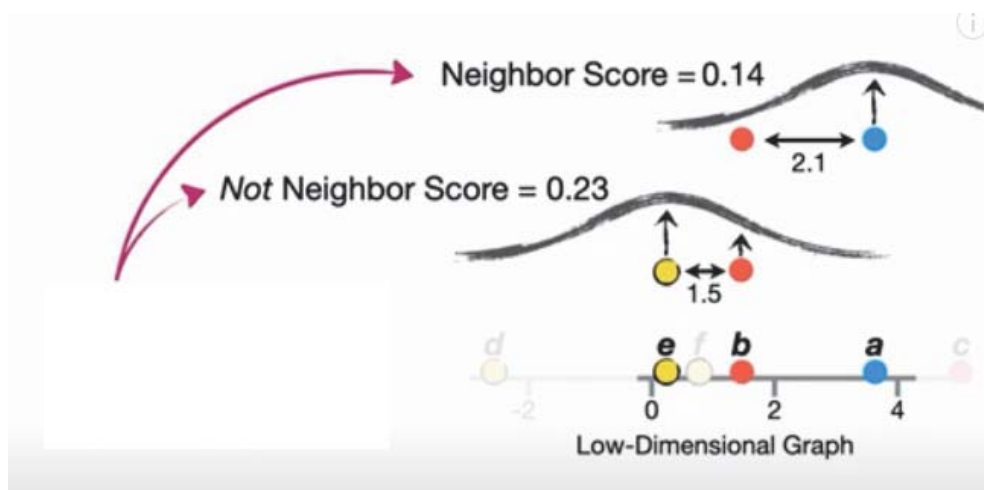
$$Cost = \ln\left(\frac{1}{0.98}\right) + \ln(1 - 0.06) = 0.08:$$

Բայց եթե b օբյեկտը տեղադրվի a -ից աջ, ինչպես նկ. 11.17-ում, ապա $Cost$ -ի արժեքը կմեծանա՝

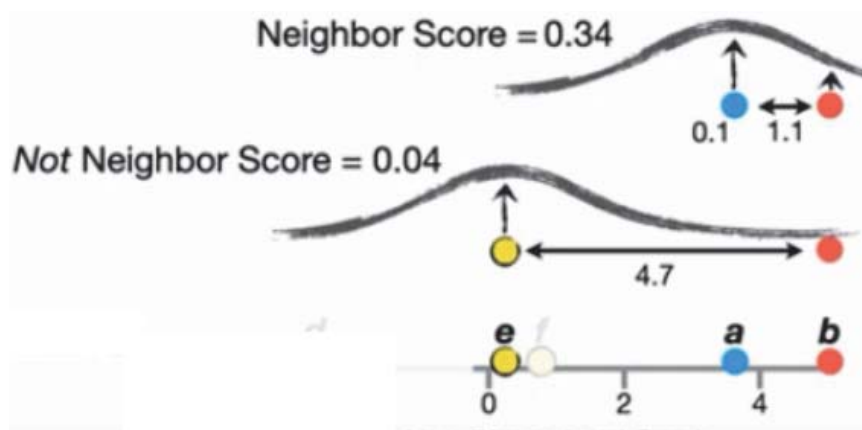
$$Cost = \ln\left(\frac{1}{0.34}\right) + \ln\left(\frac{1}{1 - 0.04}\right) = 1.12:$$

Վերջնական գրաֆը ցածր չափանի տարածությունում կձևավորվի C_{UMAP} ֆունկցիայի մինիմիզացիայի արդյունքում ստոխաստիկ գրադիենտի մեթոդով:

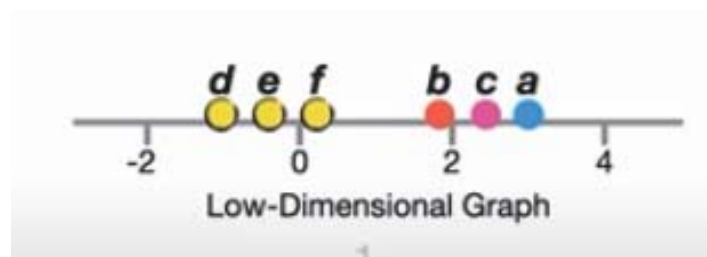
$UMAP$ ալգորիթմի աշխատանքի մանրամասնություններին կարելի է ծանոթանալ [14] հոդվածում:



Նկ. 11.15: Օբյեկտների պատկերների նմանության միավորները



Նկ. 11.16: Օբյեկտների պատկերների նմանության միավորները



Նկ. 11.17: Վերջնական գրաֆը ցածր չափանի տարածությունում

ԽՆԴԻՐՆԵՐ

1. Տրված է տվյալների հավաքածու (աղյուսակ 11.18): PCA մեթոդով գտնել գլխավոր կոմպոնենտները (ուղղությունները) և կետերի պրոյեկցիաները նրանց վրա:
2. Տրված է տվյալների հավաքածու (աղյուսակ 11.19): PCA մեթոդով գտնել գլխավոր կոմպոնենտները (ուղղությունները) և կետերի պրոյեկցիաները դրանց վրա:

x_1	4	0	-2	2
x_2	3	1	-3	-1

Նկ. 11.18: Տվյալների հավաքածու

x_1	4	0	-1	3	4
x_2	2	-3	-2	1	2
x_3	3	2	2	1	-3

Նկ. 11.19: Տվյալների հավաքածու

12

Կլաստերավորում (Կլաստերիզացիա)

Կլաստերավորումը (cluster analysis) օբյեկտների հավաքածուն ենթարգմոթյունների (կլաստերների) խմբավորման խնդիր է՝ ըստ որոշակի չափանիշի: Այդ չափանիշը պետք է լինի այնպիսին, որ մի կլաստերի առարկաները ավելի նման լինեն միմյանց, քան այլ կլաստերների օբյեկտներին՝ ըստ այդ չափանիշի:

Կլաստերավորման նպատակն է օբյեկտների դասակարգումը: Դա եղանակ է՝ հասկանալու օբյեկտների միջև կախվածությունը՝ բացահայտելով դրանց կլաստերային կառուցվածքը: Նմանատիպ օբյեկտներին մի խմբի մեջ հավաքելով՝ այն թույլ է տալիս ամեն մի կլաստերի նկատմամբ տալ տվյալների մշակման իր անալիզը:

Քննարկենք կլաստավորման առավել հայտնի մեթոդներից k -միջինների և *Dbscan* ալգորիթմները:

k -միջինների ալգորիթմի ժամանակ նախապես ֆիքսվում է կլաստերների քանակը, և այնուհետև ալգորիթմը կատարում է օբյեկտների հավաքածուի տրոհում այդ դասերի:

Իսկ *Dbscan* ալգորիթմում կլաստերների քանակը ձևավորվում է աշխատանքի ընթացքում:

12.1 k -միջինների ալգորիթմը

Թեորեմ 12.1. Դիցուք ունենք հավաքածու՝ $X = \{x_1, x_2, \dots, x_l\}$: Բաժանենք այն k խմբերի հետևյալ ալգորիթմով:

- Վերցնում ենք յուրաքանչյուր դասի համար պատահական կենտրոններ՝ $c_1^{(0)}, \dots, c_k^{(0)}$, $n = 0$:
- Բաժանենք հավաքածուն խմբերի «մոտիկ հարևանի» մեթոդով: Այսինքն՝ $X_i^{(n)}$ դասի մեջ գցում ենք այն էլեմենտները, որոնք ավելի մոտիկ են $c_i^{(n)}$ կետին: Այսպես ձևավորվում են $X_1^{(n)}, \dots, X_k^{(n)}$ դասերը:

- Հաշվարկվում են դասերի՝ կլաստերների նոր կենտրոնները հետևյալ բանաձևով՝

$$c_i^{(n+1)} = \frac{1}{|X_i^{(n)}|} \sum_{x \in X_i^{(n)}} x:$$

- Եթե $c_i^{(n+1)} = c_i^n$, $i = 1, 2, \dots, k$, ապա ալգորիթմն ավարտվում է, հակառակ դեպքում անցնում ենք երկրորդ քայլին:

Այդ դեպքում ալգորիթմը վերջավոր քայլերի ընթացքում զուգամետ է:

Ապացույց: Նշանակենք

$$A: \{1, 2, \dots, l\} \rightarrow \{1, 2, \dots, k\}:$$

Եթե $A(i) = q$, ապա սա նշանակում է, որ i օբյեկտը պատկանում է q կլաստերին: Սահմանենք կորստի ֆունկցիան (loss-function) հետևյալ կերպ՝

$$L(A, m_1, m_2, \dots, m_k) = \sum_{q=1}^k \sum_{A(i)=q} \|x_i - m_q\|^2,$$

որտեղ m_i , $i = 1, 2, \dots, k$ կլաստերների կենտրոններն են: Իրականացնում ենք հետևյալ ալգորիթմը:

- Վերցնում ենք պատահական արտապատկերում $A^{(0)}$:
- Մինիմիզացնում ենք $L(A^{(0)}, m_1, \dots, m_k)$ ֆունկցիան ըստ $m_1, m_2, \dots, m_k$ փոփոխականների և գտնում նոր արժեքներ՝ $m_1^{(0)}, \dots, m_k^{(0)}$:
- Այնուհետև $A_1^{(1)}$ -ով նշանակենք այն արտապատկերումը, որը մինիմիզացնում է

$$L(A, m_1^{(0)}, \dots, m_k^{(0)})$$

ֆունկցիան ըստ A -ի: Եվ այսպես կառուցում ենք նոր արտապատկերում՝

$$A^{(n+1)} = \arg \min_A L(A, m_1^{(n)}, \dots, m_k^{(n)}):$$

Ամեն մի կլաստերի համար սահմանենք դրա կենտրոն հետևյալ բանաձևով՝

$$\eta(A, q) = \frac{1}{|\{i: A(i) = q\}|} \sum_{A(i)=q} x_i:$$

Ճշմարիտ են հետևյալ երկու պնդումները:

Պնդում 1: Ցանկացած $m_1, m_2, \dots, m_k$ վեկտորների համար տեղի ունի

$$L(A, m_1, \dots, m_k) \geq L(A, \eta(A, 1), \dots, \eta(A, k))$$

անհավասարությունը:

Ապացույց: Դիցուք $A(i) = q$: Այդ դեպքում՝

$$\begin{aligned} \|x_i - m_q\|^2 &= \|x_i - \eta(A, q) + \eta(A, q) - m_q\|^2 = \\ &= \|x_i - \eta(A, q)\|^2 + \|\eta(A, q) - m_q\|^2 + \\ &\quad + 2(x_i - \eta(A, q), \eta(A, q) - m_q) \geq \\ &\geq \|x_i - \eta(A, q)\|^2 + 2(x_i - \eta(A, q), \eta(A, q) - m_q): \end{aligned}$$

Հետևաբար՝ քանի որ $\sum_{A(i)=q} (x_i - \eta(A, q)) = 0$, ապա

$$\begin{aligned} \sum_{A(i)=q} \|x_i - m_q\|^2 &\geq \sum_{A(i)=q} \|x_i - \eta(A, q)\|^2 + 2 \sum_{A(i)=q} (x_i - \eta(A, q), \eta(A, q) - m_q) = \\ &= \sum_{A(i)=q} \|x_i - \eta(A, q)\|^2: \end{aligned}$$

Այստեղից՝ քանի որ

$$L(A, m_1, m_2, \dots, m_k) = \sum_{q=1}^k \sum_{A(i)=q} \|x_i - m_q\|^2,$$

ապա

$$L(A, m_1, \dots, m_k) \geq L(A, \eta(A, 1), \dots, \eta(A, k)):$$

Նշանակենք՝

$$A(m_1, m_2, \dots, m_k)(i) = \arg \min_{q \in \{1, 2, \dots, k\}} \|x_i - m_q\|^2:$$

Պնդում 2: Կամայական A արտապատկերման ժամանակ տեղի ունի հետևյալ անհավասարությունը՝

$$L(A, m_1, m_2, \dots, m_k) \geq L(A(m_1, m_2, \dots, m_k), m_1, m_2, \dots, m_k):$$

Ապացույց: Ունենք $L(A, m_1, m_2, \dots, m_k) = \sum_{q=1}^k \sum_{A(i)=q} \|x_i - m_q\|^2 \geq$

$$\geq \sum_{q=1}^k \sum_{A(m_1, m_2, \dots, m_k)(i)=q} \|x_i - m_q\|^2 = L(A(m_1, m_2, \dots, m_k), m_1, \dots, m_k):$$

Այժմ անցնենք թեորեմի ապացույցին: Դիցուք

$$L_n \equiv L(A^{(n+1)}, m_1^{(n)}, \dots, m_k^{(n)}):$$

Ըստ սահմանման՝

$$m_q^{(n+1)} = \eta(A^{(n+1)}, q), \quad A^{(n+2)} = A(m_1^{(n+1)}, \dots, m_k^{(n+1)}):$$

Հետևաբար՝

$$\begin{aligned} L_n &\geq L(A^{(n+1)}, m_1^{(n+1)}, \dots, m_k^{(n+1)}) \geq \\ &\geq L(A^{(n+2)}, m_1^{(n+1)}, \dots, m_k^{(n+1)}) = L_{n+1}: \end{aligned}$$

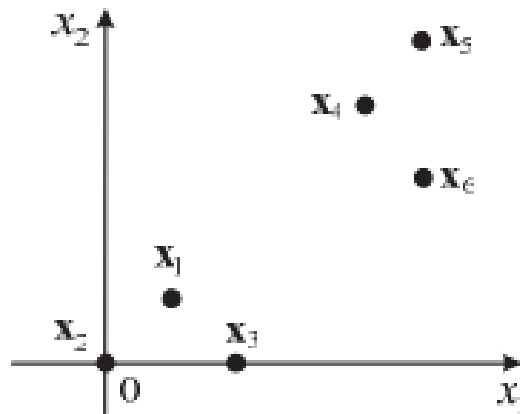
Այսպիսով $L_n \geq 0$ հաջորդականությունը մոնոտոն նվազող է և ներքևից սահմանափակ: Ուրեմն ունի սահման: Քանի որ A արտապատկերումների քանակը վերջավոր է՝ հավասար է N^K - ի, հետևաբար ալգորիթմը կգուգամիտի վերջավոր քայլերում:

Բերենք օրինակ այս ալգորիթմը մեկնաբանելու համար:

Ենթադրենք՝ հարթության վրա տրված են կետեր (տե՛ս նկ. 12.1)

$$x_1 = (1, 1), x_2 = (0, 0), x_3 = ((2, 0), x_4 = (4, 4), x_5 = (5, 5), x_6 = (5, 3):$$

Բաժանենք այս բազմությունը երկու դասերի:



Նկ. 12.1: Դասակարգման օրինակ

- Որպես սկզբնական կենտրոններ դասերի համար ընտրենք հետևյալ կետերը՝ $c_1^0 = x_1, c_2^0 = x_2$: Ըստ մոտիկ հարևանների մեթոդի՝ բազմությունը կբաժանվի հետևյալ երկու դասերի՝

$$X_1^0 = \{x_1, x_3, x_4, x_5, x_6\}, X_2^0 = \{x_2\}:$$

- Գտնենք դասերի (կլաստերների) ծանրության կենտրոնները.

$$c_1^{(1)} = (17/5, 13/5), c_2^{(1)} = (0, 0):$$

- Քանի որ $c_1^1 \neq c_1^0, c_2^0 = c_2^1$, ապա ալգորիթմը շարունակվում է:
- Բազմությունը ըստ նոր կենտրոնների կտրոհվի հետևյալ երկու կլաստերների՝

$$X_1^{(1)} = \{x_4, x_5, x_6\}, X_2^{(2)} = \{x_1, x_2, x_3\}:$$

- Նորից հաշվելով ծանրության կենտրոնները՝ ստանում ենք

$$c_1^{(2)} = (14/3, 4), \quad c_2^{(2)} = (1, 1/3):$$

- Քանի որ $c_1^{(1)} \neq c_1^{(2)}, c_2^{(1)} \neq c_2^{(2)}$, ապա ալգորիթմը շարունակվում է:
- Ստանում ենք երկու նոր կլաստերներ՝

$$x_1^{(2)} = \{x_4, x_5, x_6\}, \quad X_2^{(2)} = \{x_1, x_2, x_3\}:$$

- Նորից հաշվելով ծանրության կենտրոնները՝ ստանում ենք

$$c_1^{(3)} = c_1^{(2)}, \quad c_2^{(3)} = c_2^{(2)}:$$

Հետևաբար ալգորիթմը ավարտվում է:

k -means ալգորիթմի աշխատանքի տեսանելի (visible) տարբերակին կարելի է ծանոթանալ հետևյալ հղումով՝ <https://www.naftaliharris.com/blog/visualizing-k-means-clustering/>:

12.2 Կլաստերավորման մետրիկական մեթոդներ: Dbscan ալգորիթմը

Սկզբում նկարագրենք կլաստերավորման մի պարզագույն մետրիկական եղանակ: Դիցուկ ունենք նմուշների $\{x_1, x_2, \dots, x_N\}$ հավաքածու: Դիտարկենք երկու դասեր հետևյալ նմուշներով՝

$$\omega_1 = \{x_1, x_2, \dots, N_1\}, \quad \omega_2 = \{x_{N_1+1}, \dots, x_N\}:$$

Այդ դասերի (կլաստերների) կենտրոններն են՝

$$c_1 = \frac{1}{N_1} \sum_{i=1}^{N_1} x_i, \quad c_2 = \frac{1}{N_2} \sum_{i=1}^{N_2} x_i:$$

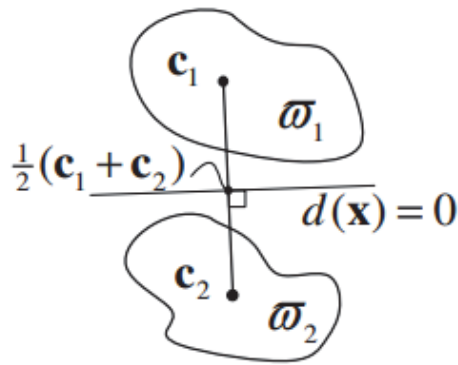
Կասենք, որ x օբյեկտը պատկանում է ω_1 դասին, եթե

$$\|x - c_1\| \leq \|x - c_2\|:$$

Հետևաբար որպես դասերը բաժանող ֆունկցիա կարելի վերցնել հետևյալը՝

$$d(x) = \|x - c_1\|^2 - \|x - c_2\|^2 = 2(c_1 - c_2)(x - \frac{1}{2}(c_1 + c_2)) \quad (\text{տե՛ս նկ. 12.2}):$$

Այսպիսով, բաժանող հիպերհարթությունը կլինի $d(x) = x - \frac{1}{2}(c_1 + c_2) = 0$ գիծը:



Նկ. 12.2: Մետրիկական կլաստերավորում

Այժմ նկարագրենք կլաստերավորման ավելի ժամանակակից և լայն տարածված Dbscan ալգորիթը:

Dbscan (Density-based spatial clustering of applications with noise)-ը «խտության» կլաստերավորման ալգորիթն է, որը խմբավորում է «խիտ փաթեթավորված» կետերը: Դիցուք ունենք X հավաքածու և

$$\epsilon > 0, \minPts > 0$$

պարամետրեր: Տանք հետևյալ սահմանումները:

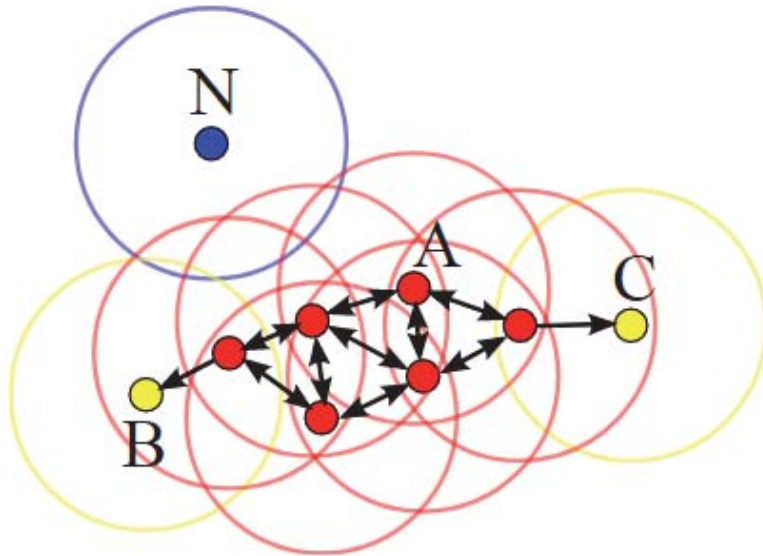
- $p \in X$ կետը կոչվում է միջուկային (core), եթե նրա ϵ շրջակայքում կան առնվազն $\minPts$ կետեր: Ատում են, որ այդ կետերը «ուղիղ հասանելի» (directly density-reachable) են p կետից:
- $q \in X$ կետը կոչվում է հասանելի (density-reachable) p կետից, եթե գոյություն ունի կետերի հաջորդականություն՝

$$p_1, p_2, \dots, p_n, p_1 = p, p_n = q,$$

այնպիսին, որ յուրաքանչյուր p_{i+1} կետ լինի ուղիղ հասանելի p_i կետից: Բոլոր կետերը, բացի q -ից, պետք է լինեն միջուկային:

- $r \in X$ կետը կոչվում է դեն նետված (outlier), եթե այն հասանելի չէ ոչ մի կետից:

Ալգորիթմի էությունն այն է, որ կետը և դրանից հասանելի բոլոր կետերը ձևավորում են կլաստեր: Կլաստերում կարող են լինել միջուկային և ոչ միջուկային կետեր: Այսպիսով ամեն մի կլաստեր պարունակում է առնվազն մեկ միջուկային կետ: Ալգորիթմն սկսվում է կամայական կետից: Եթե նրա ϵ -շրջակայքում չկան առնվազն $\minPts$ կետեր, ապա այդ կետը նշվում է որպես Noise, հակառակ դեպքում ձևավորվում է կլաստեր K , որը պարունակում է այդ շրջակայքի բոլոր կետերը: Եթե այդ շրջակայքի ինչ-որ մի կետ պատկանում է մի ուրիշ C կլաստերի, ապա այդ այդ C կլաստերի կետերը ավելացվում են K կլաստերին: Այնուհետև դիտարկվում է նոր «չայցելած» կետ, որից կձևավորվի նոր կլաստեր կամ Noise հատկանիշով կետ:



Նկ. 12.3: Dbscan

Նկար 12.3-ում օրինակով տրված են նշված կատեգորիայի կետերը:

Իսկ նկար 12.4-ում օրինակով տրված է *density – reachable* հասկոությունը:

Եթե A կետը միջուկային է, ապա այն ձևավորում է կլաստեր բոլոր այն կետերի հետ, որոնք հասանելի են այդ կետից: B և C կետերը եզրային են (edge), քանի որ մտնում են կլաստերի մեջ, բայց միջուկային չեն:

N կետը դեմ նետված է:

Dbscan ալգորիթմը հետևյալն է:

for each $p \in X$ **do**

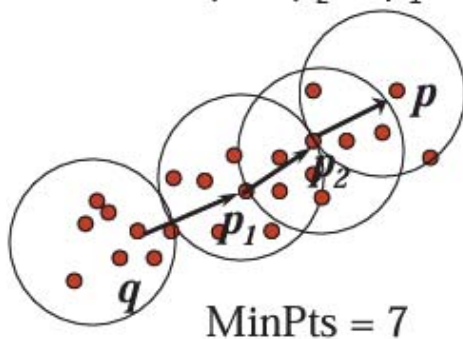
if p is not yet classified then

if p is a core-object then collect all objects density-reachable from p and assign them to a new cluster

else assign p to NOISE.

Density-reachability

- Density-Reachable (directly and indirectly):
 - A point p is directly density-reachable from p_2
 - p_2 is directly density-reachable from p_1
 - p_1 is directly density-reachable from q
 - $p \leftarrow p_2 \leftarrow p_1 \leftarrow q$ form a chain



- p is (indirectly) density-reachable from q
- q is not density-reachable from p

Uq. 12.4: Density-reachable

Այժմ նկարագրենք Dbscan ալգորիթմի կոդը:
 Dbscan (D, eps, MinPts) C = 0

```

for each unvisited point P in dataset D
mark P as visited
NeighborPts = regionQuery(P, eps)
if sizeof (NeighborPts) < MinPts
mark P as NOISE

else
C = next cluster
expandCluster (P, NeighborPts, C, eps, MinPts)
expandCluster (P, NeighborPts, C, eps, MinPts)
add P to cluster C

```

```

for each point P' in NeighborPts
if P' is not visited
mark P' as visited
NeighborPts' = regionQuery(P', eps)

```

```

if sizeof (NeighborPts') >= MinPts
NeighborPts = NeighborPts joined with
NeighborPts'

```

```

if P' is not yet member of any cluster
add P' to cluster C
regionQuery (P, eps)

```

return all points within P's eps-neighborhood (including P)

Այժմ դիտարկենք մի թվային օրինակ այս ալգորիթմի աշխատանքը մեկնաբանելու համար: Դիցուք տված է տվյալների հետևյալ բազան հարթության վրա (աղյուսակ 12.1):

Աղյուսակ 12.1(տվյալների բազա)

P1(3,7)	P2(4,6)
P3(5,5)	P4(6,4)
P5(7,3)	P6(6,2)
P7(7,2)	P8(8,4)
P9(3,3))	P10(2,6)
P11(3,5)	P12(2,4)

minPts = 4 and epsilon(ϵ) = 1.9												
	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
P1: (3, 7)	0											
P2: (4, 6)	1.41	0										
P3: (5, 5)	2.83	1.41	0									
P4: (6, 4)	4.24	2.83	1.41	0								
P5: (7, 3)	5.66	4.24	2.83	1.41	0							
P6: (6, 2)	5.83	4.47	3.16	2.00	1.41	0						
P7: (7, 2)	6.40	5.00	3.61	2.24	1.00	1.00	0					
P8: (8, 4)	5.83	4.47	3.16	2.00	1.41	2.83	2.24	0				
P9: (3, 3)	4.00	3.16	2.83	3.16	4.00	3.16	4.12	5.10	0			
P10: (2, 6)	1.41	2.00	3.16	4.47	5.83	5.66	6.40	6.32	3.16	0		
P11: (3, 5)	2.00	1.41	2.00	3.16	4.47	4.24	5.00	5.10	2.00	1.41	0	
P12: (2, 4)	3.16	2.83	3.16	4.00	5.10	4.47	5.39	6.00	1.41	2.00	1.41	0

Նկ. 12.5: Կետերի հեռավորություն

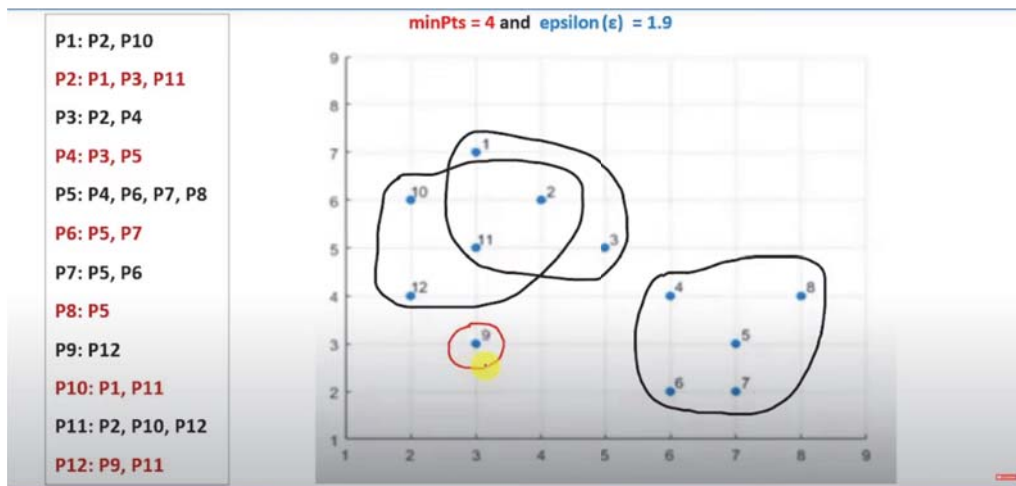
minPts = 4 and epsilon(ϵ) = 1.9		
Point	Status	
P1	Noise	Border
P2	Core	
P3	Noise	Border
P4	Noise	Border
P5	Core	
P6	Noise	Border
P7	Noise	Border
P8	Noise	Border
P9	Noise	
P10	Noise	Border
P11	Core	
P12	Noise	Border

Նկ. 12.6: Կետերի ստատուսները

Այս հավաքածուի նկատմամբ կիրառենք *Dbscan* ալգորիթը $\epsilon = 1.9$, $minPts = 4$ պարամետրերով: Կետերը դասավորենք աղյուսակի տեսքով և հաշվենք դրանց փոխադարձ հեռավորությունները, ինչպես պատկերված է նկար 12.5-ում: Այնուհետև յուրաքանչյուր պոնտակում նշում ենք բոլոր այն թվերը, որոնք փոքր են $\epsilon = 1.9$ պարամետրից: Դրանից հետո ձևավորվում են կետերի ստատուսները: Միջուկային են p_2, p_5, p_{11} կետերը, քանի որ նրանց ϵ շրջակայքում դասավորված է չորս կետ: Մնացած կետերը ստանում են *Noise* ստատուսը: Այսպիսով ձևավորվում են երեք կլաստերներ: Դրանից հետո բոլոր կետերը ստանում են *Border* ստատուսը, քանի որ մտնում են որևէ կլաստերի մեջ: Միայն P_9 կետը մնում է *Noise*, որովհետև այն ոչ մի կլաստերի մեջ չի մտնում: Նկար 12.6-ում ներկայացված են ստեղծված կլաստերները և կետերի ստատուսները:

Արդյունքում ձևավորվում են երեք կլաստեր և մի *Noise* կետ (նկ. 12.7):

Տեսանելիորեն (visible) *Dbscan* clustering ալգորիթի աշխատանքին կարելի է ծանոթանալ հետևյալ հղումով՝ <https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/>, որտեղ կետերի հավաքածուն ունի Պուլանսոնյան բաշխում:



Նկ. 12.7: Կետերի կլաստերները

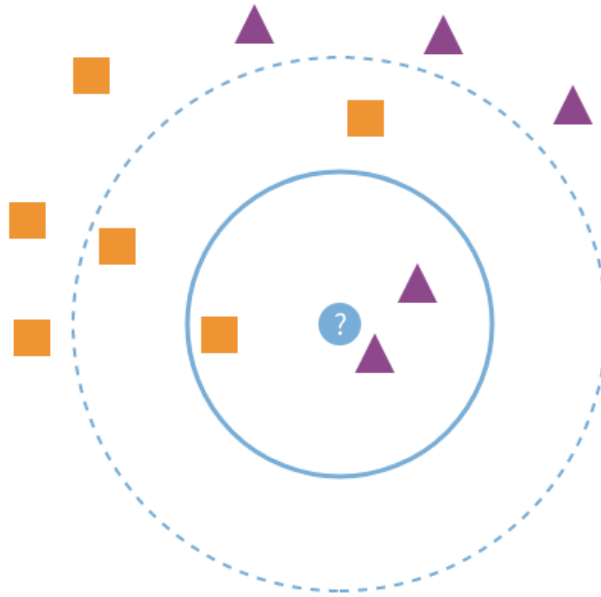
Իսկ նորմալ բաշխված կետերի Dbscan clustering՝ կարելի է տեղեկանալ նաև հետևյալ հղումով՝ <https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/>: Վերջում նշենք, որ k միջինների և Dbscan ալգորիթների տարբերությունը այն է, որ k -միջինների ժամանակ սկզբից տրվում է, թե քանի կլաստերի պիտի տրոհել օբյեկտները, իսկ Dbscan-ի աշխատանքի ընթացքում է որոշվում կլաստերների քանակը:

12.3 k -մոտիկ հարևանների մեթոդը

Դիցուք ունենք օբյեկտների S հավաքածու, որը բաժանված է $C_i, i = 1, 2, \dots, m$ դասերի: Ենթադրենք՝ կա մի նոր օբյեկտ, և ցանկանում ենք այն տեղադրել դասերից մեկի մեջ: Ենթադրենք՝ տրված է ինչ-որ մետրիկա օբյեկտների միջև: Հաշվում ենք այդ օբյեկտի հեռավորությունները հավաքածուի օբյեկտներից և ընտրում ենք այն k հատ օբյեկտները, որոնք ամենամոտիկն են այդ նոր x օբյեկտին: Օբյեկտը այնհույսեան դասում ենք այն խմբում, որի էլեմենտները ամենաշատն են այդ k հարևաններում: Օրինակ՝ նկար 12.8-ում «կապույտ օբյեկտը» (հարցական մշանով) $k = 3$ -ի դեպքում պատկանում է «եռանկյուն» դասին, իսկ $k = 5$ -ի դեպքում՝ «ուղղանկյուն» դասին:

Դիտարկենք մի թվային օրինակ: Դիցուք ունենք վարկառուների ցուցակ, որոնցից ոմանք ժամանակին չեն վերադարձրել վարկը: Նրանց տվյալները ներկայացված են աղյուսակ 12.2-ում: Դիցուք ունենք մի օբյեկտ որի կոորդինատներն են (42, 34): Ուզում ենք իմանալ, թե որ դասի մեջ այն կլինի, եթե կիրառենք k -մոտիկ հարևանների ալգորիթը $k = 3$ -ի դեպքում:

Հաշվենք այս օբյեկտի հեռավորությունները բոլոր օբյեկտներից և ռանգավորենք (աղյուսակներ 12.3-12.4): Աղյուսակ 12.4-ից երևում է, որ քննարկվող օբյեկտը ունի երեք մոտիկ հարևան, որոնցից երկուսը «այո» դասից են : Հետևաբար դիտարկվող օբյեկտը պատկանում է «այո» դասին:



Նկ. 12.8: k-մոտիկ հարևանների ալգորիթմը (k=3)

Աղյուսակ 12.2		
Տարիքը	Եկամուտը (հազար դրամ)	y-Վարկի ժամանակին վերադ.
46	40	այո (y=1)
36	54	այո (y=1)
34	29	ոչ (y=-1)
28	23	ոչ (y=-1)

Աղյուսակ 12.3				
Տարիքը	Եկամուտը(հազար դրամ)	Հեռ.	Ռանգ	Հարևան
46	40	7.2	1	ոչ
36	54	20.9	4	ոչ
34	29	9.4	2	այո
28	23	11.7	3	այո

Աղյուսակ 12.4				
Տարիքը	Եկամուտը(հազար դրամ)	Հեռ.	Ռանգ	Հարևան
46	40	7.2	1	ոչ
34	29	9.4	2	այո
38	23	11.7	3	այո

Այժմ ձևակերպենք այս ալգորիթմը և գնահատենք դրա սխալանքը: Դիցուք ունենք $S = \{(x_1, y_1), (x_2, y_2), \dots, (x_l, y_l)\}$ հավաքածու: x օբյեկտի համար $(\pi_1(x), \pi_2(x), \dots, \pi_l(x))$ վեկտորը կլինի $(1, 2, \dots, l)$ վեկտորի տեղափոխություն, եթե

$$\rho(x, x_i) \leq \rho(x, x_{i+1}), \quad i < l:$$

Այսպիսով, $k - NN$ ալգորիթմը հետևյալն է.

- Ալգորիթմի մուտքին տրվում է նմուշների հավաքածու՝ $S = (x_1, y_1), \dots, (x_l, y_l)$:
- Ալգորիթմի ելքում ամեն մի x օբյեկտի համար տրվում է այն նշիչը, որ ամենաշատն է հանդիպվում $\{y_{\pi_i(x)}, i \leq k\}$ բազմությունում:

$k = 1$ -ի դեպքում մենք կունենանք հետևյալ պրեդիկտորը՝ $h_S(x) = y_{\pi_1(x)}$: Այժմ քննարկենք, այս պրեդիկտորի իրական սխալանքի չափը բինար դասակարգման խնդրում, այսինքն՝

$$Y = \{0, 1\}, l(h, (x, y)) = \mathbb{1}_{[h(x) \neq y]}:$$

Ենթադրվում է նաև՝ $[0, 1]^d$ օբյեկտների բազմությունն է, և հայտնի նաև պայմանական հավանականությունը՝ $\eta(x) \equiv P(y = 1|x)$: Հիշենք նաև, որ օպտիմալ բայեսյան պրեդիկտորը հետևյալն է՝ $h^*(x) = \mathbb{1}_{[\eta(x) > 1/2]}$: Կենթադրենք նաև, որ $\eta(x)$ ֆունկցիան բավարարում է Լիպշիցի պայմանը $c > 0$ հաստատունով: Այսինքն՝

$$|\eta(x) - \eta(x')| \leq c||x - x'||, \forall x, x':$$

Ուրիշ խոսքով ասած՝ սա նշանակում է, որ եթե երկու վեկտորներ մոտիկ են, ապա մոտիկ կլինեն նաև դրանց նշիչները: Հետևյալ լեմման տալիս է $1 - NN$ ալգորիթմի սպասվող սխալանքի գնահատական թեստային նմուշի և հավաքածուում գտնվող դրան մոտիկ հարևանի միջև:

Լեմմա 12.1. *Դիցուք $A = [0, 1]^d$, $Y = \{0, 1\}$, իսկ D -ն հավանականային բաշխում է $A \times Y$ -ի վրա, որի դեպքում $\eta(x)$ ֆունկցիան c -լիպշիցյան է A -ի վրա: Դիցուք $S = \{(x_1, y_1), \dots, (x_l, y_l)\}$ հավաքածու է, իսկ h_S -ը դրան համապատասխանող $1 - NN$ հիպոթեզը: Դիցուք h^* -ը օպտիմալ բայեսյան պրեդիկտորն է: Այդ դեպքում տեղի ունի հետևյալ գնահատականը՝*

$$E_{S \sim D^l}[L_D(h_S)] \leq 2l_D(h^*) + cE_{S \sim D^l, x \sim D_x}[||x - x_{\pi_1(x)}||]:$$

Ապացույց: Սկզբում ցանկացած երկու x, x' օբյեկտների համար գնահատենք հետևյալ հավանականությունը՝

$$P_{y \sim \eta(x), y' \sim \eta(x')}[y \neq y']:$$

Ունենք

$$\begin{aligned} P_{y \sim \eta(x), y' \sim \eta(x')}[y \neq y'] &= \\ &= \eta(x')(1 - \eta(x)) + \eta(x)(1 - \eta(x')) = \\ &= (\eta(x) - \eta(x) + \eta(x'))(1 - \eta(x)) + \\ &\quad + (1 - \eta(x) + \eta(x) - \eta(x'))\eta(x) = \\ &= 2\eta(x)(1 - \eta(x)) + (\eta(x) - \eta(x'))(2\eta(x) - 1): \end{aligned}$$

Հաշվի առնելով, որ $|2\eta(x) - 1| \leq 1$, ապա ստացված անհավասարությունից կստանանք՝

$$P_{y \sim \eta(x), y' \sim \eta(x')} [y \neq y'] \leq 2\eta(x)(1 - \eta(x)) + c||x - x'||:$$

Ունենք նաև

$$\begin{aligned} E_S[L_D(h_S)] &= E_{S_x \sim D_x^l, y \sim \eta(x), y' \sim \eta(\pi_1(x))} [\mathbb{1}_{[y \neq y']}] = \\ &= E_{S_x \sim D_x^l, x \sim D_x} [P_{y \sim \eta(x), y' \sim \eta(\pi_1(x))} [y \neq y']]: \end{aligned}$$

Հաշվի առնելով այս ներկայացումը՝ կստանանք

$$E_S[L_D(h_S)] \leq E_x[2\eta(x)(1 - \eta(x))] + cE_{S,x}[||x - x_{\pi_1(x)}||]:$$

Այժմ ցույց տանք՝ օպտիմալ բայեսյան h^* պրեդիկտորի համար տեղի ունի հետևյալ ներկայացումը՝

$$L_D(h^*) = E_x[\min\{\eta(x), (1 - \eta(x))\}]:$$

Իրոք, բայեսյան օպտիմալ պրեդիկտորի սխալանքի համար ճշմարիտ են հավասարությունների հետևյալ շղթան՝

$$\begin{aligned} E[h^*(x) \neq y] &= E[h^*(x) = 1][y = -1] + E[h^*(x) = -1][y = 1] = \\ &= E_x E[h^*(x) = 1][y = -1|x] + E_x E[h^*(x) = -1][y = 1|x] \\ &= E[h^*(x) = 1]E[y = -1|x] + E[h^*(x) = -1]E[y = 1|x] \\ &= E[h^*(x) = 1](1 - \eta(x)) + E[h^*(x) = -1]\eta(x) = \\ &= E[\eta(x) \geq 1/2](1 - \eta(x)) + E[\eta(x) < 1/2]\eta(x) \\ &= E_x[\min\{\eta(x), 1 - \eta(x)\}]: \end{aligned}$$

Վերջնականում կստանանք

$$\begin{aligned} E_S[L_D(h_S)] &\leq 2E_x(2\eta(x)(1 - \eta(x)) + cE_{x,S}[||x - \pi_1(x)||] \\ &\leq 2E_x(\min\{\eta(x), (1 - \eta(x))\} + cE_{x,S}[||x - \pi_1(x)||] = L_D(h^*) + cE_{x,S}[||x - \pi_1(x)||]: \end{aligned}$$

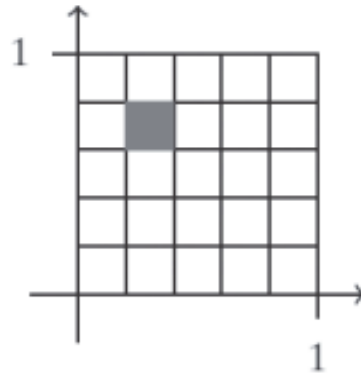
Թեորեմ 12.2. Դիցուք տեղի ունեն լեմմա 12.1-ի պայմանները: Այդ դեպքում ճշմարիտ է հետևյալ անհավասարությունը՝

$$E_{S \sim D^l} [L_D(h_S)] \leq 2L_D(h^*) + \frac{4c\sqrt{d}}{l^{1/(d+1)}}:$$

Ապացույց: Ֆիքսենք $\epsilon = 1/T$, որտեղ T որևէ բնական թիվ է: Դիցուք $r = T^d$: Տրոհենք օբյեկտների $A = [0, 1]^d$ բազմությունը $C_1, C_2, \dots, C_r$ հիպերխորանարդների՝ ϵ երկարության կողերով: Այսինքն՝ ցանկացած $(\alpha_1, \alpha_2, \dots, \alpha_d) \in [T]^d$ ինդեքսի համար գոյություն ունի

$$C_i = \{x : x_j \in [(\alpha_j - 1)/T, \alpha_j/T], \forall j\}$$

տիպի բազմություն:



Նկ. 12.9

Նկար 12.9- ում պատկերված են $d = 2, T = 5$ պարամետրերի դեպքը և այն քառակուսին, որը համապատասխանում է $\alpha = (2, 4)$ ինդեքսին: Եթե x պատկանում է C_i խորանարդին, որը պարունակում է S հանույթից որևէ կետ, ապա

$$\|x - x_{\pi_1(x)}\| \leq \sqrt{\sum_{i=1}^r \epsilon_i^2} = \epsilon\sqrt{d}:$$

Եթե x պատկանում է C_i խորանարդին, որը չի պարունակում է S հանույթից որևէ կետ, ապա

$$\|x - x_{\pi_1(x)}\| \leq \sqrt{d}:$$

Այսպիսով,

$$E\left[\|x - x_{\pi_1(x)}\|\right] = E\left[\epsilon\sqrt{d} \sum_{i=1}^r [C_i \cap S \neq \emptyset][x \in C_i] + \sqrt{d} \sum_{i=1}^r [C_i \cap S = \emptyset][x \in C_i]\right]:$$

Նկատենք, որ

$$\sum_{i=1}^r [C_i \cap S \neq \emptyset][x \in C_i] \in \{0, 1\}$$

(քանի որ միայն մեկ ինդեքսի համար տեղի ունի $x \in C_i$ պայմանը): Հետևաբար՝

$$E\left[\sum_{i=1}^r [C_i \cap S \neq \emptyset][x \in C_i]\right] \leq 1:$$

Քանի որ x օրյեկտը և S հավաքածուն իրարից անկախ են, ապա

$$E\left[[C_i \cap S = \emptyset][x \in C_i]\right] =$$

$$E[C_i \cap S = \emptyset]E[x \in C_i] =$$

$$= P(C_i \cap S = \emptyset)P(x \in C_i):$$

Մյուս կողմից, քանի որ S հանույթը բաղկացած է իրարից անկախ l օրյեկտներից,

ապա

$$P(C_i \cap S = \emptyset) = P(x_1 \notin C_i, x_2 \notin C_2, \dots, x_l \notin C_l) = P(x \notin C_i)^l = (1 - P(x \in C_i))^l \leq e^{-lP(x \in C_i)}:$$

Նշանակելով $p_i = P(x \in C_i)$ ՝ կունենանք՝

$$\begin{aligned} E[||x - x_{\pi_1(x)}||] &\leq \epsilon\sqrt{d} + d \sum_{i=1}^r e^{-lp_i} p_i \leq \\ &\leq \epsilon\sqrt{d} + \sqrt{d} \sum_{i=1}^r \max_{p \in [0,1]} p e^{-pl} = \epsilon\sqrt{d} + r\sqrt{d} \max_{p \in [0,1]} p e^{-pl}: \end{aligned}$$

Քանի որ $f(p) = pe^{-lp}$ ֆունկցիան գոգավոր է, ապա նրա մաքսիմումը հասանելի է $p = 1/l$ կետում: Հետևաբար՝

$$E[||x - x_{\pi_i}||] \leq \epsilon\sqrt{d} + \frac{r\sqrt{d}}{le} = \sqrt{d}(\epsilon + \frac{1}{le}r), \quad r = (\frac{1}{\epsilon})^d:$$

Այս անհավասարության աջ մասը մինիմալ արժեք ընդունում է $\epsilon^* = l^{-1/(d+1)}$ կետում: Տեղադրելով այն անհավասարության աջ մասում՝ հեշտ է նկատել, որ այն փոքր է դառնում

$$\frac{4\sqrt{d}}{l^{1/(d+1)}}:$$

Թեորեմն ապացուցվեց:

Այժմ ենթադրենք, $L_D(h^*) = 0$: Ընտրենք l -ը այնպիսին, որ

$$\frac{4c\sqrt{d}}{l^{1/(d+1)}} < \epsilon,$$

որտեղ $\epsilon > 0$ նախապես տրված ճշտություն է: Այստեղից կստանանք $l \geq (\frac{4c\sqrt{d}}{\epsilon})^{d+1}$: Հետևաբար ուսուցանող S հավաքածուի l երկարությունը էքսպոնենցիալ աճում է տարածության d չափողականության մեծացմանը զուգահեռ, երբ ֆիքսված է հավանականային D բաշխումը:

12.4 Կլաստերավորման տրամաբանական մեթոդներ: Որոշիչ ծառեր

Դիտարկենք քինար ծառ, որի

- յուրաքանչյուր ներքին v գագաթում գործում է $\beta_v : X \rightarrow \{0, 1\}$ ֆունկցիա (պրեդիկատ):
- v գագաթին համապատասխան տերևին վերագրված է $c_v \in Y$ նշիչ (կանխատեսում):

Ենթադրենք՝ ծառի v_0 գագաթից սկսում է աշխատել $a(x)$ ալգորիթմը և հաշվում է β_{v_0} պրեդիկատի արժեքը: Եթե այն հավասար է զրոյի, ապա անցում է կատարվում

«ձախ» գազաթին, հակառակ դեպքում՝ «աջին»: Այս պրոցեսը շարունակվում է մինչև հասնում է տերևային գազաթի: Ալգորիթմը վերադարձնում է այն դասը, որը վերագրված է այդ տերևին: Այս ալգորիթմը կոչվում է **բինար որոշիչ ծառ**: Հաճախ օգտագործվում են միաչափ պրեդիկատներ, որոնք ունեն հետևյալ տեսքը՝ $\beta_v(x, j, t) = [x_j < t]$: Երբեմն էլ օգտագործում են բազմաչափ պրեդիկատներ: Օրինակ՝

- գծային՝ $\beta_v(x) = [(w, x)] < t]$,
- մետրիկական՝ $\beta_v(x) = [\rho(x, x_v)] < t$, որտեղ x_v հավաքածուի որևէ ֆիքսած օբյեկտ է:

Այժմ կոնկրետ օրինակների վրա քննարկենք բինար որոշիչ ծառերի կառուցման որոշ սկզբունքներ: Այնուհետև կանրադառնանք կառուցման ընդհանուր ալգորիթմներին:

Դիտարկենք կլաստերավորման մի օրինակ՝ երկու հարցումներից հետո: Այդ տրամաբանական հարցումների շղթայից հետո տվյալ օբյեկտը դասում են այս կամ այն կլաստերի մեջ:

Դիտարկենք երեք բինար հարաբերություններ:

•

$$f_1(x) = \begin{cases} 0, & \text{չի սիրում Python,} \\ 1, & \text{սիրում է Python:} \end{cases}$$

•

$$f_2(x) = \begin{cases} 0, & \text{չի սիրում «քոն ֆու պանդա»,} \\ 1, & \text{սիրում է «քոն ֆու պանդա» :} \end{cases}$$

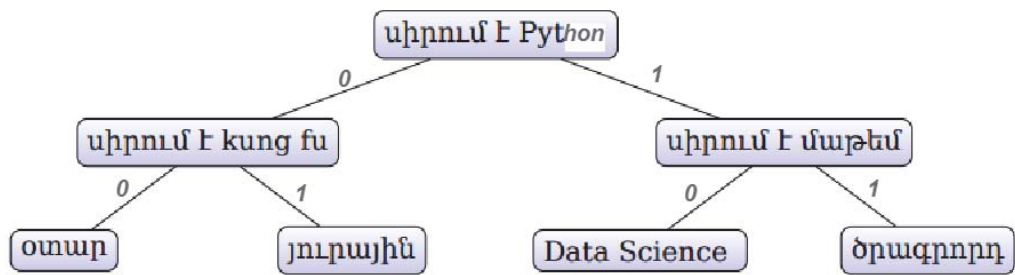
•

$$f_3(x) = \begin{cases} 0, & \text{տանելի չէ մաթեմատիկան,} \\ 1, & \text{տանելի է մաթեմատիկան:} \end{cases}$$

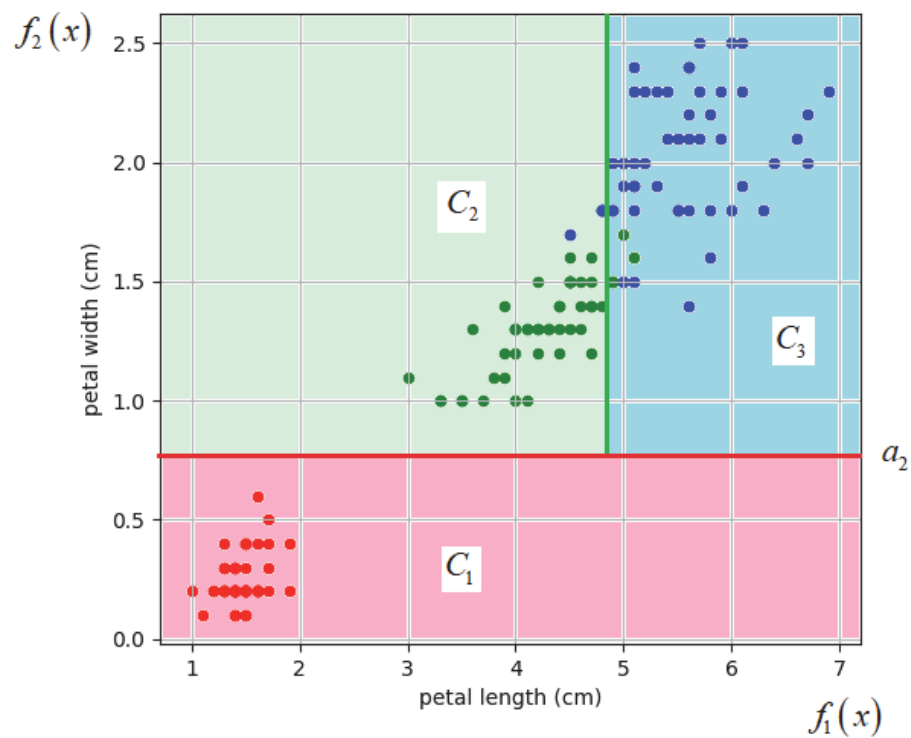
Այս հայտանիշների օգնությամբ ձևակերպվում են տրամաբանական եզրակացություններ (պրեդիկատներ) օրինակ՝

- $B_1(x) = f_1(x) \wedge \overline{f_2(x)} \rightarrow$ «ծրագրորդ»,
- $B_2(x) = \overline{f_1(x)} \wedge f_2(x) \rightarrow$ «յուրային »:

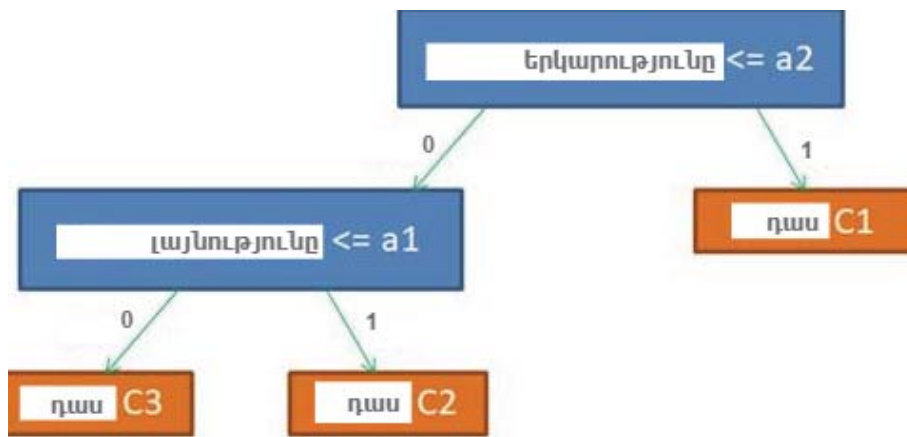
Այսպիսի պրեդիկատների օգնությամբ կառուցվում է բինար ծառ (տե՛ս նկ. 12.10): Նման ծառեր գրականությունում կոչվում են «որոշիչ» ծառեր կամ վճիռների ծառեր: Այժմ դիտարկենք կլաստերավորման ևս մեկ օրինակ՝ ավելի ընդհանուր պրեդիկատների օգնությամբ: Նկար 12.11-ում պատկերված են երեք տիպի (գույնի) բազմություններ:



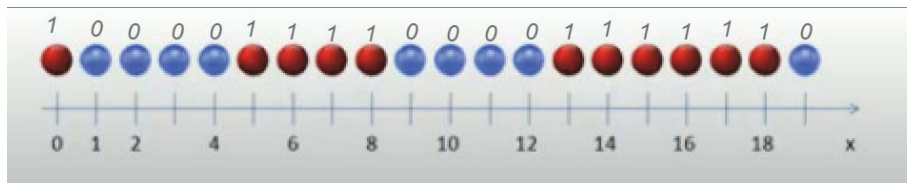
Նկ. 12.10: Կլաստերավորում երկու հարցումից հետո



Նկ. 12.11: Կլաստերավորում պրեդիկատների օգնությամբ



Նկ. 12.12: Բինար ծառի կառուցում պրեդիկատների օգնությամբ



Նկ. 12.13: Որոշիչ ծառի կառուցման պարզ օրինակ

Կատարենք կլաստերավորում երեք R_1, R_2, R_3 պրեդիկատների օգնությամբ: Դիցուք $f_1(x)$ բնութագրիչը լայնությունն է, իսկ $f_2(x)$ -ը՝ բարձրությունը, իսկ պրեդիկատները հետևյալներն են՝

•

$$R_1(x) = [f_2(x) \leq a_2] \rightarrow \begin{cases} 1, & \text{կլաստեր } C_1, \\ 0, & \text{ուրիշ կլաստերներ:} \end{cases}$$

•

$$R_2(x) = \overline{[f_2(x) \leq a_2]} \wedge [f_1(x) \leq a_1] \rightarrow \begin{cases} 1, & \text{կլաստեր } C_2, \\ 0, & \text{ուրիշ կլաստերներ:} \end{cases}$$

•

$$R_3(x) = \overline{[f_2(x) \leq a_2]} \wedge \overline{[f_1(x) \leq a_1]} \rightarrow \begin{cases} 1, & \text{կլաստեր } C_3, \\ 0, & \text{ուրիշ կլաստերներ:} \end{cases}$$

Այս պրեդիկատների օգնությամբ կառուցվում է որոշիչ ծառ, և օբյեկտները խմբավորում են երեք կլաստերներում՝ ինչպես ցույց է տրված նկ.12.12-ում:

Այժմ մի պարզ օրինակի միջոցով ցույց տանք, թե ինչ սկզբունքներով ու հայտանիշներով են ընտրվում պրեդիկատները բինար որոշիչ ծառի կառուցման համար: Պետք է տրված «կապույտ-0» և «կարմիր-1» գնդիկներից բաղկացած հաջորդականությունը բաժանել երկու դասերի՝ «կարմիր-1» և «կապույտ-0» (նկ.12.13): Դիցուք $x = [x_1]$ վեկտորը ունի մի կոմպոնենտ (հատկանիշ), որը գնդիկի համարն է այդ հաջորդականությունում: Բինար ծառի յուրաքանչյուր գագաթում գործելու է հետևյալ պրեդիկատը՝ $B_{1,t}(x) = [x_1 \leq t]$: Նշված հաջորդականությունը տրվում է բինար ծառի արմատին և այս պրեդիկատի օգնությամբ այն բաժանվում

է երկու ենթահաջորդականությունների՝ ձախ և աջ: Բայց նախապես պետք է տալ t շեմ: Այնուհետև յուրաքանչյուր միջանկյալ գագաթում պետք է նույն տիպի պրեդիկատ ինչ-որ շեմով և այսպես շարունակ: Օրինակ՝ որպես բաժանման չափանիշ վերցնենք այն շեմը, որի բաժանման արդյունքում մի մասում լինեն նույն գույնի ավելի շատ գնդիկներ: Եվ այս սկզբունքը կիրառվում է յուրաքանչյուր միջանկյալ գագաթի համար: Օգտագործելով այս սկզբունքը՝ կրճատվում է որոշիչ ծառի խորությունը:

Կա ես մի կարևոր սկզբունք՝ չափանիշ: Դա Էնտրոպիան է: Էնտրոպիան ընդհանրապես արտահայտում է համակարգի քառայնության չափանիշը: Այսինքն՝ ինչքան շատ է բազմազանությունը, այնքան մեծ է Էնտրոպիան: Էնտրոպիան արտահայտվում է հետևյալ բանաձևով՝

$$S = - \sum_{i=1}^N p_i \log_2 p_i,$$

որտեղ N -ը դասերի քանակն է, մեր օրինակում $N = 2$ -ի, իսկ p_1 -ը «կարմիր» գնդիկի հանդես գալու հավանականությունն է, p_2 -ը՝ «սպիտակ» գնդիկի հանդես գալու հավանականությունն է խմբում: Մեր օրինակում

$$p_1 = \frac{11}{20}, p_2 = \frac{9}{20}:$$

Հետևաբար Էնտրոպիան (impurity) վերը նշված հաջորդականության համար հավասար է՝

$$S_0 = -p_1 \log_2 p_1 - p_2 \log_2 p_2 = 0.993:$$

Մեր օրինակում որպես բաժանման շեմ սկզբում վերցված է $t = 12$: Հաշվենք ամեն մի խմբում Էնտրոպիան.

$$S_1 = -\frac{5}{13} \log_2 \frac{5}{13} - \frac{8}{13} \log_2 \frac{8}{13} = 0.96,$$

$$S_2 = -\frac{6}{7} \log_2 \frac{6}{7} - \frac{1}{7} \log_2 \frac{1}{7} = 0.6:$$

Ինչպես նկատելի է, երկրորդ ենթահաջորդականությունում Էնտրոպիան շատ ավելի փոքր է, քան սկզբնական հաջորդականությունում: Դա հենց այն սկզբունքն է, որով առաջնորդվելով՝ մենք պիտի ընտրենք բաժանման շեմը: Այժմ մենք ունենք հաջորդականությունը բնորոշող երեք բնութագրիչներ՝ S_0, S_1, S_2 : Ելնելով այս երեք բնութագրիչներից՝ մենք պետք է ձևավորենք ընդհանուր մի կրիտերիա, որով կրնոթագրվի հաջորդականության՝ երկու մասերի բաժանման «որակը»:

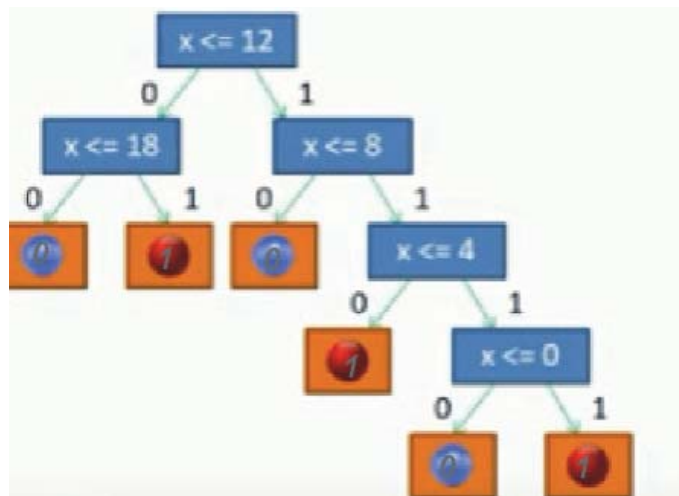
Դրա համար հաճախ օգտագործվում է հետևյալ քանաձևը, որը կոչվում է ինֆորմացիոն ձեռքբերում (information gain).

$$IG(Q(t)) = S_0 - \sum_{i=1}^2 \frac{N_i}{N} S_i,$$

որտեղ N_i -ը i -րդ խմբի էլեմենտների քանակն է: Մեր օրինակի համար ունենք

$$IG(x \leq 12) = S_0 - \frac{13}{20}S_1 - \frac{7}{20}S_2 = 0.16:$$

Եթե մենք հաշվենք այս ինֆորմացիոն ձեռքբերումը ուրիշ շեմերի համար, ապա կտեսնենք, որ այն կլինի ավելի փոքր, քան $t = 12$ -ի համար (նկ.12.14):



Նկ. 12.14: Բինար որոշիչ ծառի օրինակ

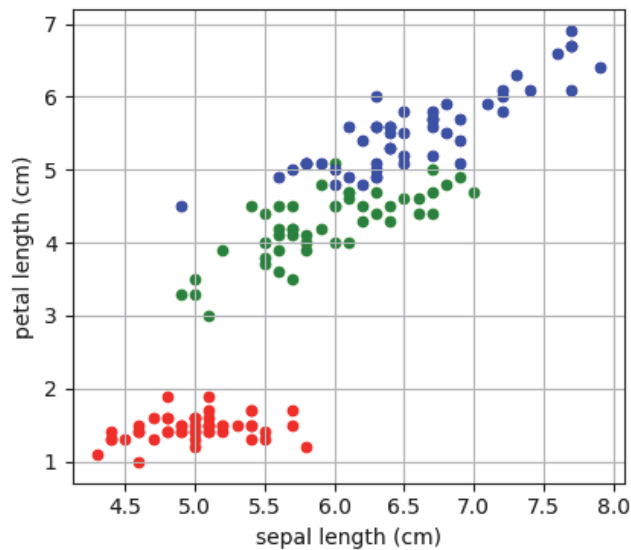
Այսպիսով, յուրաքանչյուր ընթացիկ գազաթի համար օպտիմալ շեմի ընտրության համար լուծվում է հետևյալ խնդիրը՝ $t^* = \arg \max_{t \in T} IG(t)$:

Բինար ծառի կառուցումը ավարտավում է, երբ *impurity*-ին փոքր է նախապես տրված ճշտությունից կամ երբ ֆիքսվում է ծառի խորության չափը:

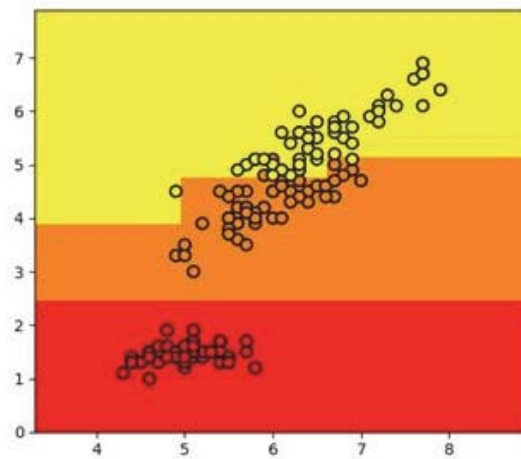
Այժմ բերենք որոշիչ ծառի կառուցման ևս մի օրինակ: Դիցուք ունենք երեք տիպի օբյեկտներ: Յուրաքանչյուրից հիսունը դասավորված է այնպես ինչպես պատկերված է նկար 12.15-ում:

Այդ օբյեկտները բնորոշվում են երկու բնութագրիչով, որոնք դրանց կոորդինատներն են: Սկսած գազաթից՝ մենք գտնում ենք այն հայտանիշը՝ j -ն, և շեմը՝ t -ն, որոնց համար տեղի ունի $(j^*, t^*) = \arg \max_{j \in J, t \in T} Q(R_m, j, t)$ հավասարությունը: Այնուհետև այս քայլերը կրկնում ենք յուրաքանչյուր հաջորդ միջանկյալ գազաթի համար: Համարում ենք, որ ծառի մաքսիմալ խորությունը պետք է լինի հավասար չորսի: Նկար 12.17-ում պատկերված է ալգորիթմի աշխատանքի արդյունքում ձևավորված բինար ծառը:

Ակնհայտ երևում է, որ սկզբնական քայլում բաժանումը կատարվում է $B_{j,t}(x) = [x_1 \leq 2.45]$ պրեդիկատի օգնությամբ: Արդյունքում ձախ մասում ընկնում է հիսուն «կարմիր» օբյեկտ գրոյական էնտրոպիայով, և հետևաբար այլևս այստեղից բաժանում չի կատարվում, այսինքն՝ այս գազաթը համարվում է ծառի տերև:

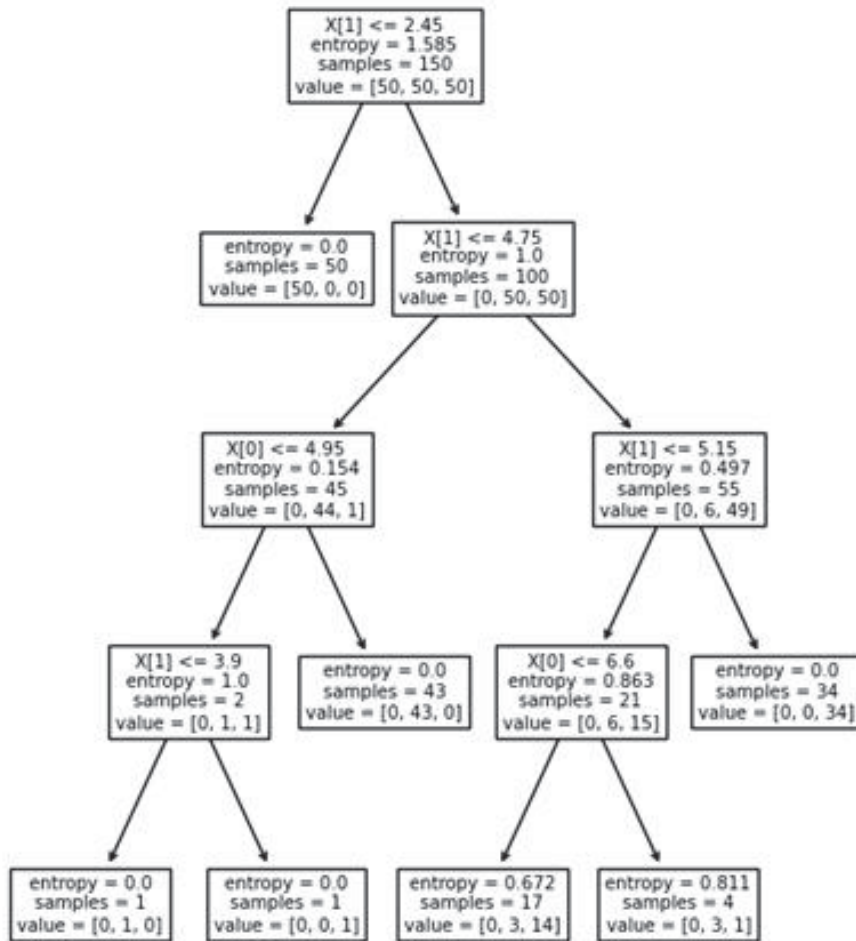


Նկ. 12.15: Օբյեկտների դասավորվածության տիրույթը



Նկ. 12.16: Օբյեկտների բաժանումը ըստ դասերի

Աջ գագաթում ընկնում են մնացած հարյուր օբյեկտները, որտեղ Էնտրոպիան հավասար է մեկի: Այնուհետև մենք օբյեկտները բաժանում ենք ըստ երկրորդ հայտանիշի, իսկ շեմ վերցրած է 4.75: Ձևավորվում են հաջորդ երկու գագաթները և այսպես շարունակ, մինչև չի ձևավորվում գրոյական Էնտրոպիա, կամ ծառի խորությունը դառնում է չորս: Նկար 12.16-ում կատարված է օբյեկտների տրոհում ըստ կլաստերների:



Նկ. 12.17: Բինար որոշիչ ծառի օրինակ

12.5 Որոշիչ (վճիռների) ծառեր գծային ռեգրեսիայի և դասակարգման խնդիրներում

Այժմ քննարկենք այն ընդհանուր սկզբունքը, որով կատարվում է ինֆորմացիայի տրոհում որոշիչ ծառերի գազաթններում: Այնհույսեաա այն մասնավորացնենք գծային ռեգրեսիայի խնդիրների համար:

Ենթադրենք՝ որոշիչ ծառի ինչ-որ գազաթին տրվում է տվյալների X_m հավաքածու, և պիտի կատարվի օպտիմալ բաժանում՝

$$X_m = X_l \cup X_r:$$

Դա կատարվում է հետևյալ սկզբունքով: Դիցուք $L(y, c)$ -ը որևէ կորստի ֆունկցիա է: c հաստատունը վերցնում են այնպես, որը մինիմիզացնի միջին կորուստը՝

$$\frac{1}{|X_m|} \sum_{(x_i, y_i) \in X_m} L(y_i, c) \rightarrow \min:$$

Օպտիմալ

$$H(X_m) = \min_{c \in Y} \frac{1}{|X_m|} \sum_{(x_i, y_i) \in X_m} L(y_i, c)$$

արժեքը կոչվում է *impurity*:

Իսկ

$$H(X_m) - \frac{|X_l|}{|X_m|} H(X_l) - \frac{|X_r|}{|X_m|} H(X_r)$$

մեծությունը կոչվում է *information gain*: Բաժանման կրիտերիան կլինի մինիմիզացնել այդ մեծությունը, կամ որ նույնն է՝ լուծել հետևյալ խնդիրը՝

$$\text{branch}(X_m, j, t) = |X_m| H(X_m) - |X_l| H(X_l) - |X_r| H(X_r) \rightarrow \min_{j, t}$$

Այժմ քննարկենք՝ թե ինչպիսին կլինի այդ կրիտերիան կոնկրետ խնդիրների դեպքում:

Դիցուք ունենք ռեգրեսիայի խնդիրը, որտեղ կորստի ֆունկցիան միջին քառակուսային շեղումն է՝ $L(y, c) = (y - c)^2$: Այդ դեպքում տերևի *impurity*-ին արտահայտվում է հետևյալ կերպ՝

$$H(X_m) = \frac{1}{|X_m|} \min_c \sum_{(x_i, y_i) \in X_m} (y_i - c)^2:$$

Պարզ է, որ այս խնդրի մինիմումի կետն է

$$c = \frac{\sum_{i: (x_i, y_i) \in X_m} y_i}{|X_m|}:$$

Տեղադրելով այն *impurity* ֆունկցիոնալի մեջ՝ ստանում ենք

$$H(X_m) = \sum_{(x_i, y_i) \in X_m} \frac{(y_i - \bar{y})^2}{|X_m|},$$

որտեղ

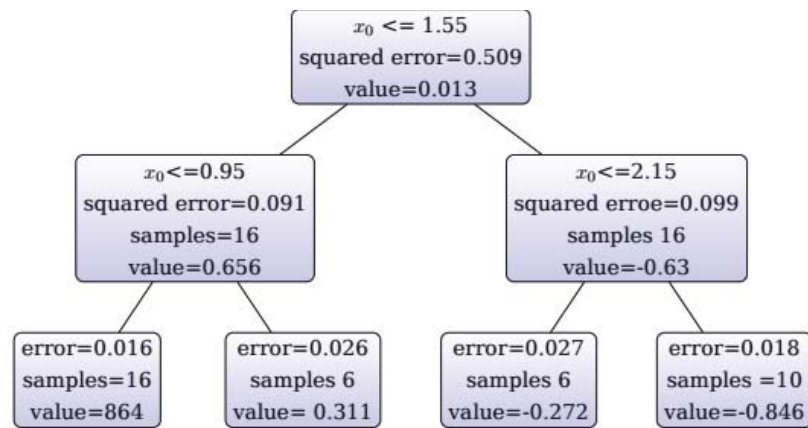
$$\bar{y} = \frac{1}{|X_m|} \sum_{i: (x_i, y_i) \in X_m} y_i:$$

Այսպիսով, $\text{branch}(X_m, j, t)$ բանաձևի համար վերջնականում կստանանք հետևյալ բանաձևը՝

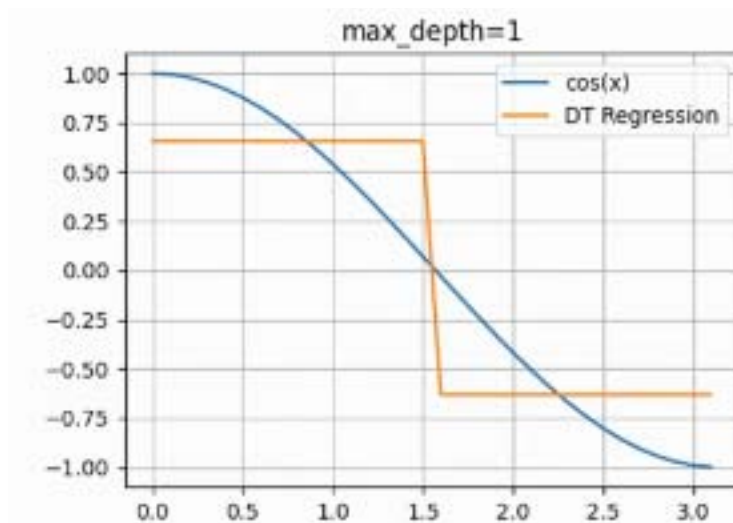
$$\begin{aligned} \text{branch}(X_m, j, t) &= \\ &= \sum_{(x_i, y_i) \in X_m} (y_i - \bar{y})^2 - \sum_{(x_i, y_i) \in X_l} (y_i - \bar{y})^2 - \sum_{(x_i, y_i) \in X_r} (y_i - \bar{y})^2: \end{aligned}$$

Ուստի պետք է կատարել այնպիսի բաժանում, որ աջ և ձախ տերևների պատասխանների դիսպերսիաների գումարը լինի մինիմում:

Դիտարկենք կոնկրետ օրինակ: Դիցուք ունենք ուսուցանող $X = \{(x_i, y_i = \cos(x_i)), i = 1, 2, \dots, l\}$ հավաքածու, որը $y = \cos x$ ֆունկցիայի գրաֆիկի կետեր է (նկ.12.18):



Նկ. 12.18: $\cos(x)$ ֆունկցիան մոտարկող որոշիչ ծառ



Նկ. 12.19: $\cos(x)$ ֆունկցիան մոտարկող երկու տերևանի որոշիչ ծառ

Եթե որոշիչ ծառի խորությունը մեկ է, այսինքն՝ ունի երկու տերև, ապա նկար 12.19-ում պատկերված է այդպիսի ծառով $y = \cos(x)$ ֆունկցիայի գրաֆիկի մոտարկումը: Իսկ նկար 12.20-ից երևում է, որ չորս տերևով որոշիչ ծառը ավելի ճշգրիտ է մոտարկում այդ ֆունկցիայի գրաֆիկը:

Այժմ դիտարկենք *impurity*-ի որոշման հարցը դասակարգման խնդիրներում: Դիցուք ունենք K դասեր, և ընթացիկ X_m գազաթում k դասի էլեմենտների մասնաբաժինը հավասար է p_k -ի, այսինքն՝

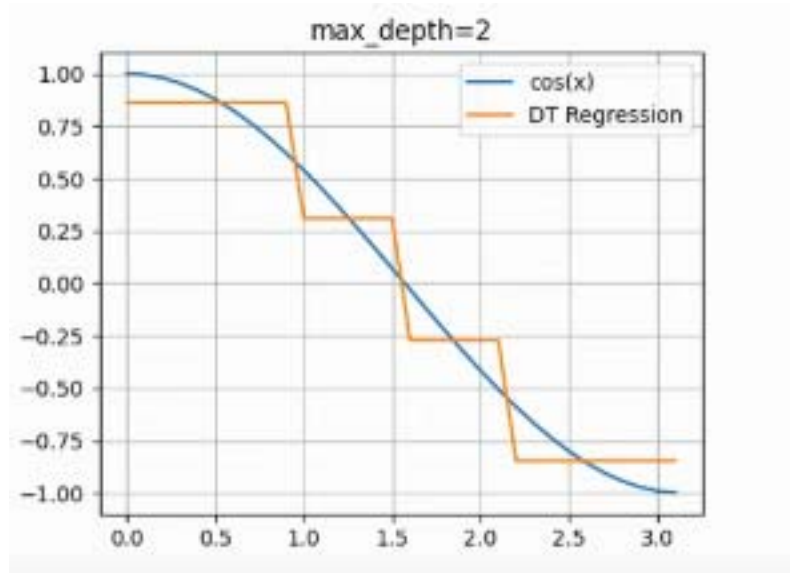
$$p_k = \frac{1}{|X_m|} \sum_{(x_i, y_i) \in X_m} [y_i = k]:$$

Այստեղ որպես կորստի ֆունկցիա հանդես է գալիս հետևյալը՝

$$L(y, c) = [y \neq c]:$$

Այդ դեպքում այդ տերևի վրա *impurity*-ին հավասար կլինի՝

$$H(X_m) = \min_{c \in Y} \frac{1}{|X_m|} \sum_{(x_i, y_i) \in X_m} [y_i \neq c]:$$



Նկ. 12.20: $\cos(x)$ ֆունկցիան մոտարկող չորս տերևանի որոշիչ ծառ

Ենթադրենք՝ դիտարկվող տերևի վրա ամենամեծ հաճախությունը ունի k^* դասը: Այդ դեպքում *impurity*-ն կարտահայտվի հետևյալ բանաձևով՝

$$H(X_m) = \frac{1}{|X_m|} \sum_{(x_i, y_i) \in X_m} [y_i = k^*] = 1 - p_{k^*}:$$

Այս մեծությունը կոչվում է *misclassification error*:

Այժմ դիտարկենք կորստի լոգարիթմական հետևյալ ֆունկցիան՝

$$H(X_m) = \min_{\sum_k c_k = 1} \left(-\frac{1}{|H_m|} \sum_{(x_i, y_i) \in X_m} \sum_{k=1}^K [y_i = k] \ln c_k \right):$$

Օպտիմալ c_k գործակիցները գտնելու համար կիրառենք Լագրանժի անորոշ գործակիցների մեթոդը: Այսինքն՝ պետք է լուծել մինիմիզացիայի հետևյալ խնդիրը՝

$$\Lambda(c, \lambda) \equiv -\frac{1}{|H_m|} \sum_{(x_i, y_i) \in X_m} \sum_{k=1}^K [y_i = k] \ln c_k + \lambda \sum_{k=1}^K c_k \rightarrow \min_{c_k}:$$

Դիֆերենցելով Լագրանժի ֆունկցիան ըստ c_k փոփոխականների՝ կստանանք

$$\frac{\partial}{\partial c_k} \Lambda(c, \lambda) = -\frac{1}{|X_m|} \sum_{(x_i, y_i) \in H_m} [y_i = k] \frac{1}{c_k} + \lambda = 0 = -\frac{p_k}{c_k} + \lambda = 0:$$

Այստեղից, $c_k = p_k / \lambda$: Գումարելով այս հավասարությունները ըստ k -ի՝ կստանանք

$$1 = \sum_{k=1}^K c_k = \frac{1}{\lambda} \sum_{k=1}^K p_k = \frac{1}{\lambda}:$$

Այսինքն՝ $\lambda = 1$, և հետևաբար մինիմումը հասանելի է $c_k = p_k$ արժեքների դեպքում:

Տեղադրելով այս արժեքները *impurity*-ի բանաձևի մեջ՝ կստանանք

$$H(X_m) = - \sum_{k=1}^K p_k \ln(1 - p_k),$$

որը Էնտրոպիայի հայտնի բանաձևն է:

12.6 Կլաստերավորման GMM մոդելը (gaussian mixture model)

Այս պարագրաֆում օրինակի միջոցով քննարկելու ենք օբյեկտների կլաստերավորման խնդիրը, երբ մոդելները ունեն գաուսյան նորմալ բաշխումներ:

Դիցուք ունենք տվյալների հավաքածու՝

$$X = \{0, 1, 2, 2.5, 3, 4, 4.5, 5, 6, 7, 8, 10, 12\} \text{ (տե՛ս նկ. 12.21):}$$

Կիրառելով k -միջինների ալգորիթմը $k = 2$ -ի դեպքում՝ հավաքածուն կբաժանի երկու կլաստերների, ընդ որում առաջինում կհայտնվեն $\{0, 1, 2, 2.5, 3, 4, 4.5, 5\}$ թվերը, իսկ երկրորդում՝ $\{6, 7, 8, 10, 12\}$: Այժմ Մեր հավաքածուն բաժանենք երկու խմբերի, որոնցում կետերը նորմալ են բաշխված: Այսինքն՝

$$Class_0 \sim N(\mu_1, \sigma_1),$$

$$Class_1 \sim N(\mu_2, \sigma_2) :$$

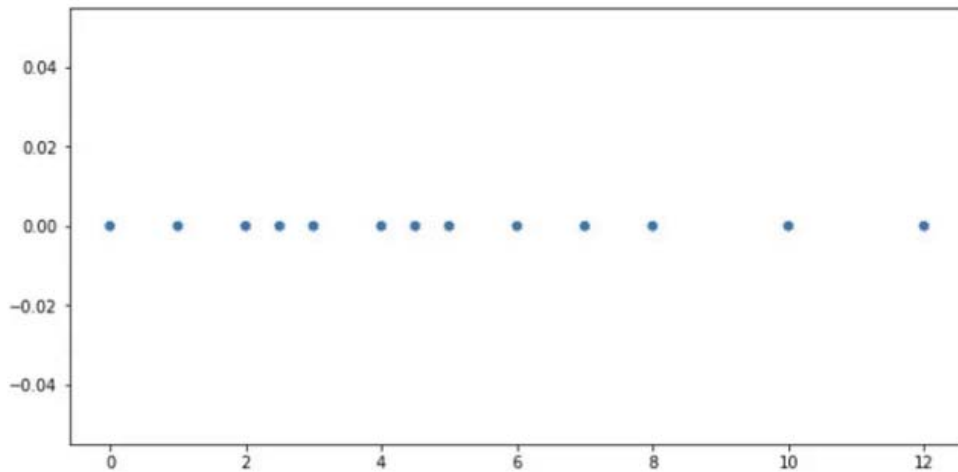
Մենք ուզում ենք կառուցել նոր մոդել, որը կոչվում է Gaussian mixture model(GMM): Դիցուք պարամետրերին տանք սկզբնական պատահական արժեքներ: Օրինակ՝ $\mu_1 = 0, \sigma_1 = 1, \mu_2 = 1, \sigma_2 = 1$: Այնուհետև նկարենք այդ պարամետրերով գաուսյանների գրաֆիկները (տե՛ս նկ.12.22): Ըստ այդ բաշխումների՝ հաշվենք $p(X = x|Class_0)$, $p(X = x|Class_2)$ պայմանական հավանականությունները ամբողջ հավաքածուի համար: Այնուհետև Բայեսի բանաձևով հաշվարկենք տվյալ օբյեկտը այս կամ այն դասին պատկանելու հավանականությունները՝

$$p(Class_0|x) = \frac{p(x|Class_0)p(Class_0)}{p(x|Class_0)p(Class_0) + p(x|Class_1)p(Class_1)},$$

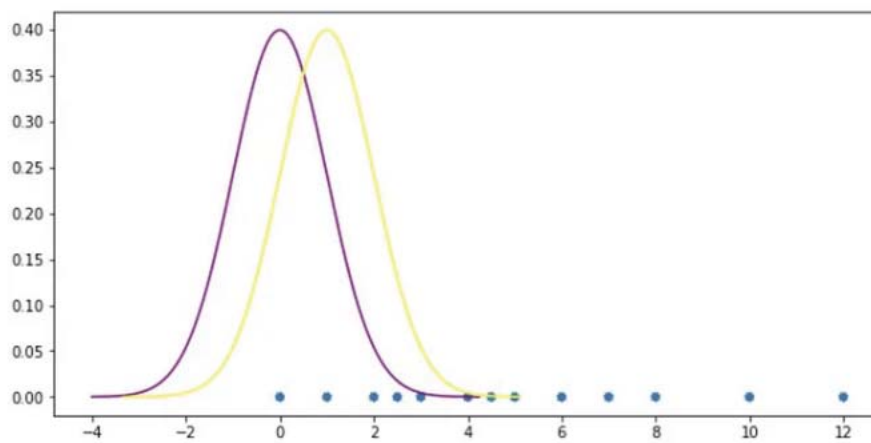
$$p(Class_1|x) = \frac{p(x|Class_1)p(Class_1)}{p(x|Class_0)p(Class_0) + p(x|Class_1)p(Class_1)} :$$

Նախօրոք ենթադրենք, որ $p(Class_0) = p(Class_1) = \frac{1}{2}$:

Այսպիսով, $Class_0$ -ի համար գաուսյան բաշխման սկզբնական ֆունկցիան ունի $\frac{1}{2\pi}e^{-\frac{x^2}{2}}$ տեսքը: Իսկ երկրորդ դասի համար այն արտահայտվում է հետևյալ բանաձևով՝ $\frac{1}{2\pi}e^{-\frac{(x-1)^2}{2}}$:



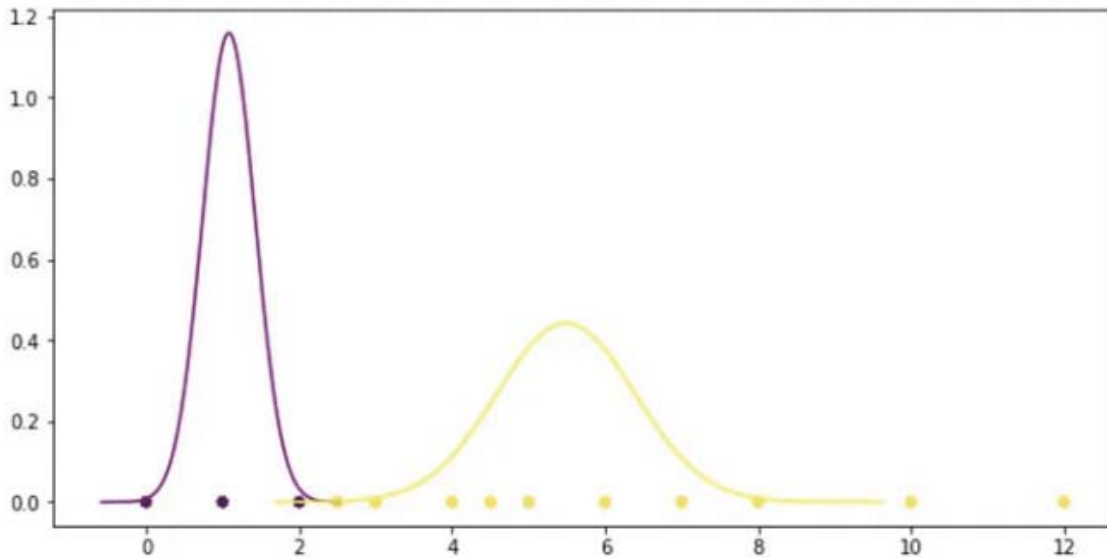
Նկ. 12.21: Կետերի հավաքածու



Նկ. 12.22: Գաուսյանների գրաֆիկները

	samples	p(X=x class0)	p(X=x class1)	p(class0 x)	p(class1 x)	p(class0 x)_standardized	p(class1 x)_standardized
0	0.0	3.990434e-01	2.420321e-01	0.622459	0.377541	0.431670	0.032665
1	1.0	2.420321e-01	3.990434e-01	0.377541	0.622459	0.261821	0.053855
2	3.0	4.432972e-03	5.400466e-02	0.075858	0.924142	0.052607	0.079957
3	4.0	1.338642e-04	4.432972e-03	0.029312	0.970688	0.020328	0.083984
4	5.0	1.487097e-06	1.338642e-04	0.010987	0.989013	0.007619	0.085569
5	6.0	6.077424e-09	1.487097e-06	0.004070	0.995930	0.002823	0.086168
6	7.0	9.137037e-12	6.077424e-09	0.001501	0.998499	0.001041	0.086390
7	2.0	5.400466e-02	2.420321e-01	0.182426	0.817574	0.126510	0.070737
8	2.5	1.753275e-02	1.295504e-01	0.119203	0.880797	0.082666	0.076207
9	4.5	1.598779e-05	8.729040e-04	0.017986	0.982014	0.012473	0.084964
10	8.0	5.053552e-15	9.137037e-12	0.000553	0.999447	0.000383	0.086472
11	10.0	7.696550e-23	1.028238e-18	0.000075	0.999925	0.000052	0.086514
12	12.0	2.146928e-32	2.119357e-27	0.000010	0.999990	0.000007	0.086519

Նկ. 12.23: Հավանականությունների աղյուսակ



Նկ. 12.24: GMM-ի գրաֆիկը մեկ քայլից հետո

12.23 աղյուսակի վերջին երկու սյուները հաշվված են հետևյալ բանաձևերով՝

$$p(Class_k|x_i)_{standardized} = \frac{p(Class_k|x_i)}{\sum_{i=1}^{13} p(Class_k|x_i)}, \quad k = 0, 1 :$$

Հաջորդ քայլում կատարում ենք $\mu_1, \sigma_1, \mu_2, \sigma_2$ պարամետրերի արժեքների թարմացում պարագրաֆ 8.5-ում նկարագրված EM ալգորիթմով՝ օգտագործելով այնտեղ ստացված (8.3) բանաձևը:

$$\mu_1^{new} = \sum_{i=1}^{13} x_i p(class_0|x_i)_{standardized} =$$

$$0 \times 0.43 + 1 \times 0.26 + \dots 12 \times 12 \times 0.000007 = 1.08,$$

$$variance_1^{new} = \sum_{i=1}^{13} ((x_i - \mu_1^{new})^2 p(Class_0|x_i)_{standardized}) =$$

$$(0 - 1.08)^2 \times 0.43 + (1 - 1.08)^2 \times 0.026 + \dots (12 - 1.08)^2 \times 0.000007 = 1.54,$$

$$\sigma_1^{new} = \sqrt{variance_1^{new}/13} = 0.34:$$

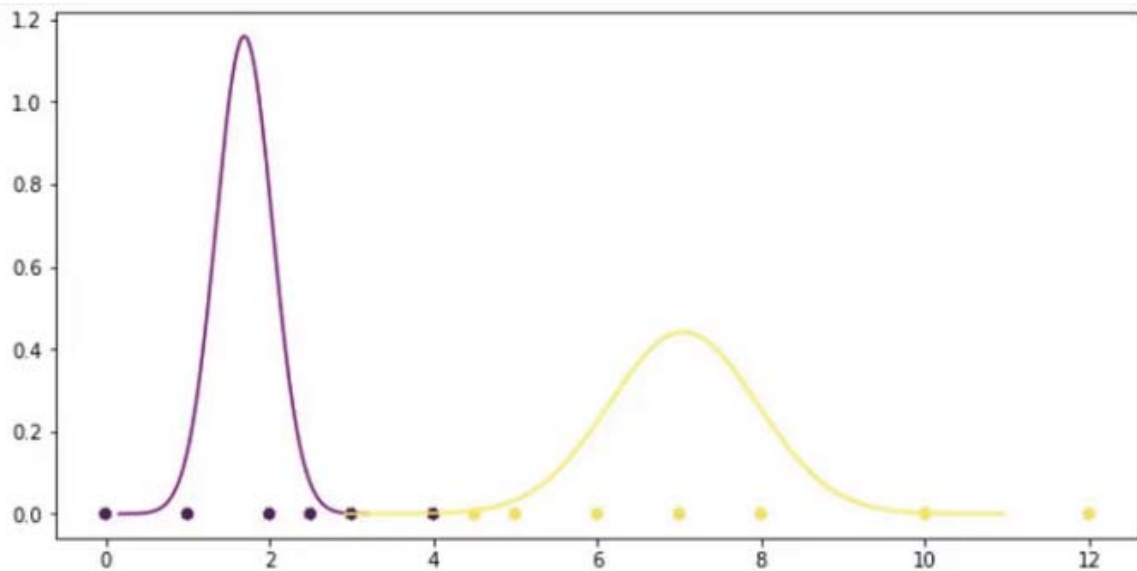
Նույն կերպ կատանանք μ_2, σ_2 պարամետրերի նոր արժեքներ՝ $\mu_2 = 5.48, \sigma_2 = 0.9$:

Այսպես շարունակելով՝ տասը իտերացիայից հետո GMM-ը կունենա նկար 12.25-ում պատկերված տեսքը: Այստեղ

$$Class_0 \sim N(1.69, 0.34),$$

$$Class_1 \sim N(7.04, 0.9):$$

Հաշվարկները կատարված են Python-ով:



Նկ. 12.25: GMM-ի գրաֆիկը տասը քայլից հետո

Այսպիսով, *GMM* ալգորիթմը հավաքածուն բաժանում է երկու կլաստերների՝

$$Class_0 = \{0, 1, 2, 2.5, 3, 4\},$$

$$Class_1 = \{4.5, 5, 6, 7, 8, 10, 12\} :$$

Ի տարբերություն *k*-միջինների ալգորիթմի այս դեպքում 4, 4.5 օբյեկտները դասվել են *Class*₁-ում:

12.7 Apriori ալգորիթմը և ասոցիատիվ կանոններ

Ասոցիացիայի կանոնների իմացությունը մեքենայական ուսուցման կարևոր բաղադրիչներից է: Այն օգտագործվում է շուկայական գամբյուղը վերլուծելիս և կայքերի այցելումների վիճակագրություն կատարելիս: Շուկայական գամբյուղի վերլուծությունը մեթոդ է, որը կիրառվում է խոշոր մանրածախ ցանցերում՝ ապրանքների միջև նույնանուն կապերը բացահայտելու համար: Օրինակ՝ սուպերմարկետում, բոլոր գնված ապրանքները դրված են կողք կողքի: Սուպերմարկետի խնդիրն է որոշել, թե հաճախ որ ապրանքներն են գնվում միասին, տեղադրել դրանք կողքի կողքի՝ վաճառքը ավելացնելու համար:

Apriori-ի ալգորիթմի աշխատանքի սկզբունքները ներկայացնենք օրինակի միջոցով: Ենթադրենք ունենք տվյալների հավաքածու, որը ներկայացված է նկ. 12. 26-ում: Այդտեղ կան *A, B, C, D, E* տիպի ապրանքներ: Աղյուսակի առաջին տողը ցույց է տալիս, որ գնվել են *A, B* ապրանքները նշված հերթականությամբ: Երկրորդ անգամ՝ *B, D* ապրանքները և այսպես շարունակ: Աղյուսակում ներկայացված են իննը նման վիճակագրական տվյալներ: Անվանենք այն շուկայական գամբյուղ:

TID	ITEMSETS
T1	A, B
T2	B, D
T3	B, C
T4	A, B, D
T5	A, C
T6	B, C
T7	A, C
T8	A, B, C, E
T9	A, B, C

Նկ. 12.26: Տվյալների բազա

Յուրաքանչյուր ապրանքի համար սահմանենք մի չափանիշ, որը կոչվում է *support*: Այն ցուց է տալիս թե տվյալ ապրանքը քանի անգամ է հայտնվել գամբյուղում: Օրինակ, *A* ապրանքը կա *T1, T4, T5, T7, T8, T9* դարակներում: Հետևաբար՝ $support(A) = 6$: Այսպես հաշվելով յուրաքանչյուր ապրանքի *support* չափանիշը՝ այն ներկայացնենք աղյուսակի տեսքով (նկ. 12.27):

Itemset	Support_Count
A	6
B	7
C	5
D	2
E	1

Նկ. 12.27: *support* չափանիշներ

Ենթադրենք *support* չափանիշի մինիմալ սահմանը պետք է լինի երկուսը: Այդ դեպքում աղյուսակից պետք է հեռացվի *E* ապրանքը, և ձևավորվի նոր աղյուսակ՝ (տե՛ս նկ.12.28):

Itemset	Support_Count
A	6
B	7
C	5
D	2

Նկ. 12.28: *support* չափանիշների նոր աղյուսակ

Այնուհետև կառուցվում է իրերի հավաքածուների (itemsets) նոր աղյուսակ՝ նկ.12.29:

Itemset	Support_Count
{A, B}	4
{A,C}	4
{A, D}	1
{B, C}	4
{B, D}	2
{C, D}	0

Նկ. 12.29: Itemsets

Այս աղյուսակում ներկայացված են նաև հավաքածու ապրանքների *support* չափանիշները:

Օրինակ, $support(A, B) = 4$, որովհետև այս զույգ հերթականությունն կա $T1, T3, T8, T9$ դարակներում (տե՛ս նկ. 12.29-ը): Այս աղյուսակից հեռացնենք (A, D) , (C, D) զույգերը, որովհետև դրանց *support*-ները փոքր են մինիմալ չափանիշից (տե՛ս նկ. 12.30): Դրանից հետո կազմվում է եռյակ իրերի հավաքածուների աղյուսակ, և հաշվում են դրանց *support* չափանիշները (նկ.12.31):

Itemset	Support_Count
{A, B}	4
{A, C}	4
{B, C}	4
{B, D}	2

Նկ. 12.30: Itemsets

Itemset	Support_Count
{A, B, C}	2
{B, C, D}	1
{A, C, D}	0
{A, B, D}	0

Նկ. 12.31: Եռյակ Itemsets և նրանց *support* չափանիշները

Այստեղից՝ միայն $\{A, B, C\}$ եռյակն է բավարարում մինիմալ չափանիշը: Սրանից հետո այս եռյակի միջոցով կազմվում են բոլոր հնարավոր ասոցիատիվ կանոնները A, B, C ապրանքների միջև և հաշվում, այսպես ասած, այդ կանոնների վստահության (*confidence*) չափանիշը, և ընտրվում են այն կանոնները, որոնց համար այդ չափանիշը մեծ կամ հավասար է նախօրոք դրված մինիմալ սահմանից: Ներկայացնենք այդ գաղափարը օրինակով: $A \wedge B \rightarrow C$ գրառությունը

նշանակում է, որ եթե գնվել են A, B ապրանքները, ապա կգնվի նաև C ապրանքը: Այս կանոնը կոչվում է ասոցիատիվ կանոն, որի համար սահմանվում է վստահության(confidence) չափանիշ հետևյալ բանաձևով`

$$confidence(A \wedge B \rightarrow C) = \frac{support(A, B, C)}{support(A, B)}:$$

Այս չափանիշը իրականում C ապրանքի գնման հավանականությունն է (A, B) զույգի գնման դեպքում` նշված հերթականությամբ: Եվ ընդհանրապես`

$$confidence(X \rightarrow Y) = \frac{support(X, Y)}{support(X)}:$$

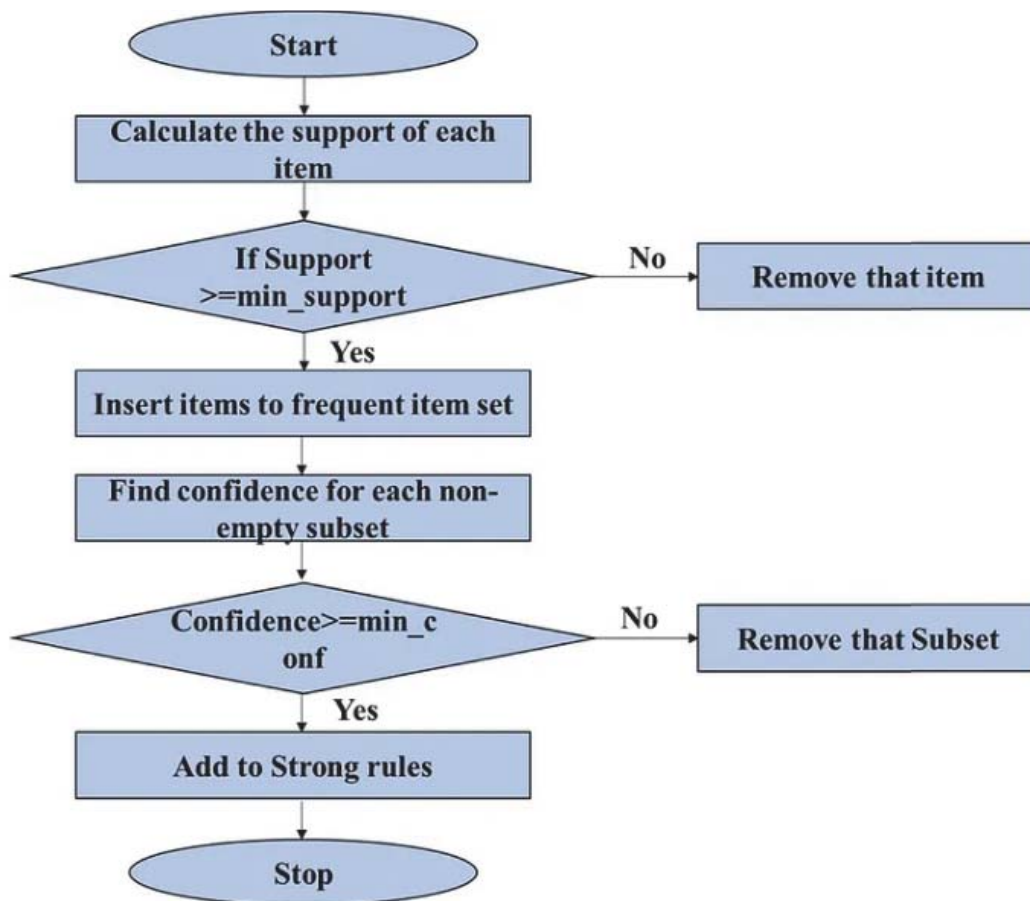
Մեր օրինակում մինիմալ սահմանը 50 տոկոս է: Այդ պայմանը բավարարում են հետևյալ երեք կանոնները`

$$A \wedge B \rightarrow C, B \wedge C \rightarrow A, A \wedge C \rightarrow B:$$

Նկ.12.33 -ում ներկայացված է *apriori*-ի ալգորիթմի բլոկ սխեման:

Rules	Support	Confidence
$A \wedge B \rightarrow C$	2	$Sup\{(A \wedge B) \wedge C\}/sup(A \wedge B)=2/4=0.5=50\%$
$B \wedge C \rightarrow A$	2	$Sup\{(B \wedge C) \wedge A\}/sup(B \wedge C)=2/4=0.5=50\%$
$A \wedge C \rightarrow B$	2	$Sup\{(A \wedge C) \wedge B\}/sup(A \wedge C)=2/4=0.5=50\%$
$C \rightarrow A \wedge B$	2	$Sup\{(C \wedge (A \wedge B))\}/sup(C)=2/5=0.4=40\%$
$A \rightarrow B \wedge C$	2	$Sup\{(A \wedge (B \wedge C))\}/sup(A)=2/6=0.33=33.33\%$
$B \rightarrow B \wedge C$	2	$Sup\{(B \wedge (B \wedge C))\}/sup(B)=2/7=0.28=28\%$

Նկ. 12.32: Ասոցիատիվ կանոնները



Նկ. 12.33: Ապրիորի ալգորիթմի բլոկ սխեման

Առաջադրանք: Դիցուք տրված է ուսուցանող հավաքածու. տվյալները ներկայացված են նկ.12.34-ում: Աղյուսակում կան չորս օբյեկտներ, որոնք ունեն A, B, C հատկանիշներ: Պահանջվում է՝

- կառուցել որոշիչ ծառ որպես ինֆորմատիվության չափանիշ ընտրելով Ջինի ինդեքսը, որը որոշվում է հետևյալ առնչությամբ՝

$$Gini(x) = p(x=0)p(Y=0|x=0)p(Y=1|x=0) + p(x=1)p(Y=0|x=1)p(Y=1|x=1),$$

- կառուցված ծառի միջոցով դասակարգել նոր օբյեկտներ, օրինակ՝ օբյեկտ, որը տրված է նկ. 12.35-ում:

Աղյուսակ					
N	A	B	C	Y	
1	1	1	1	0	
2	1	1	0	1	
3	1	1	1	1	
4	0	0	0	1	

Նկ. 12.34: Չորս օբյեկտներ A, B, C հատկանիշներով

Աղյուսակ					
N	A	B	C	Y	
k	1	0	1	1	

Նկ. 12.35: Դասակարգել k օբյեկտը

13

Մոդելների գնահատման չափանիշներ

13.1 Խաչաձև ստուգում (k-fold cross validation)

Կոնկրետ խնդիրների մոդելների կառուցման ժամանակ հարց է առաջանում թե որ մոդելը ընտրել և ինչպես գնահատել դրա որակը տվյալ դեպքում: Դա արվում է, այսպես ասած, *k – fold cross validation* մեթոդով: Այսինքն՝ տվյալները բաժանվում են ենթաբազմությունների: Դրանցից մեկը վերցվում է որպես *test* խումբ, իսկ մնացածները կոչվում են *train* խմբեր: Նկար 13.1-ում տվյալները բաժանվել են $k = 5$ խմբերի: Աղյուսակում նշված է թե որ իտերացիայում որն է թեստային խումբը: Ենթադրենք մենք օգտագործում են հենման վեկտորների մեթոդը և ստանում ենք արդյունքներ, որոնք ներկայացված են 13.2 աղյուսակում: Այստեղ ամեն մի իտերացիայում հաշված է $accuracy_i$ -ի մեծությունը՝

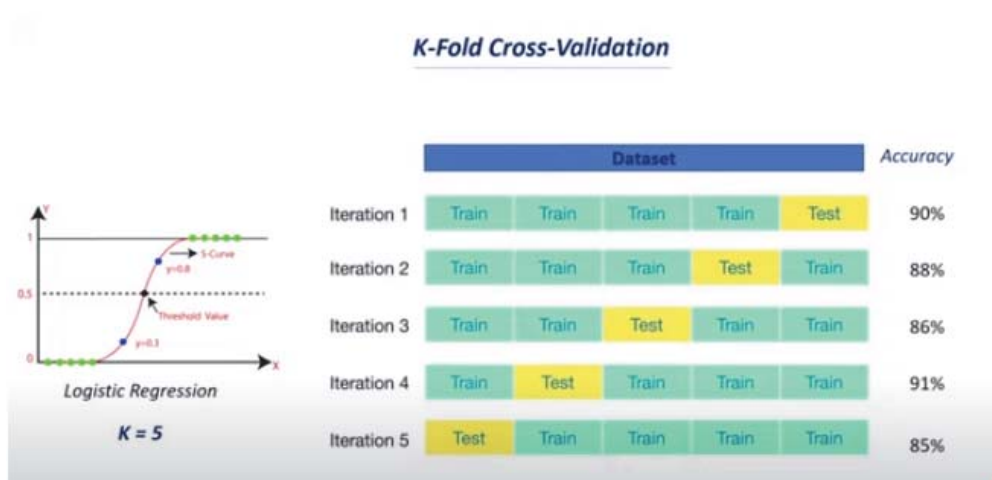
$$accuracy_i = \frac{1}{|Test_i|} \sum_{j \in Test_i} [a(x_j) = y_j], \quad i = 1, 2, \dots, k:$$



Նկ. 13.1: cros validation



Նկ. 13.2: cros validation, SVD



Նկ. 13.3: cros validation, logistic

Տվյալները ներկայացված են նկար 13.2-ում: Այնուհետև վերջում հաշվում ենք

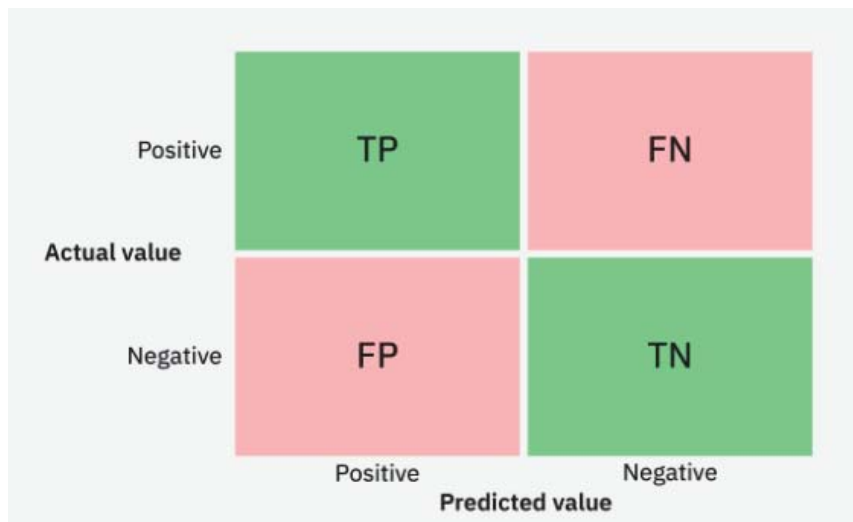
$$\text{Means accuracy} = \frac{1}{k} \sum_{i=1}^k \text{accuracy}_i := \frac{88 + 83 + 86 + 81 + 84}{5} = 84.4:$$

Իսկ *logistic* ռեգրեսիայի դեպքում տվյալները ներկայացված են նկար 13.3-ում:

Քանի որ *means accuracy*-ին *logistic*-ի դեպքում ավելի մեծ է, քան *SVD*-ի պարագայում, ապա տվյալների դիտարկված օրինակի համար *logistic* ռեգրեսիան ավելի լավ է:

13.2 Դասակարգման մոդելների որակի գնահատման չափանիշներ

Այն բանից հետո, երբ ուսուցման տվյալների հիման վրա կառուցվում է մոդելը, հարց է առաջանում որքան «լավ» է մոդելը: Մոդելի որակի գնահատման տարբեր ցուցանիշներ և չափանիշներ կան:



Նկ. 13.4: Սխալների մատրից (confusion matrix)

	Y=1	Y=-1
al(x)=1	80	20
al(x)=-1	20	80

Նկ. 13.5: Սխալների մատրից 1

Դասակարգման դեպքում լայնորեն օգտագործվում են գնահատման հետևյալ գործիքները:

- Սխալների մատրից,
- ճշգրտություն,
- լրիվություն,
- ROC կոր:

Պարզության դեպքում՝ բինար դասակարգման դեպքը: Դիտարկենք նկ. 13.4-ում պատկերված մատրիցը, որը կոչվում է սխալների մատրից (confusion matrix) : Այստեղ

- False Negative (FN)-ն այն օբյեկտների քանակն է, որոնք իրականում պատկանում են «-1» դասին, բայց մոդելը դրանց տեղադրել է «1» դասում:
- False Positive (FP)-ն այն օբյեկտների քանակն է, որոնք իրականում պատկանում են «+1» դասին, բայց մոդելը դրանց տեղադրել է «-1» դասում:
- True Positive (TP)-ը «+1» դասի այն օբյեկտների քանակն է, որոնք գտել է մոդելը:
- True Negative (TN)-ը «-1» դասի այն օբյեկտների քանակն է, որոնք գտել է մոդելը:

Դիտարկենք երկու օրինակ: Առաջին մոդելը (նկ.13.5) «լավ» է ճանաչում և՛

	Y=1	Y=-1
a2(x) = 1	48	2
a2(x) = -1	52	98

Նկ. 13.6: Միալների մատրից 2

առաջին դասի, և՛ երկրորդ դասի օբյեկտները: Նա սխալվում է հավասարաչափ երկու դասերը որոշելիս:

Երկրորդ մոդելը (նկ. 13.6) բավականին ճշգրիտ ճանաչում է «+1» դասի օբյեկտները, բայց միաժամանակ բավականին շատ օբյեկտներ առաջին դասից վերագրում է երկրորդին:

Այժմ քննարկենք «ճշտության» ցուցանիշը: Այն տրվում է հետևյալ բանաձևով՝

$$precision(a, X) = \frac{TP}{TP + FP} = \frac{\sum_{i=1}^l [a(x_i) = 1][a(x_i) = y_i]}{\sum_{i=1}^l [a(x_i) = +1]} :$$

Ինչպես երևում է, այս բանաձևի համարիչը «+1» դասի այն օբյեկտների քանակն է, որոնք գտել է մոդելը, իսկ հայտարարը՝ այն օբյեկտների քանակն է, որոնք մոդելը տվել է «+1» նշիչ:

Լրիվության չափանիշը որոշվում է հետևյալ բանաձևով՝

$$recall(a, X) = \frac{TP}{TP + FN} = \frac{\sum_{i=1}^l [a(x_i) = 1][a(x_i) = y_i]}{\sum_{i=1}^l [y_i = +1]} :$$

Հետևաբար այդ ցուցանիշը մոդելի հայտնաբերած և՛ «1» դասին պատկանող օբյեկտների քանակի, և՛ այդ դասի հզորության հարաբերությունն է:

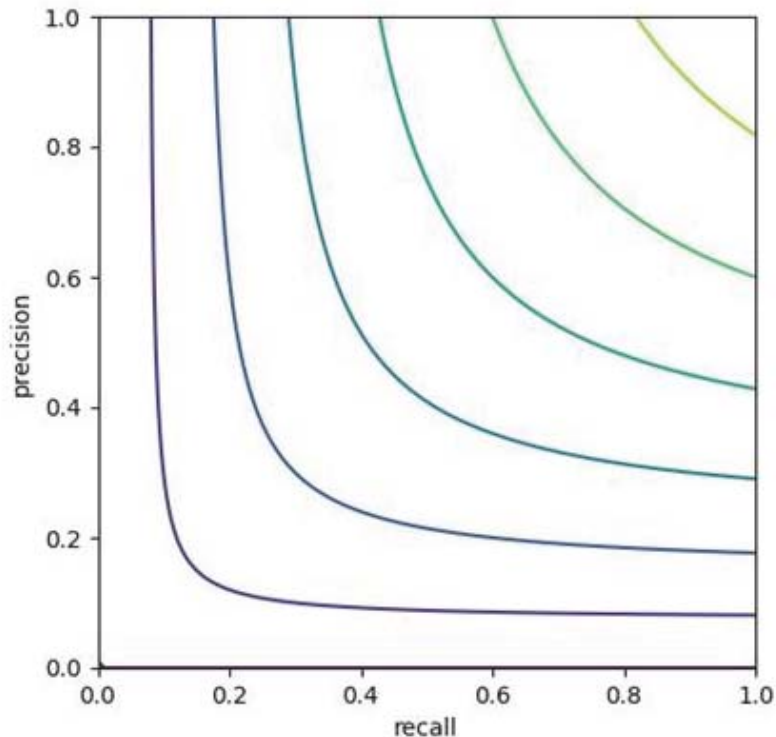
Վերադառնալով վերևում բերված օրինակներին՝ ստանում ենք՝

$$precision(a_1, X) = 0.8, recall(a_1, X) = 0.8,$$

$$precision(a_2, X) = 0.96, recall(a_2, X) = 0.48:$$

Ինչպես երևում է, առաջին ալգորիթմը ունի մեծ ճշտություն և լրիվություն: Այն լավ ալգորիթմ է, և կարելի է օգտագործել: Իսկ երկրորդի ճշտությունը ավելի մեծ է, բայց լրիվությունը՝ բավականին ցածր: Այն կարելի է օգտագործել, եթե անհրաժեշտ է գտնել օբյեկտներ առաջին դասից:

Այժմ քննարկենք, թե որն է կոմպրոմիսը այս երկուսի միջև: Իհարկե, լավ մոդելը պետք է բարձր արժեքներ ցույց տա ինչպես ճշգրտության, այնպես էլ լրիվության պարամետրերում: Այնուամենայնիվ, գործնականում ավելի մեծ նախապատվություն են տալիս մեկ ցուցանիշին և ավելի քիչ՝ մյուսին, քանի որ հաճախ շատ դժվար է միաժամանակ առավելագույնի հասցնել երկուսը:



Նկ. 13.7: F ֆունկցիայի մակարդակի գծեր

Եվ այստեղ առաջանում է հաջորդ խնդիրը՝ միավորել այս երկու բնութագրերը մեկի մեջ, որպեսզի ելքի ժամանակ ունենանք մեկ թիվ, որով կարող ենք դատել ստացված մոդելի որակի մասին:

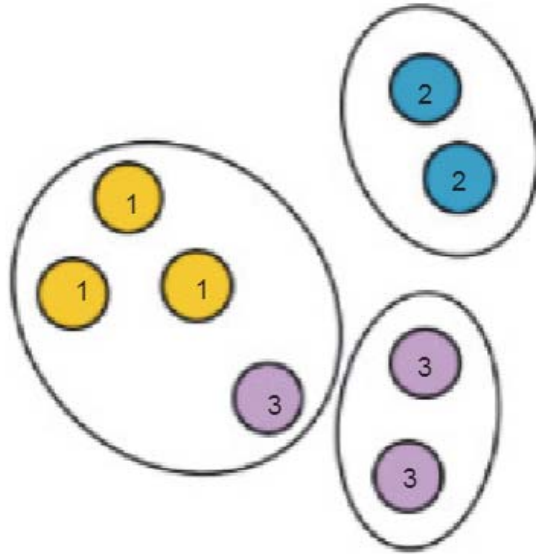
Որպես մոդելի որակի գնահատման չափանիշ վերցնում են հարմոնիկ միջինը՝

$$F = \frac{2 \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}:$$

Նկար 13.7-ում տրված են F ֆունկցիայի մակարդակի գծերը: Գծագրից երևում է, որ F -ի ֆիքսած արժեքի դեպքում recall պարամետրի մեծացնելուց precision պարամետրը աճում է, և ընդհակառակը:

Այժմ քննարկենք ընդհանուր դեպք, երբ կա առնվազն երեք ստանդարտային դասեր: Դիտարկվում են օբյեկտների զույգերը, որոնք համարվում են «+1» դասի օբյեկտ, եթե այդ զույգի երկու օբյեկտներն էլ պատկանում են միևնույն ստանդարտ դասին: Կանխատեսումը համարում ենք «+1» այն և միայն այն դեպքում, երբ զույգի յուրաքանչյուր օբյեկտ պատկանում է միևնույն ալգորիթմական կլաստերին: Այսինքն՝ խնդիրը բերվում է բինար դասակարգման դեպքին:

Դիցուք ունենք ութ օբյեկտներ, որոնք բաժանված են երեք ստանդարտ (Էտալոնային) դասերի՝ 3 օբյեկտ («դեղին-1» դաս), 3 օբյեկտ («մանուշակագույն-3» դաս), 2 օբյեկտ («կապույտ-2» դաս) (նկ. 13.8): Ենթադրենք՝ կլաստերավորման արդյունքում «կապույտ-2» օբյեկտները հավաքվել են նույն դասում:



Նկ. 13.8: Օբյեկտներ և դասեր

Բայց ահա ալգորիթմը սխալվել է «մանուշակագույն-3» մի օբյեկտի վրա և այն դասել է «դեղին-1» խմբում: Այստեղ կա, $8 \times 8 = 64$ զույգ օբյեկտներ, որոնցից

$$3 \times 3 + 3 \times 3 + 2 \times 2 = 22$$

պատկանում են «+» դասին, իսկ մնացած 42 զույգը՝ «-1» դասին: Դրանք ներկայացված են նկար 13.9-ում այսպես ասած «Էտալոնային» մատրիցի տեսքով: Բինար կլաստերավորման տերմիններով .

- $TP = 18$ զույգերը դրական են, և դրանք ալգորիթմական դրական են:
- $FP = 6$ զույգերը դրական են, բայց ալգորիթմական բացասական են:
- $TN = 36$ զույգերը բացասական են, և ալգորիթմական նույնպես բացասական են:
- $FN = 4$ զույգերը բացասական են, բայց դրանք ալգորիթմական դրական են:

Այս տվյալները ներկայացված են նկար 13.10-ում մատրիցի տեսքով: Այստեղ

- TP -ն կանաչ վանդակներն են ,
- FN -ը «կարմիր» վանդակներն են,
- FP -ն շտրիխով վանդակներն են,
- TN -ը մնացած վանդակներն են:

Հաշվելով համապատասխան որակի ցուցանիշները՝ կստանանք՝

$$precision = \frac{TP}{TP + FP} = 0.75, recall = \frac{TP}{TP + FN} = 0.82, F = 0.78:$$

1	1	1					
1	1	1					
1	1	1					
			3	3	3		
			3	3	3		
			3	3	3		
						2	2
						2	2

Նկ. 13.9: Էտալոնային մատրից

TP	TP	TP	FP				
TP	TP	TP	FP				
TP	TP	TP	FP				
FP	FP	FP	TP	FN	FN		
			FN	TP	TP		
			FN	TP	TP		
						TP	TP
						TP	TP

Նկ. 13.10: Միավների մատրիցի օրինակ

13.3 Վարկանիշային (ռանժիրովկայի) մեթոդներ և մոդելների գնահատման որոշ չափանիշ- ներ

Դիցուք տրված է օբյեկտների ուսուցանող հավաքածու՝ $X = \{x_1, x_2, \dots, x_l\}$: Այստեղ օբյեկտները ունեն երկու բնութագրիչներ՝ $x = (q, d)$: Առաջին բաղադրիչը հարցման բաղադրիչն է, իսկ երկրորդը՝ փաստաթղթի: Հարցումների q փոփոխականը փոխվում է Q բազմության վրա, իսկ d փոփոխականը ընդունում է արժեքներ D բազմությունից: $D_q \subset D$ -ով նշանակենք այն փաստաթղթերը, որոնք ստացվել են q հարցման արդյունքում: Այսինքն՝

$$(q, d) \in X \Leftrightarrow q \in Q, d \in D_q:$$

Պատասխանների $Y = \{y\}$ բազմությունը կարգավորված է, և այն կոչվում է ռեյտինգների բազմություն: Փաստաթղթերի միջև կարգի հարաբերությունը սահմանվում է հետևյալ կերպ՝

$$(q, d) < (q, d') \Leftrightarrow y(q, d) < y(q, d'):$$

Այս դեպքում ասում են, որ d' փաստատուողը ավելի «նախընտրելի» (relevant) է, քան d -ն (միևնույն q հարցման դեպքում):

Գոյություն ունեն վարկանիշման մի քանի մեթոդներ: Առաջին մոտեցման դեպքում վերցվում է w պարամետրից կախված մի մոդել և որոնվում է պարամետրի այնպիսի արժեք, որ հետևյալ միջին քառակուսային շեղումը լինի մինիմալ.

$$\sum_{i=1}^l ((a(w, q_i, d_i)) - y(q_i, d_i))^2 \rightarrow \min_w:$$

Որպես մոդել սովորաբար վերցնում են գծային ֆունկցիա՝ $a(w, x) = (w, x)$:

Երկրորդ մոտեցման դեպքում w պարամետրի ընտրությունը կատարվում է հետևյալ ֆունկցիայի մինիմիզացիայի արդյունքում՝

$$\sum_{x_i < x_j} [a(w, x_j) - a(w, x_i)] < 0] \rightarrow \min_w:$$

Ինչպես նկատելի է, այս ֆունկցիոնալը խզվող է, և այստեղ չեն աշխատում գրադիենտային մինիմիզացիոն մեթոդները: Այդ պատճառով կատարվում է վերևից ողորկ մոտարկում, և մինիմիզացվում է դիֆերենցելի այդ ֆունկցիան՝

$$\sum_{x_i < x_j} L(a(w, x_j) - a(w, x_i)) \rightarrow \min_w:$$

Որպես ողորկ մոտարկող L ֆունկցիա սովորաբար վերցնում են $L(M) = \ln(1 + e^{-M})$ լոգիստիկ ռեգրեսիայի ֆունկցիան: Այս մեթոդը կոչվում է RankNet:

Օպտիմիզացիոն այս խնդիրը լուծում են ստոխաստիկ գրադիենտային մեթոդով, որը այս դեպքում կոմենսա հետևյալ տեսքը՝

$$w^{k+1} = w^k + \eta \frac{x_j - x_i}{1 + \exp((w^k, x_j - x_i))} :$$

Այժմ քննարկենք, թե ինչպես կարելի է չափել վարկանշման որակը: Դիտարկենք դա օրինակի միջոցով: Դիցուք ունենք փաստաթղթերի $S = \{D_1, D_2, D_3, D_4, D_5, D_6\}$ բազմությունը: Ընդ որում այդ փաստաթղթերի վարկանիշը գնահատված է 0, 1, 2, 3 սանդղակում: Ենթադրենք այդ փաստաթղթերի վարկանիշային աղյուսակը հետևյալն է $R = \{3, 2, 3, 0, 1, 2\}$: Ենթադրենք՝ հարցման արդյունքում տրվում է փաստաթղթերի ցուցակը վերը նշված հերթականությամբ: Վերցնենք հետևյալ բնութագրիչը՝

$$DCG(6) = \sum_{i=1}^6 \frac{rel_i}{\log_2(i+1)} =$$

$$= 3 + 1.262 + 1.5 + 0.387 + 0.712 = 6.861 \text{ (նկ. 13.11):}$$

Այժմ ենթադրենք հարցման դեպքում փաստաթղթերը տրվում են հետևյալ

i	rel_i	$\log(i+1)$	$\frac{rel_i}{\log(i+1)}$
1	3	1	3
2	2	1.585	1.262
3	3	2	1.5
4	0	2.322	0
5	1	2.585	0.387
6	2	2.807	0.712

Նկ. 13.11: Ցուցակի վարկանիշային գնահատում

ցուցակով՝ $\{D_1, D_3, D_2, D_6, d_5, D_4\}$: Այդ դեպքում վարկանիշի գնահատման վերը նշված հայտանիշը կընդունի հետևյալ արժեքը՝

$$DCG(6) = 3 + 1.893 + +1 + 0.861 + 0.387 + 0 = 7.141 \text{ (նկ. 13.12):}$$

Նկատենք, որ երկրորդ ցուցակի դեպքում DCG հայտանիշը ընդունում է մաքսիմալ արժեք:

Սովորաբար որպես վարկանիշի գնահատման ընդհանուր հայտանիշ (կրիտերիա) վերցնում են հետևյալ արտահայտությունը՝

$$DCG@k(q) = \sum_{i=1}^k \frac{2^{y(q, d_q^i)} - 1}{\log_2(i+1)} :$$

i	rel_i	$\log(i+1)$	$\frac{rel_i}{\log(i+1)}$
1	3	1	3
2	3	1.585	1.893
3	2	2	1
4	2	2.322	0.861
5	1	2.585	0.387
6	0	2.807	0

Նկ. 13.12: Ցուցակի վարկանիշային գնահատում

Այս բանաձևից երևում է, որ որքան մեծ է փաստաթղթի «relevant»-ը, այնքան մեծ է կոտորակի համարիչը: Իսկ հայտարարը բնութագրում է փաստաթղթի դիրքը ցուցակում: Խնդիրը այն է, որ պետք է ընտրել k երկարությամբ վարկանիշային այն ցուցակը, որը մաքսիմում է հաղորդում $DCG@k(q)$ արտահայտությանը:

Շարունակենք ծանոթանալ բինար դասակարգման համար մոդելային գնահատականների որակի չափանիշներին: Այժմ կանդիդատնանք վարկանիշավորման մոդելների որակի գնահատման շատ կարևոր խնդիրն: Բայց եկեք նախ սահմանենք, թե ինչ դասի մոդելներ են դրանք: Դիտարկենք բջջային օպերատորի կամ այլ կազմակերպության հաճախորդների պահպանման դասական օրինակ: Ենթադրենք՝ մեզ հանձնարարվել է մշակել մի մոդել, որը որոշ ինֆորմացիայի հիման վրա կկանխատեսի, թե որ հաճախորդը շուտով կլքի հաստատությունը, և որը՝ ոչ:

$$a(x) = \begin{cases} +1, & \text{կլքի հաստատությունը} \\ -1, & \text{չի լքի հաստատությունը:} \end{cases}$$

Իհարկե, մոդելը միայն ենթադրություն է անում. մենք չգիտենք՝ պատվիրատուն իրականում կհեռանա՞, թե՞ ոչ: Մենք կարող ենք պարզել դա որոշ ժամանակ անց: Բայց ամեն դեպքում լավ կլինի նման «անորոշ» հաճախորդներին զանգահարել և «հաճելի» առաջարկ անել՝ նրանց պահելու համար: Այնուամենայնիվ, ցանկացած կազմակերպության ռեսուրսները սահմանափակ են, և բոլոր նման հաճախորդներին զանգահարելը կարող է ծախսատար լինել: Ուստի այստեղ ավելի լավ է առանձնացնել ամենաանվտանգներին, ովքեր, ամենայն հավանականությամբ, պատրաստվում են անցնել այլ կազմակերպություն: Եվ հարց է առաջանում՝ ինչպե՞ս բացահայտել այդ «անորոշներին»: Դա անում են հետևյալ կերպ: Տեսակավորում են բոլոր հաճախորդներին ըստ սկայար արտադրյալի արժեքների՝ (w, x_i) , $i = 1, 2, \dots$: Ստացվում է մի աղյուսակ (տե՛ս 13.1. վարկանիշային աղյուսակը), որտեղ ամենաանվտանգներն են հենց առաջին (թոփ) հաճախորդները՝ նրանք, ովքեր ամենայն հավանականությամբ մոտ ապագայում կլքեն հաստատությունը, և աղյուսակի ներքևում հայտնված

հաճախորդները ավելի հավատարիմ են: Ավելին, մենք կարող ենք օգտագործել պատմական տվյալները (կամ որոշ ժամանակ անց)՝ որոշելու, թե իրականում ինչպես է իրեն պահում այս կամ այն հաճախորդը:

Վարկանիշային աղյուսակ	
(w, x)	իրական եկամուտ
100	1
98	1
76	1
75	-1
...	...
50	1
25	1
10	-1
6	-1
4	+1
1	-1
-8	+1
-10	-1
-14	-1

Վարկանիշային աղյուսակ 13.1

Այսպիսով, «գանգերի» կենտրոնի սահմանափակ ռեսուրսների դեպքում տրամաբանական կլինի այս աղյուսակում առանձնացնել առաջին մի քանի հաճախորդներին և համոզել նրանց, որ չիրաժարվեն և հետևաբար՝ պահպանել նրանց: Մաթեմատիկորեն սա կարելի է գրել հետևյալ կերպ.

$$a(x, t) = \text{sign}((w, x) - t):$$

Այստեղ t -ն որոշակի շեմ է (թիվ), որով հաճախորդները բաժանվում են անհավատարիմ (+1) և հավատարիմ (-1) դասերի: Սահմանելով, օրինակ, այս շեմը զրոյից մեծ՝ մենք կնշենք ամենաքիչ հավատարիմ հաճախորդներին մոդելի տեսանկյունից: Ավելին, շեմային արժեքը փոփոխական է և կարող է փոփոխվել՝ բացահայտելով քիչ թե շատ հավատարիմ հաճախորդներին, օրինակ՝ կախված գանգերի կենտրոնի ծանրաբեռնվածությունից: Սա վարկանիշային ալգորիթմների հիմնական առանձնահատկությունն է: Մենք պետք է մշակենք մոդել առանց որոշակի շեմային արժեք իմանալու: Այսինքն՝ այն պետք է ընդհանուր առմամբ (ցանկացած շեմերի համար) լավ կանխատեսումներ տա օգտագործողի վարքագծի վերաբերյալ: Օրինակ, եթե որոշակի t շեմի դեպքում սկայյար արտադրյալի արժեքները, որոնք դասավորված են նվազման կարգով, համապատասխանում են պատմական տվյալների թիրախային արժեքների փոփոխությանը, ապա մոդելը կատարում է վատ կանխատեսում, և հաստատության կատարողականը, հավանական է, կտրուկ անկում ապրի:

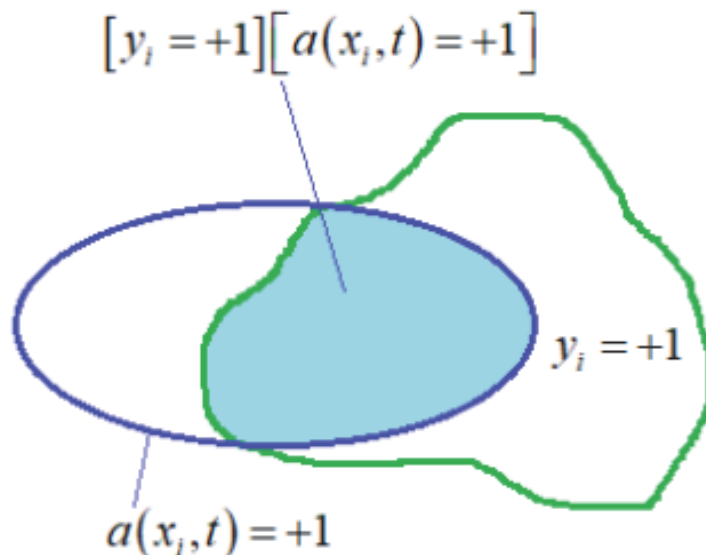
Հետևաբար նախքան մոդելը գործնականում օգտագործելը, այն պետք է գնահատվի, և սովորաբար դա արվում է՝ օգտագործելով վարկանիշային որակի մշակված չափանիշները: Այդպիսի չափանիշ է *ROC*-կորը: Ընդհանուր առմամբ դասակարգման իդեալական մոդել է, երբ տեսակավորումից հետո թիրախային y_i արժեքները նույնպես դասակարգվում են ոչ աճման կարգով (նախ բոլորը՝ +1, ապա բոլորը՝ -1): Այնուհետև ցանկացած t շեմի համար մենք կստանանք բարձրորակ արդյունքներ: Բայց սա իդեալական է: Իրականում կան սխալներ, և այս պայմաններում անհրաժեշտ է հասկանալ, թե ինչպես է մոդելը կանխատեսումներ անում տարբեր շեմերի դեպքում: Դա արվում է մոդելի երկու բնութագրիչի օգնությամբ:

- True Positive Rate - ճիշտ դրական տոկոս կամ համամասնություն (ալգորիթմի կայունությունը).

$$FPR(t) = \frac{FP(t)}{FP(t) + TN(t)} = \frac{\sum_{i=1}^l [y_i = 1][a(x_i, t) = +1]}{\sum_{i=1}^l [y_i = 1]} \quad (\text{նկ.13.13}):$$

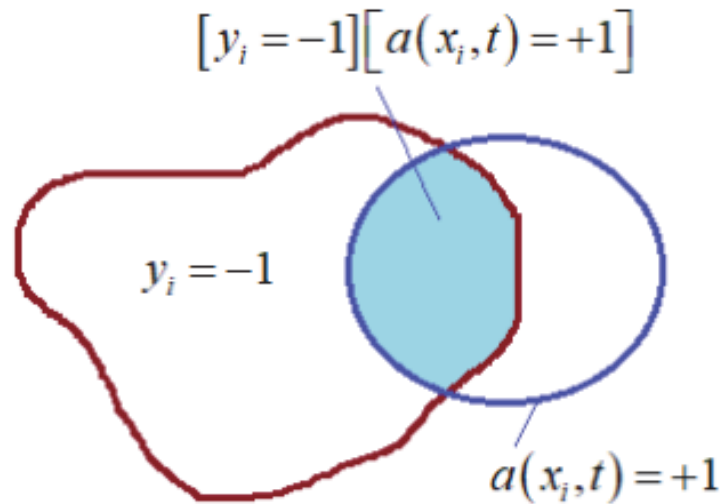
- FPR- կեղծ դրական տոկոս կամ համամասնություն.

$$FPR(t) = \frac{TP(t)}{TP(t) + FN(t)} = \frac{\sum_{i=1}^l [y_i = -1][a(x_i, t) = +1]}{\sum_{i=1}^l [y_i = -1]} \quad (\text{նկ. 13.14}):$$



Նկ. 13.13: TPR

Այսպիսով մենք ունենք երկու բնութագրիչ՝ TPR , FPR , որոնք կախված են t պարամետրից: Ինչպե՞ս կարող ենք այժմ դրանք օգտագործել վարկանիշավորման խնդրի մոդելը գնահատելու համար. կառուցվում է գրաֆիկ, որը կոչվում է



Նկ. 13.14: FPR

Receiver operating characteristic կամ կարճ ROC կոր: ROC կորի գրաֆիկը գծագրվում է TPR/FPR առանցքներով (նկ. 13.15):

Նախ ընտրում ենք t_1 շեմը ավելի մեծ, քան նմուշի սկայար արտադրյալներից ամենամեծը՝

$$t_1 > b(x_1) \geq t_2 \geq b(x_2) \geq t_3 \geq b(x_3) \geq \dots t_l \geq b(x_l) \geq t_{l+1}:$$

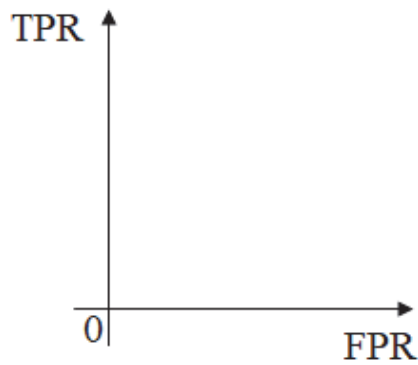
Այդ դեպքում՝

$$TPR(t_1) = \frac{TP(t_1)}{TP(t_1) + FN(t_1)} = \frac{0}{\sum_{i=1}^l [y_i = +1]} = 0,$$

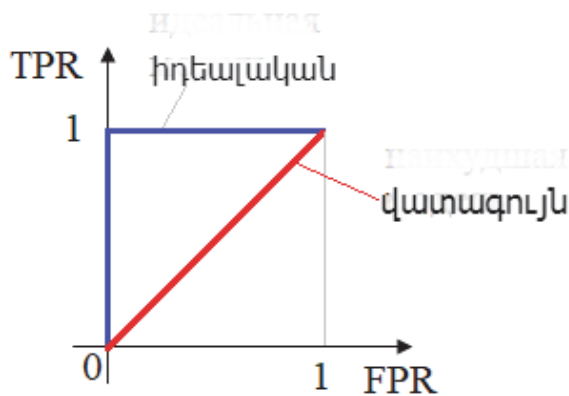
$$FPR(t_1) = \frac{FP(t_1)}{FP(t_1) + TN(t_1)} = \frac{0}{\sum_{i=1}^l [y_i = -1]} = 0:$$

Հետևաբար ROC կորի գրաֆիկը սկսվում է գրոյից (0; 0): Ենթադրենք՝ ունենք իդեալական մոդել, այսինքն՝ այն չի սխալվում (ամբողջովին համապատասխանում է թիրախային արժեքներին): Այնուհետև հաջորդ t_2 շեմի համար մենք կստանանք $TPR(t_2) = 1/l$, իսկ $FPR(t_2) = 0$: Անընդհատ բարձրացնելով շեմերը՝ ակնհայտորեն բարձրանում է TPR-ը՝ պահպանելով գրոյական FPR արժեքը: Դրական դասի բոլոր օբյեկտները թվարկելուց հետո TPR-ը կհասնի 1-ի առավելագույն արժեքի, իսկ FPR-ը դեռ զրո կլինի: Հետագայում, անցնելով բացասական դասերի, ամեն ինչ հակառակը կլինի. TPR-ը մնում է անփոփոխ՝ հավասար 1-ի, իսկ FPR-ն աստիճանաբար կաճի՝ հասնելով 1-ի: Արդյունքում, իդեալական մոդելի համար ROC կորի գրաֆիկը կունենա հետևյալ տեսքը՝ (տե՛ս նկ.13.16):

Ամենավատ մոդելի դեպքում, երբ այն կանխատեսում է բինար դասերը $1/2$ հավանականությամբ, ROC կորը մոտավորապես դասավորված է անկյունագծի վրա: Ինչպես արդեն հասկանալի է, իրական մոդելները փոքր թվով կանխատեսման սխալներով գտնվում են «կարմիր» (վատագույն) և «կապույտ» (իդեալական) կորերի միջև (տես նկ. 13.17):



Նկ. 13.15: TPR



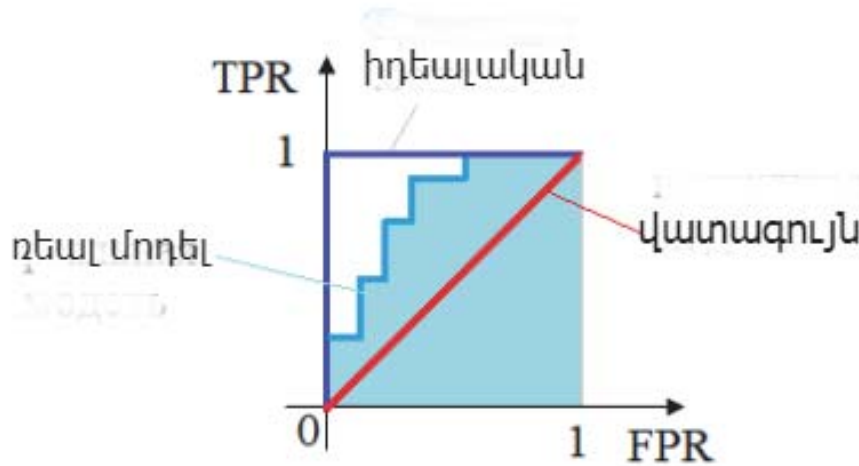
Նկ. 13.16: TPR

Այսպիսով մենք սովորեցինք, թե ինչպես կառուցել ROC կորը, բայց ինչպե՞ս կարող ենք այժմ օգտագործել այն մոդելի որակի թվային բնութագիրը ստանալու համար: Այստեղ երկու ընդհանուր մոտեցում կա. առաջին դեպքում պարզապես վերցվում է ROC կորի տակ գտնվող մակերեսը: Այս հատկանիշը կոչվում է AUC-ROC, և որքան մեծ է այն, այնքան լավ մոդել է: Իդեալական մոդելի համար մակերեսը 1 է, իսկ ավելի վատ մոդելի համար՝ 1/2: Լավն են համարվում 0,9-ից ավելի ցածր տարածքներով մոդելները: Այսինքն՝ AUC-ROC ցուցիչը ցույց է տալիս, թե որքան լավ է մոդելը դասակարգում օբյեկտները:

Երկրորդ դեպքում իրական և վատագույն մոդելների ROC կորերի միջև մակերեսների տարբերության չափումը հաշվարկվում է հետևյալ բանաձևով՝

$$G = 2((AUC - ROC) - \frac{1}{2}):$$

Այս թիվը կոչվում է Ջինի ինդեքս:



Նկ. 13.17: TPR

13.4 Կլաստերավորման մոդելների գնահատման մետրիկական չափանիշներ

Դիտարկենք կլաստերավորման որակի գնահատման մի քանի մետրիկական չափանիշներ:

Առաջին չափանիշը միջին ներդասային հեռավորությունն է՝

$$F_0 = \frac{\sum_{i < j} \rho(x_i, x_j) [a(x_i) = a(x_j)]}{\sum_{i < j} [a(x_i) = a(x_j)]} :$$

Այսինքն՝ միևնույն կլաստերի կետերի միջև եղած հեռավորությունների գումարը բաժանվում է նույն կլաստերին պատկանող գույգերի թվին:

Համանմանությամբ ներմուծվում է երկրորդ մետրիկական չափանիշը՝ միջին միջդասային հեռավորություն:

$$F_1 = \frac{\sum_{i < j} \rho(x_i, x_j) [a(x_i) \neq a(x_j)]}{\sum_{i < j} [a(x_i) \neq a(x_j)]} :$$

Այսպիսով ունենք երկու հայտանիշ: Դրվում են դիսկրետ օպտիմիզացիայի հետևյալ երկու խնդիրները՝ օպտիմալ է համարվում այն a մոդելը, որը լուծում է հետևյալ խնդիրներից մեկը. $F_0 \rightarrow \min$, կամ $F_1 \rightarrow \max$:

14

Ուժեղացում (Boosting algorithms)

Դասակարգման, ռեգրեսիայի և կանխատեսման բարդ խնդիրներ լուծելիս հաճախ պարզվում է, որ ալգորիթմներից և ոչ մեկը չի ապահովում ցանկալի որակ (կախվածության վերականգնում): Նման դեպքերում իմաստ ունի կառուցել ալգորիթմների կոմպոզիցիաներ, որոնցում առանձին ալգորիթմների սխալները փոխադարձաբար չեզոքացվում են: Կոմպոզիցիաների ամենահայտնի օրինակներն են պարզ և կշռադատված քվեարկությունը:

Այս բաժնում ներմուծվում է «թույլ» պրեդիկտորի գաղափար: Այնուհետև նկարագրվում են երեք ալգորիթմներ՝ *Adaboost*, *Anyboost* և *Gradient boosting*, որոնք իրենց մուտքում ստանալով այդ թույլ պրեդիկտորները՝ ելքում տալիս են կոռեկտ պրեդիկտորներ: Այդ ալգորիթմները աշխատում են վերը նշված սկզբունքով:

Որպես թույլ պրեդիկտորներ սովորաբար հանդես են գալիս որոշիչ ծառերը: Հիշենք, որ կոռեկտ ալգորիթմի գաղափարը տրված է 2.3 պարագրաֆում:

14.1 Թույլ հիպոթեզների դասեր

Դիցուք H -ը հիպոթեզների բազմությունն է, D -ն նմուշների Z բազմության վրա տրված հավանականային բաշխում է, իսկ $f : X \rightarrow \{-1, +1\}$ -ը՝ նշիչային այնպիսի ֆունկցիա, որ $L_D(f) = 0$: A ալգորիթմը կոչվում է γ -«թույլ» ($\gamma > 0$), եթե ցանկացած $\delta \in (0, 1)$ թվի համար գոյությունն ունի $m_H(\delta) \in \mathbb{N}$ համար, այնպիսին, որ կիրառելով այդ ալգորիթմը $m \geq m_H(\delta)$ երկարությամբ հավաքածուի վրա՝ այն կվերադարձնի այնպիսի $h = A(S) \in H$ հիպոթեզ, որ $L_D(A(S)) \leq 1/2 - \gamma$ անհավասարությունը տեղի կունենա առնվազն $1 - \delta$ հավանականությամբ:

Հիպոթեզների H դասը կոչվում է « γ -թույլ», եթե դրա համար գոյություն ունի γ -թույլ ալգորիթմ:

Այժմ բերենք թույլ հիպոթեզների դասի օրինակ:

Դիտարկենք հիպոթեզների հետևյալ բազմությունը՝

$$H = \{h_{\theta_1, \theta_2, b} : \theta_1 \leq \theta_2, b \in \{-1, +1\}\},$$

որտեղ

$$h_{\theta_1, \theta_2, b}(x) = \begin{cases} +b, & x < \theta_1 \text{ or } x > \theta_2, \\ -b, & \theta_1 \leq x \leq \theta_2: \end{cases} :$$

Այժմ ցույց տանք, որ եթե որպես A վերցնենք ERM ալգորիթմը, ապա այն կվերադարձնի

$$\hat{h} \in B = \{x \rightarrow \text{sign}(x - \theta).b : \theta \in R, b \in \{-1, +1\} \subset H$$

հիպոթեզը, այնպիսին, որ $L_D(\hat{h}) \leq 1/3$: Այսինքն՝ այս դեպքում $\gamma = 1/12$: Իրոք, H դասի ամեն մի դասակարգիչ բաղկացած է երեք մասից ($(-\infty, \theta_1)$, $[\theta_1, \theta_2]$, $(\theta_2, +\infty)$): Ընդ որում կլասիֆիկատորը յուրաքանչյուր մասին վերագրում է նշիչ՝ ըստ հերթականության: Օրինակ՝ $+1, -1, +1$ (տե՛ս նկ.14.1):



Նկ. 14.1: «Թույլ» հիպոթեզի օրինակ

Եթե D -ն որևէ հավանականային բաշխում է ամբողջ առանցքի վրա, ապա կգտնվի մի ինտերվալ, որի վրա D բաշխման կշիռը փոքր կլինի $1/3$ -ից: Դիցուք այժմ $h \in B$ -ն այն հիպոթեզն է, որի նշիչները համընկնում են $\hat{h}$ -ի նշիչների հետ երկու ինտերվալներում և տարբերվում են այն ինտերվալում, որում D -ի կշիռը փոքր է կամ հավասար $1/3$ -ից: Քանի որ B դասի VC չափողականությունը հավասար է 2 -ի, ապա կարող ենք կիրառել մեքենայական ուսուցման ֆունդամենտալ թեորեմը: Համաձայն այդ թեորեմի գոյություն ունի այնպիսի m_H թիվ, որ կամայական $m \geq m_H$ երկարությամբ հավաքածուի վրա կիրառելով ERM ալգորիթմը, այն վերադարձնում է $\hat{h} \in H$ հիպոթեզ այնպիսին, որ

$$L_D(\hat{h}) \leq \min_{h \in B} L_D(h) + \epsilon \leq \frac{1}{3} + \epsilon:$$

Դիցուք այժմ $\epsilon = 1/12$: Այդ դեպքում քանի որ

$$1/3 + 1/12 = 1/2 - 1/12,$$

ապա $\gamma = 1/12$:

14.2 Բուստինգ դասակարգման խնդիրներում, Adaboost ալգորիթմը

Դիտարկենք բինար դասակարգման խնդիրը՝ $Y = \{-1, +1\}$: Ենթադրենք ունենք հիպոթեզների մի բազմություն H , որի ամեն մի $h \in H$ հիպոթեզ յուրաքանչյուր x օբյեկտի վրա վերադարձնում է 1 կամ -1 արժեքներ: Դիտարկենք հետևյալ հիպոթեզը՝

$$a(x) = \text{sign}\left(\sum_{t=1}^T \alpha_t h_t(x)\right), \quad h_t \in H, \quad x \in X,$$

որտեղ α_t , $t = 1, 2, \dots, T$ դրական թվեր են: Այնուհետև դիտարկենք բինար դասակարգման հենքային կորստի ֆունկցիայի մինիմիզացիայի հետևյալ խնդիրը՝

$$Q_T = \sum_{i=1}^l [y_i \sum_{t=1}^T \alpha_t h_t(x_i) < 0] \rightarrow \min_{\alpha_t, h_t}:$$

Սա օպտիմիզացիայի բարդ խնդիր է, քանի որ ֆունկցիոնալը խզվող է: Որպեսզի կատարենք մինիմիզացիա ըստ α_t պարամետրի ֆունկցիոնալի մեջ հենքային ֆունկցիան վերևից մոտարկենք ողորկ ֆունկցիայով: Դիտարկենք էքսպոնենցիալ մոտարկումը՝ $E(z) = e^{-z}$: Գնահատենք Q_T ֆունկցիոնալը վերևից՝

$$\begin{aligned} Q_T &\leq \tilde{Q}_T = \sum_{i=1}^l \exp\left(-y_i \sum_{t=1}^T \alpha_t h_t(x_i)\right) = \\ &= \sum_{i=1}^l \exp\left(-y_i \sum_{t=1}^{T-1} \alpha_t h_t(x_i)\right) \exp(-y_i \alpha_T h_T(x_i)): \end{aligned}$$

Նշանակենք

$$w_i \equiv \exp\left(-y_i \sum_{t=1}^{T-1} \alpha_t h_t(x_i)\right):$$

Այս բանաձևից հետևում է w_i կշռային գործակիցը կախված չէ α_T, h_T պարամետրերից: Հետևաբար այդ գործակիցը կարելի է հաշվել նախքան h_T պրեդիկտորի կառուցումը: Կատարենք այդ գործակիցների նորմավորումը՝

$$\tilde{W}^l = (\tilde{w}_1, \dots, \tilde{w}_l),$$

որտեղ

$$\tilde{w}_i = \frac{w_i}{\sum_{i=1}^l w_i}:$$

Տեղի ունի հետևյալ պնդումը:

Թեորեմ 14.1. Դիցուք

$$\min_{h \in H} Q(h, \tilde{W}^l) < \frac{1}{2} - \gamma, \quad \forall \tilde{W}^l, \quad \|\tilde{W}^l\| = 1:$$

Այդ դեպքում $\tilde{Q}_T$ ֆունկցիոնալի մինիմումը հասանելի է հետևյալ պարամետրերի դեպքում՝

- $h_T = \arg \min_h \tilde{Q}_l(h, \tilde{W}^l);$
- $\alpha_T = \frac{1}{2} \ln \frac{1-Q(T, \tilde{W}^l)}{Q(T, \tilde{W}^l)}:$

Ապացույց: Կամայական $\alpha \in R$ և $\beta \in \{0, 1\}$ թվերի համար տեղի ունի հետևյալ ակնհայտ հավասարությունը՝

$$e^{-\alpha\beta} = e^{-\alpha}[\beta = 1] + e^{-\alpha}[\beta = -1]:$$

Հետևաբար $\tilde{Q}_T$ ֆունկցիոնալը կարելի է ներկայացնել հետևյալ տեսքով՝

$$\begin{aligned} \tilde{Q}_T &= (e^{-\alpha_T} \sum_{i=1}^l \tilde{w}_i [h_T(x_i) = y_i] + e^{\alpha_T} \sum_{i=1}^l \tilde{w}_i [h_T(x_i) \neq y_i]) \sum_{i=1}^l w_i = \\ &= (e^{-\alpha_T} (1 - Q) + e^{\alpha_T} Q) \tilde{Q}_{T-1}, \end{aligned}$$

որտեղ $Q \equiv Q(h_T, \tilde{W}^l)$: Ածանցելով $\tilde{Q}_T$ ֆունկցիոնալը ըստ α_T պարամետրի և հավասարեցնելով այն զրոյի՝ կստանանք այդ պարամետրի օպտիմալ արժեքը

$$\alpha_T = \frac{1}{2} \ln \frac{1-Q}{Q}:$$

Տեղադրելով այս արժեքը $\tilde{Q}_T$ ֆունկցիոնալի մեջ՝ կստանանք

$$\tilde{Q}_T = \tilde{Q}_{T-1} \sqrt{4Q(1-Q)}:$$

Դիտարկենք հետևյալ ալգորիթմը, որը կոչվում է *adaboost*.

- կատարվում է կշիռների ինիցիալիզացիա՝ $w_i = 1/l$, $i = 1, 2, \dots, l$,
- կամայական $t = 1, 2, \dots, T$ գտնում ենք h_t և α_t պարամետրերը հետևյալ քանաձևերով՝

$$h_t = \arg \min_h Q(h, W^l), \alpha_t = \frac{1}{2} \ln \frac{1 - Q(h_t, W^l)}{Q(h_t, W^l)},$$

- կատարվում է կշիռների վերահաշվարկ՝

$$w_i := w_i \exp(-\alpha_t y_i h_t(x_i)), \quad i = 1, 2, \dots, l,$$

և այնուհետև կշիռների նորմավորում՝

$$w_i := \frac{w_i}{\sum_{i=1}^l w_i}, \quad i = 1, 2, \dots, l:$$

Տեղի ունի *Adaboost* ալգորիթմի գուգամիտության հետևյալ թեորեմը:

Թեորեմ 14.2. Դիցուք գոյություն ունի $\gamma > 0$ հաստատուն այնպիսին, որ $Q(h_t, W^l) < 1/2 - \gamma \quad \forall t$: Այդ դեպքում Adaboost ալգորիթը վերջավոր քայլերի ընթացքում կատուցում է $a(x)$ կոռեկտ ալգորիթը, այսինքն՝ որոշակի քայլերից հետո (ինչ-որ պահի) այդ ալգորիթի սխալանքը հավաքածուի վրա հավասար կլինի զրոյի:

Ապացույց: Ըստ ենթադրության $Q(h_t, W^l) \leq 1/2 - \gamma$: Մյուս կողմից քանի որ $g(a) = a(1 - a)$ ֆունկցիան մոնոտոն աճող է $[0, 1/2]$ հատվածի վրա, ապա

$$Q_{T+1} \leq \tilde{Q}_{T+1} = \tilde{Q}_T \sqrt{4Q(1-Q)} \leq \tilde{Q}_T \sqrt{4(\frac{1}{2} - \gamma)(\frac{1}{2} + \gamma)} \leq \tilde{Q}_1 (1 - 4\gamma^2)^{T/2} :$$

Այս անհավասարությունից հետևում է, որ ինչ-որ մի T պահի անհավասարության աջ մասը դառնում է փոքր մեկից, ինչը նշանակում է, որ Q_T -ն դառնում է զրո, քանի որ այն ընդունում է ոչ բասական ամբողջ արժեքներ:

Այժմ քննարկենք թեորեմի $Q(h_t, W^l) < 1/2 - \gamma$ պայմանը:

$S = \{(x_1, y_1), \dots, (x_l, y_l)\}$ հավաքածուի վրա h կլասիֆիկատորի որակը գնահատելու համար դիտարկենք հետևյալ երկու ֆունկցիոնալները՝

$$N(h, W^l) = \sum_{i=1}^l w_i [h(x_i) = -y_i],$$

$$P(h, W^l) = \sum_{i=1}^l w_i [h(x_i) = y_i]:$$

Առաջինը այն օբյեկտների կշիռների գումարն է, որոնց վրա h կլասիֆիկատորը սխալ կանխատեսում է արել: Երկրորդը այն օբյեկտների կշիռների գումարն է, որոնց վրա h կլասիֆիկատորը ճիշտ կանխատեսում է արել: Նկատենք, որ $N + P = 1$: Հետևաբար՝ թեորեմի պայմանը նշանակում է, որ գոյություն ունի $\gamma > 0$ դրական թիվ, այնպիսին, որ

$$P(h_t, \tilde{W}^l) - N(h_t, \tilde{W}^l) \leq \gamma:$$

Այժմ մի պարզ օրինակի վրա մեկնաբանենք այս ալգորիթի աշխատանքը:

Դիցուք ունենք յոթ օբյեկտներ՝ իրենց նշիչներով և երեք կլասիֆիկատորներ h_1, h_2, h_3 , որոնց կանխատեսման արդյունքները այդ հավաքածուի վրա ներկայացված է 14.1 աղյուսակում:

Աղյուսակ 14.1, Boosting ալգորիթ

օբյեկտ	h_1	h_2	h_3	y
1	1	-1	-1	1
2	-1	1	-1	1
3	-1	-1	1	1
4	-1	1	1	1
5	1	-1	1	1
6	1	1	-1	1
7	1	1	1	1

Առաջին իտերացիայում կատարվում է

- կշիռներին սկզբնական արժեքների վերագրում (ինիցիալիզացիա)՝

$$W = (\frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}):$$

- Յուրաքանչյուր դասակարգչի համար հաշվում ենք նրա կատարած սխալ կանխատեսումների կշռային գումարը: Քանի որ h_1 կլասիֆիկատորը սխալվել է երկրորդ, երրորդ և չորրորդ օբյեկտների վրա, ապա

$$N(h_1, W) = \frac{1}{7} + \frac{1}{7} + \frac{1}{7} = \frac{3}{7} \approx 0.42:$$

Նման կերպ

$$N(h_2, W) = \frac{1}{7} + \frac{1}{7} + \frac{1}{7} = \frac{3}{7} \approx 0.42,$$

$$N(h_3, W) = \frac{1}{7} + \frac{1}{7} + \frac{1}{7} = \frac{3}{7} \approx 0.42:$$

- Ընտրում ենք h_1 կլասիֆիկատորը հետևյալ կշռային գումարով՝

$$\begin{aligned} \alpha_1 &= \frac{1}{2} \ln \frac{1 - N(h_1, W)}{N(h_1, N)} \\ &= \frac{1}{2} \ln \frac{1 - 0.42}{0.42} = 0.06: \end{aligned}$$

- Կատարվում է օբյեկտների կշիռների վերահաշվարկ:

$$w_1 := w_1 e^{-\alpha_1 y_1 h_1(x_1)} = 0.13,$$

$$w_2 := W_2 e^{-\alpha_1 y_2 h_1(x_2)} = 0.15,$$

$$w_3 := w_3 e^{-\alpha_1 y_3 h_1(x_3)} = 0.15,$$

$$w_4 = 0.15, \quad w_5 = 0.13, \quad w_6 = 0.13, \quad w_7 = 0.13:$$

Այնուհետև կատարվում է նորմավորում՝

$$w_1 := \frac{w_1}{\sum_{i=1}^7 w_i} = \frac{0.13}{0.93} = 0.14:$$

Նման կերպ

$$w_2 = 0.16, \quad w_3 = 0.16, \quad w_4 = 0.16, \quad w_5 = 0.14, \quad w_6 = 0.14, \quad w_7 = 0.14:$$

Երկրորդ իտերացիայում

- օբյեկտների կշռային արժեքներն են

$$W = (0.14, 0.16, 0.16, 0.16, 0.14, 0.14, 0.14):$$

- Նորից յուրաքանչյուր կլասիֆիկատորի համար հաշվենք նրա սխալ կանխատեսած օբեկտների նոր կշռային արժեքները.

$$N(h_1, W) = 0.16 + 0.16 + 0.16 = 0.48,$$

$$N(h_2, W) = 0.14 + 0.16 + 0.14 = 0.44,$$

$$N(h_3, W) = 0.14 + 0.16 + 0.14 = 0.44:$$

- Ուրեմն կարող ենք ընտրել երկրորդ կամ երրորդ կլասիֆիկատորը: Վերցնենք երկրորդը:
- Հաշվենք h_2 կլասիֆիկատորի կշռային արժեքը՝

$$\alpha_2 = \frac{1}{2} \ln \frac{1 - N(h_2, w)}{N(h_2, w)} = \frac{1}{2} \ln \frac{1 - 0.44}{0.44} = 0.06:$$

Այսպիսով, երկու իտերացիայից հետո կառուցվել է

$$a(x) \equiv \alpha_1 h_1(x) + \alpha_2 h_2(x)$$

պրեդիկտորը: S հավաքածուի վրա նրա կանխատեսումները հետևյալ թվերն են՝

$$a(1) = 0.06(h_1(1) + h_2(1)) = 0, a(2) = 0, a(3) = -0.12, a(4) = 0, a(5) = 0,$$

$$a(6) = 0.12, a(7) = 0.12:$$

Այդ դեպքում՝

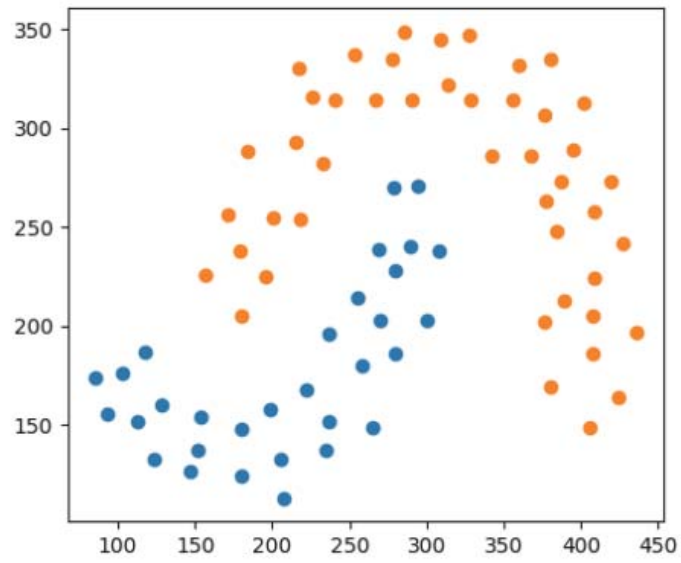
$$Q_T = \sum_{i=1}^l [y_i \sum_{t=1}^T \alpha_t h_t(x_i) < 0] = 0 \quad (T = 2, l = 7):$$

Այժմ կոնկրետ օրինակի վրա դիտարկենք այս ալգորիթմի իրականացումը *Python*-ով:

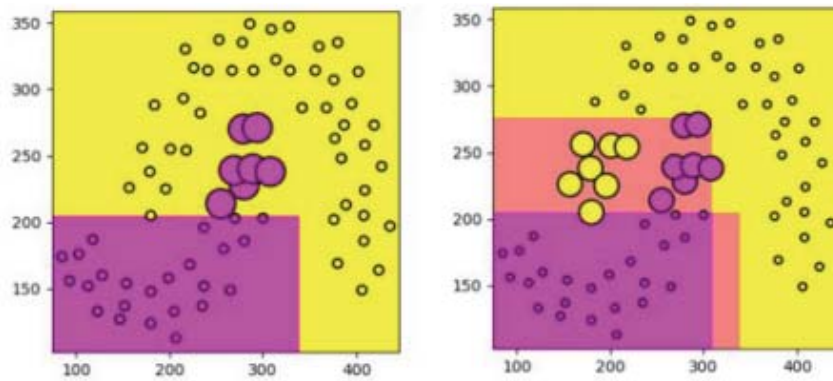
Դիտարկենք երկու դասերից բաղկացած հավաքածու (նկ.14.2):

Դիտարկում ենք երկու խորությամբ որոշիչ ծառեր: (<https://github.com/sefvedu-rus/machinelearning>): Օգտագործելով նշված դասի բազային ալգորիթմները՝ *Boosting*-ը տալիս է արդյունքներ, որոնք ներկայացված են նկարներ 14.3 և 14.4-ում:

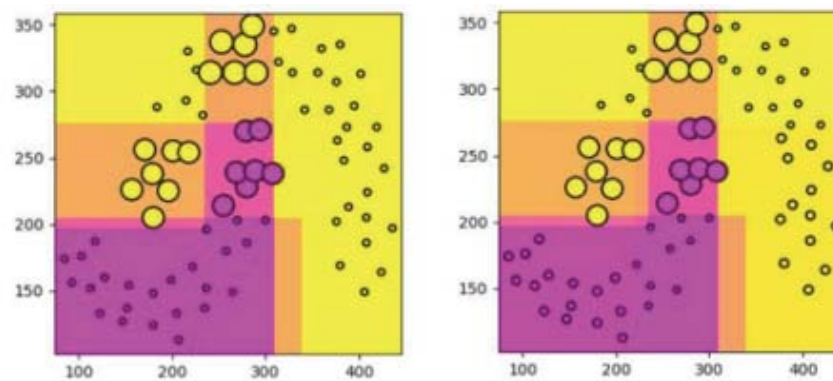
Ինչպես նկատելի է ծառերի քանակի ավելացումով սխալ կլասիֆիկացված օբյեկտների քանակը նվազում է:



Նկ. 14.2: Adaboosting դասակարգում



Նկ. 14.3: Adaboosting-ի դասակարգման արդյունքները (որոշիչ ծառեր 1-2)



Նկ. 14.4: Adaboosting կլասիֆիկացիայի արդյունքները (որոշիչ ծառեր 3-4)

14.3 Boosting ռեգրեսիայի խնդիրներում

Նախորդ բաժնում մենք քննարկեցինք *boosting* ալգորիթմը բինար դասակարգման խնդիրներում:

Այժմ այդ մոտեցումը ներկայացնենք ռեգրեսիայի խնդիրներում: Ուրեմն մենք կառուցելու ենք $h_1(x), \dots, h_T(x)$ ալգորիթմներ և հետո դրանք գումարում ենք իրար համապատասխան $w_1, \dots, w_T$ կշռային գործակիցներով: Ընդ որում այդ պարամետրերը կառուցվելու են ռեկուրսիվ: Ֆիքսվում են նախօրոք գտած էլեմենտները՝ $h_1(x), \dots, h_{T-1}(x), \alpha_1, \dots, \alpha_{T-1}$, և այնուհետև որոնվում են ընթացիկ $h_T(x)$, α_T պարամետրերը: Քանի որ մենք դիտարկում ենք ռեգրեսիայի խնդիրը, ապա կորստի ֆունկցիան կունենա հետևյալ տեսքը՝

$$Q_T = \frac{1}{2} \sum_{i=1}^l (y_i - a(x_i))^2:$$

Այստեղ $a(x)$ ալգորիթմը փնտրվում է հետևյալ տեսքով՝ $a(x) = \sum_{t=1}^T w_t h_t(x)$: Տեղադրելով այն կորստի ֆունկցիայի բանաձևում՝ կստանանք

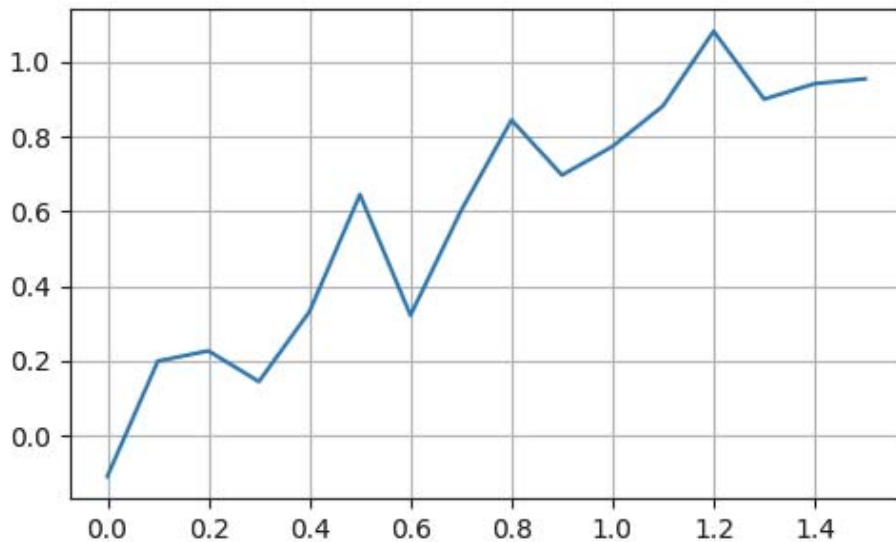
$$Q_T = \frac{1}{2} \sum_{i=1}^l (y_i - \sum_{t=1}^{T-1} \alpha_t h_t(x_i) - \alpha_T h_T(x_i))^2:$$

Նշանակենք՝

$$s_{i,T-1} = y_i - \sum_{t=1}^{T-1} \alpha_t h_t(x_i), \quad i = 1, 2, \dots, l:$$

Այստեղ որպես h_t ալգորիթմներ վերցվում են որոշիչ ծառերը, ինչպես դասակարգման խնդիրներում, թեև կարելի է դրանք ընտրել գծային ալգորիթմների ինչ-որ ընտանիքից: Այստեղ α_t կշռային գործակիցները կարելի է վերցնել հավասար մեկի:

Այսպիսով, եթե ընտրում ենք որոշիչ ծառերը, ապա արդեն հայտնի է, որ այդ ծառերի տերմինների վրա պահպանվում է այդտեղ հայտնված օբյեկտների y_i նշիչների միջին թվաբանականը:



Նկ. 14.5: $y = \sin(x) + \eta$

Adaboost ալգորիթը ռեգրեսիայի դեպքում վերջնականում կունենա հետևյալ տեսքը՝

- կատարվում է $s_{i,0}$, $i = 1, 2, \dots, l$ մնացորդների սկզբնական ինիցիալիզացիա՝

$$s_{i,0} = y_i, \quad i = 1, 2, \dots, l,$$

- յուրաքանչյուր t պահի h_t ալգորիթը որոնվում է հետևյալ պայմանից՝ $h_t = \arg \min_h Q_t$, $t = 1, 2, \dots, T$,
- կատարվում է մնացորդների թարմացում՝

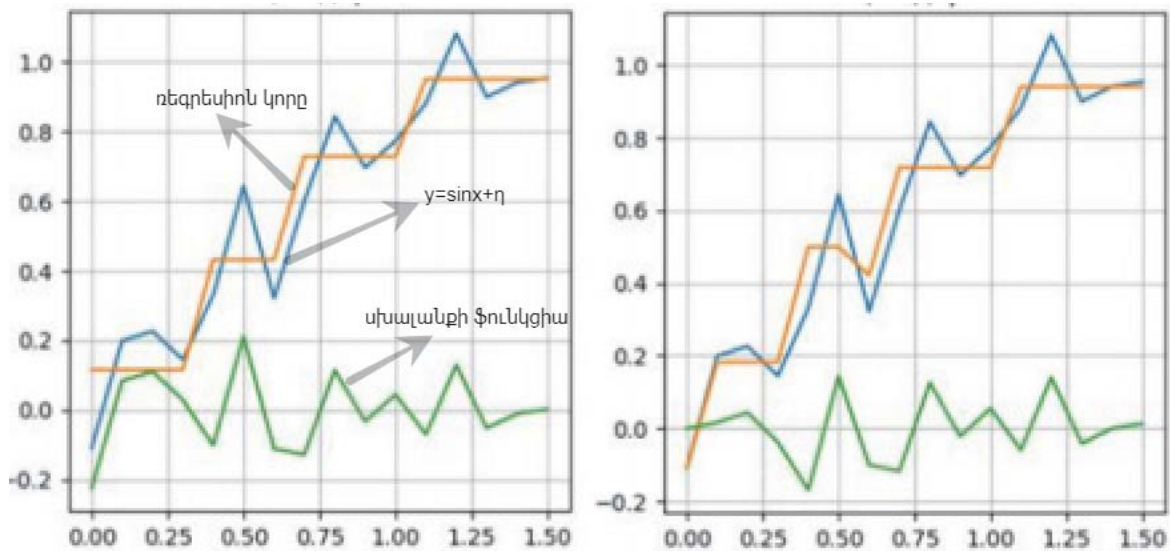
$$s_{i,t} = s_{i,t-1} - h_t(x_i) \quad i = 1, 2, \dots, l:$$

Այժմ օրինակի վրա տանք այս ալգորիթի իրականացումը *Python*-ով: Դիցուք տրված է հավաքածու՝

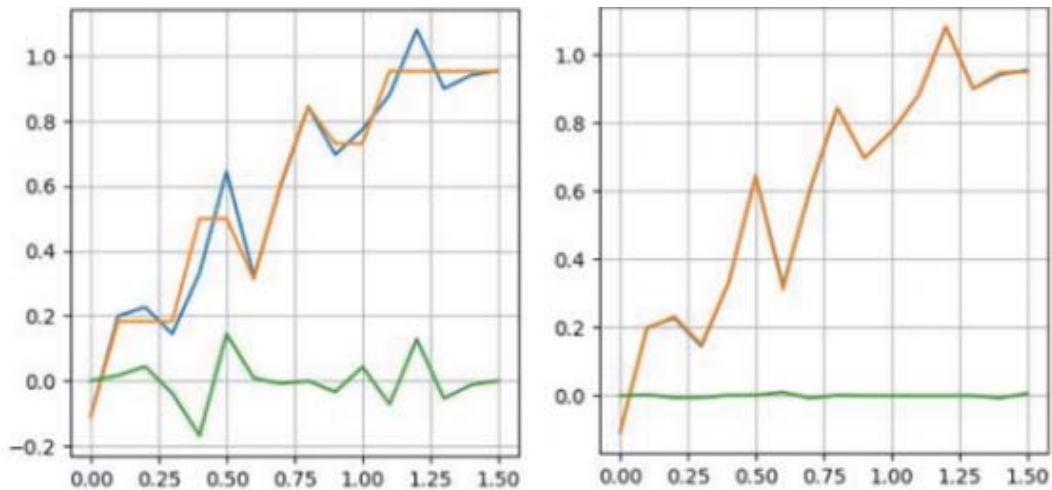
$$\{(x_i, y_i = \sin(x_i) + \eta_i), \quad i = 1, 2, \dots, l\},$$

որտեղ η_i պատահական մեծություններ են, որոնց միջինը հավասար է զրոյի, իսկ դիսպերսիան՝ 0.1 (նկ. 14.5): Այդ նկարում ներկայացված է $y = \sin x + \eta$ ֆունկցիայի գրաֆիկը:

Adaboost ալգորիթով մոտարկենք այս ֆունկցիան: Որպես բազային ալգորիթներ հանդես են գալու որոշիչ ծառեր, որոնց գրադարանը ունի հետևյալ հղումը՝ <https://github.com/selfedu-rus/machinelearning>: Կիրառելով նշված ալգորիթը՝ ստանում ենք սկզբնական ազդանշանի վերականգնման հետևյալ արդյունքները՝ (նկ. 14.6, 14.7): 14.6-ի առաջին նկարում կոմպոզիցիան բաղկացած է մեկ ծառից, և ներկայացվում է ալգորիթով գեներացված գրաֆիկը (ռեգրեսիոն կորը), որը մոտարկում է սկզբնական ֆունկցիան:



Նկ. 14.6: Adaboost ալգորիթը $y = \sin(x) + \eta$ ֆունկցիայի դեպքում



Նկ. 14.7: Adaboost ալգորիթմը $y = \sin(x) + \eta$ ֆունկցիայի դեպքում

Պատկերված է նաև սխալանքի ֆունկցիան, այսինքն՝ ռեգրեսիոն կորի և սկզբնական ֆունկցիայի տարբերության գրաֆիկը: Ինչպես նկատելի է, այս դեպքում մոտարկումը բավականին վատ է: Երկրորդ նկարում կոմպոզիցիան բաղկացած է երեք ծառերից:

Նկար 14.7-ի առաջին նկարում վերցրած են հինգ, իսկ երկրորդում տասը որոշիչ ծառեր: Ինչպես երևում է նկարից, այս դեպքում ալգորիթմի կառուցած ֆունկցիան համարյա համընկնում է սկզբնական ֆունկցիայի հետ, և սխալանքը դառնում է զրո:

Հետևաբար բազային ալգորիթմների քանակի ավելացմանը զուգընթաց՝ մեծանում է մոտարկման ճշտությունը:

14.4 AnyBoost ալգորիթմը

Դիցուք նորից $Y = \{-1, +1\}$, $h_t : X \rightarrow R$, $L(M)$ - հենքային M ֆունկցիայի վերևից մոտարկող ողորկ ֆունկցիա է, $M_T(x_i) = y_i \sum_{t=1}^T \alpha_t h_t(x_i)$ -կոմպոզիցիայի շեղումն է x_i

օբյեկտի վրա: Խնդիր է դրվում մինիմիզացնել կոմպոզիցիայի սխալները ամբողջ հավաքածուի վրա: Ունենք

$$Q \leq \tilde{Q} = \sum_{i=1}^l L(M_{T-1}(x_i) + \alpha_T y_i h_T(x_i)):$$

Կորստի L ֆունկցիան դիտարկենք որպես α_T պարամետրից կախված ֆունկցիա՝

$$\lambda(\alpha_T) = L(M_{T-1}(x_i) + y_i \alpha_T h_T(x_i)):$$

Գծայնացնենք այս ֆունկցիան $\alpha_T = 0$ կետի շրջակայքում, վերլուծելով այն Թեյլորի բանաձևի և վերցնելով միայն գծային մասը՝

$$\lambda(\alpha_T) \approx \lambda(0) + \alpha_T \lambda'(0):$$

Այստեղից հետևում է, որ

$$\tilde{Q}_T \approx \sum_{i=1}^l L(M_{T-1}(x_i)) - \alpha_T \sum_{i=1}^l -L'(M_{T-1}(x_i)) y_i h_T(x_i):$$

Նշանակենք

$$w_i = -L'(M_{T-1}(x_i)) y_i h_T(x_i),$$

որոնք կոչվում են օբյեկտների կշռային գործակիցներ:

Եթե α_T պարամետրը ֆիքսված է, ապա $\tilde{Q}_T$ մինիմիզացիայի համար պետք է գտնել այնպիսի բազային h_T ալգորիթմ, որ հանդիսանա հետևյալ խնդրի լուծումը՝

$$\sum_{i=1}^l w_i y_i h(x_i) \rightarrow \max_h:$$

Այնուհետև $\alpha_T > 0$ պարամետրի օպտիմալ արժեքը գտնելու համար պետք է լուծել միաչափ մինիմիզացիայի խնդիր $\tilde{Q}$ ֆունկցիայի նկատմամբ:

Նշենք, որ մասնավոր դեպքում, երբ

$$h_t : X \rightarrow \{-1, 0, +1\}, \quad L(M) = e^{-M},$$

AnyBoost ալգորիթմը վեր է ածվում *Adaboost*-ի:

Այսպիսով, *Anyboost* ալգորիթմը կատարվում է քայլերի հետևյալ հերթականությամբ:

- Ալգորիթմի մուտքին տրվում է S հավաքածուն և T պարամետրը:
- Ելքում պետք է ստանալ բազային ալգորիթմները և կշիռները՝ $\alpha_t, \quad h_t, \quad t = 1, 2, \dots, T$:
- Բոլոր $t = 1, 2, \dots, T$ իտերացիաների ժամանակ կատարել հետևյալ քայլերը:

- Կատարվում է լուսանցքների ինիցիալիզացիա՝

$$M_i = 0, \quad i = 1, 2, \dots, l:$$

- Հաշվում ենք օբյեկտների կշիռները՝

$$w_i = -L'(M_i), \quad i = 1, 2, \dots, l:$$

- Գտնում ենք բազային h_t ալգորիթը՝

$$h_t = \arg \max_h \sum_{i=1}^l w_i y_i h(x_i):$$

- Լուծում ենք միաչափ օպտիմիզացիայի խնդիր՝

$$\alpha_t = \arg \min_{\alpha > 0} \sum_{i=1}^l L(M_i + \alpha h_t(x_i) y_i):$$

- Կատարվում է լուսանցքների թարմացում՝

$$M_i := M_i + \alpha_t h_t(x_i) y_i, \quad i = 1, 2, \dots, l:$$

Այժմ դիտարկենք մասնավոր դեպք և որպես կորստի ֆունկցիա վերցնենք

$$L(z) = (1 - z)^2:$$

Դիտարկենք *adaboost* ալգորիթումում դիտարկված օրինակը:

- Կատարվում է լուսանցքների ինիցիալիզացիա՝ $M_i = 0, i = 1, 2, \dots, 7:$
- Հաշվում ենք օբյեկտների կշիռները՝

$$w_i = -L'(M_i) = 2(1 - M_i) = 2, \quad i = 1, 2, \dots, 7:$$

- Կառուցում ենք բազային h_1 ալգորիթը՝

$$h_1 = \arg \max_h \sum_{i=1}^l w_i y_i h(x_i):$$

Adaboost ալգորիթումում քննարկված օրինակի դեպքում կունենանք

$$\sum_{i=1}^7 w_i y_i h_1(x_i) = 2(1 - 1 - 1 - 1 + 1 + 1 + 1) = 2:$$

Նման կերպ

$$\sum_{i=1}^7 w_i y_i h_2(x_i) = \sum_{i=1}^7 w_i y_i h_3(x_i) = 2 :$$

Այսպիսով, որպես h_1 կարող ենք ընտրել h_1, h_2, h_3 պրեդիկտորներից որևէ մեկը: Ընտրենք առաջինը:

- Գտնում ենք h_1 պրեդիկտորի կշռային α_1 գործակիցը՝

$$\alpha_1 = \arg \min_{\alpha > 0} \sum_{i=1}^7 L(M_i + \alpha h_1(x_i) y_i) =$$

$$\arg \min_{\alpha > 0} (\alpha h_1(x_i) - 1)^2 = \arg \min_{\alpha > 0} 4(\alpha - 1)^2 + 3(\alpha + 1)^2 = \frac{1}{7} :$$

- Կատարվում է լուսանցքների թարմացում՝

$$M_1 := M_1 + \alpha_1 y_1 h_1(x_1) = 0 + \frac{1}{7} = 0.13 :$$

Նման կերպ ստանում ենք

$$M_2 = M_3 = M_4 = -0.13, \quad M_5 = M_6 = M_7 = 0.13 :$$

Անցնում ենք **երկրորդ** իտերացիային:

- Հաշվում ենք օբյեկտների նոր կշիռները՝

$$w_1 = w_5 = w_6 = w_7 = 1.74, \quad w_2 = w_3 = w_4 = 2.26 :$$

- Կառուցում ենք b_2 բազային պրեդիկտորը: Ունենք

$$\sum_{i=1}^7 w_i y_i h_1(x_i) = 4 \times 1.74 - 3 \times 2.26 = 0.18,$$

$$\sum_{i=1}^7 w_i y_i h_2(x_i) = (-1.74 - 1.74 - 2.26 + 2.26 + 2.26 + 1.74 + 1.74) = 2.26,$$

$$\sum_{i=1}^l w_i y_i h_3(x_i) = -1.74 - 2.26 - 1.74 + 2.26 + 2.26 + 1.74 + 1.74 = 2.26 :$$

Այսպիսով, որպես երկրորդ h_2 պրեդիկտոր կարող ենք վերցնել h_2 պրեդիկտորը:

- Հաշվենք այդ պրեդիկտորի կշռային α_2 գործակիցը:

Ունենք

$$\sum_{i=1}^7 L(M_i + \alpha h_2(x_i)) = 4(\alpha - 0.87)^2 + 6(\alpha - 1.13)^2 :$$

Հետևաբար՝

$$\alpha_2 = 0.98:$$

Այսպիսով, երկու իտերացիայից հետո կառուցվում է

$$a(x) = \alpha_1 h_1(x) + \alpha_2 h_2(x) = \frac{1}{7} h_1(x) + 0.98 h_2(x):$$

ալգորիթմը: Այս նոր ալգորիթմի համար հաշվենք նրա կորստի չափը ամբողջ հավաքածուի վրա.

$$a(x_1) = 0.13 h_1(x_1) + 0.98 h_2(x_1) =$$

$$= 0.13 - 0.98 = -0.85,$$

$$a(x_2) = -0.13 + 0.98 = 0.85,$$

$$a(x_3) = -0.13 - 0.98 = -1.11,$$

$$a(x_4) = -0.13 + 0.98 = 0.85,$$

$$a(x_5) = 0.13 - 0.98 = -0.85,$$

$$a(x_6) = 0.13 + 0.98 = 1.11,$$

$$a(x_7) = 0.13 + 0.98 = 1.13:$$

Այստեղից՝

$$Q_2 = \sum_{i=1}^7 [y_i \sum_{t=1}^2 \alpha_t h_t(x_i) < 0] = 2:$$

Նշենք նաև, որ h_1, h_2, h_3 ալգորիթմներից յուրաքանչյուրը հավաքածուի վրա կատարում է երեք սխալ կանխատեսում: Իսկ *Anyboost*-ով կառուցած պրեդիկտորը սխալվում է երկու անգամ: Իսկ *adaboost*-ով կառուցված պրեդիկտորը հավաքածուի վրա երկու իտերացիայից հետո ոչ մի սխալ կանխատեսում չի անում:

14.5 Գրադիենտային բուստինգ

Դիտարկենք էմպիրիկ ռիսկի մինիմիզացիայի խնդիրը հիպոթեզների ինչ-որ H բազմության վրա, որի յուրաքանչյուր էլեմենտ ունի հետևյալ տեսքը՝

$$a(x) = \sum_{t=1}^T \alpha_t h_t(x):$$

Այսպիսով, ունենք հետևյալ խնդիրը՝

$$\sum_{i=1}^l L(y_i, a(x_i)) \rightarrow \min_{a \in H}:$$

Այս խնդրի լուծման համար կառուցվում է հետևյալ ռեկուրենտ առնչությունը՝

$$a_k = a_{k-1} + \alpha_k h_k,$$

որտեղ (α_k, h_k) գույգը ընտրվում է հետևյալ պայմանից՝

$$(\alpha_k, h_k) \in \arg \min_{\alpha > 0, h \in H} \sum_{i=1}^l L(y_i, a_{k-1}(x_i) + \alpha h(x_i)):$$

Այս խնդիրը լուծվում է մոտավոր: $\mathbb{R}^n$ տարածության $a_{k-1}(x_i)$ կետից կատարում ենք գրադիենտային իջեցում $-(L'_y(y_i, a_{k-1}(x_i)))$ վեկտորի ուղղությամբ: Այնուհետև բազային ալգորիթների դասից ընտրում ենք այն պրեդիկտորը, որը միջին քառակուսայինի իմաստով լավ մոտարկում է այդ գրադիենտին՝

$$h_k \in \arg \min_{h \in H} \sum_{i=1}^l (l'_y(y_i, a_{k-1}(x_i)) + h(x_i))^2:$$

Այնուհետև α_k քայլի ընտրման համար լուծվում է միաչափ օպտիմիզացիայի հետևյալ խնդիրը՝

$$\alpha_k \in \arg \min_{\alpha \geq 0} \sum_{i=1}^l L(y_i, a_{k-1}(x_i) + \alpha h_k):$$

Այսպիսով, գրադիենտային բուստինգի ընդհանուր ալգորիթմը հետևյալն է՝

- Վերցնում ենք որևէ պրեդիկտոր $a_0 \in H$ (օրինակ $a_0 = 0$):
- Կառուցում ենք a_{k+1} պրեդիկտորը հետևյալ ռեկուրենտ առնչությամբ՝

$$a_{k+1} = a_k + \alpha_k h_k, \quad k = 0, 1, 2, \dots, T-1,$$

որտեղ α_k, h_k -ն ընտրում ենք վերը նշված առնչություններով: Նշենք, որ գոյություն ունեն α_k քայլի ընտրության տարբեր եղանակներ, որոնք մանրամասնորեն քննարկվել են գրադիենտային իջեցման մեթոդներում:

Դիտարկենք մասնավոր դեպք, երբ կորստի ֆունկցիան ունի հետևյալ տեսքը՝

$$L(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2:$$

Քանի որ $L'_y = \hat{y} - y$, ապա

$$h_k \in \arg \min_{h \in H} \sum_{i=1}^l (y_i - a_{k-1}(x_i) - h(x_i))^2:$$

Սա նշանակում է՝ յուրաքանչյուր խտրացիայում նոր պրեդիկտորը կոմպետացնում է ընթացիկ սխալանքը՝ $(y_i - a_{k-1}(x_i))_{i=1}^l$: Իսկ քայլի համար ստանում

ենք հետևյալ բանաձևը՝

$$\alpha_k \in \arg \min_{\alpha \geq 0} \sum_{i=1}^l (y_i - a_{k-1}(x_i) - \alpha h(x_i))^2 =$$

$$= \frac{\sum_{i=1}^l (y_i - a_{k-1}(x_i)) h_k(x_i)}{\sum_{i=1}^l h_k(x_i)^2} :$$

Դիտարկենք գրադիենտային բուստինգի մի օրինակ, որտեղ հիպոթեզների բազմության տարրերը որոշիչ ծառեր են: Տվյալները ներկայացված են աղյուսակ 14.2- ում:

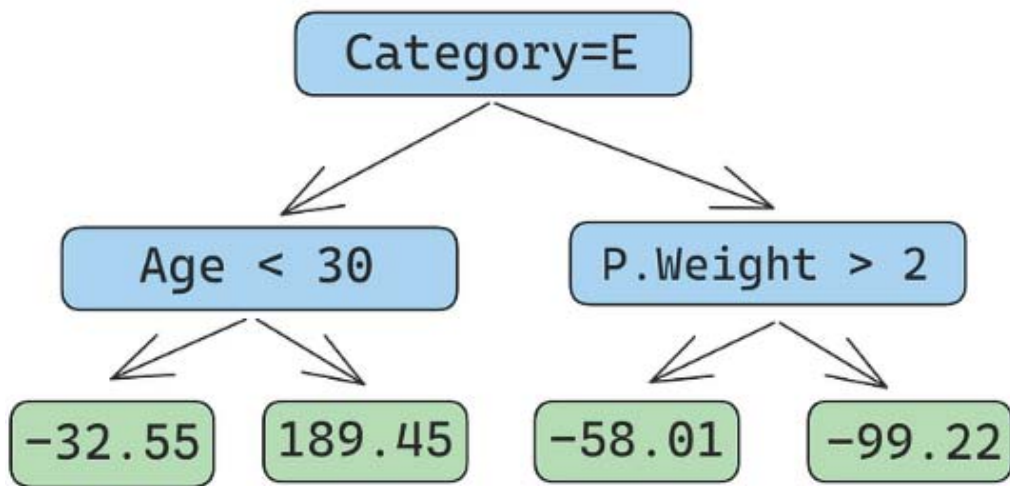
Աղյուսակ 14.2			
Տարիքը age	Տեսակը Category	Քաշը Weight(kg)	Գումարը amount (USD)
25	Էլեկտրոնիկա	2.5	123.45
34	հագուստ	1.3	56.78
42	Էլեկտրոնիկա	5.0	345.67
19	կենցաղ. ապրանք	3.2	98.01

Որպես սկզբնական պրեդիկտոր վերցնում ենք այն ալգորիթմը, որը բոլոր օբյեկտների վրա տալիս է ռեալ արժեքների միջինին հավասար կանխատեսում՝

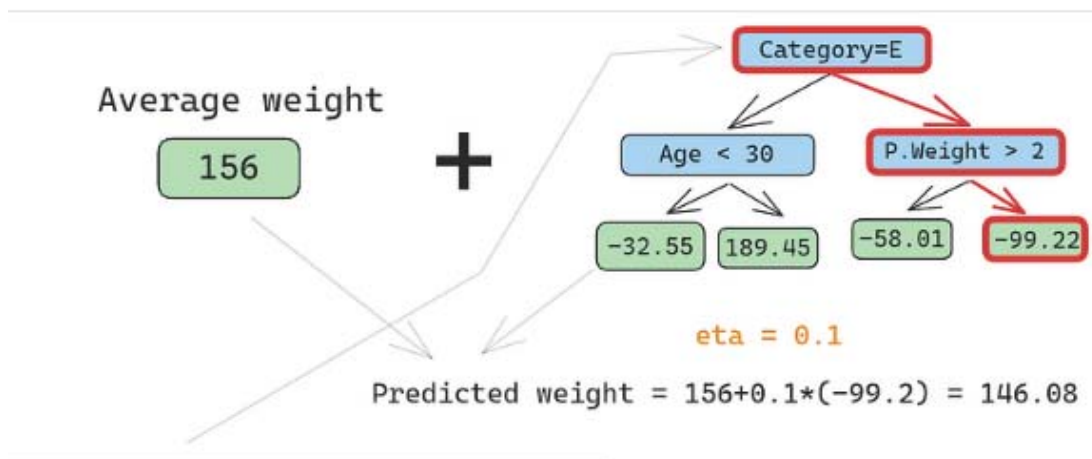
$$a_0(x) = \frac{1}{4}(123.45 + 56.78 + 345.67 + 98.01) = 156:$$

Այնուհետև հաշվում ենք պրեդիկտորի սխալանքները յուրաքանչյուր օբյեկտի վրա: Այն տրված է աղյուսակ 14.3-ում:

Աղյուսակ 14.3, կանխ. սխալանքները				
Տարիքը age	Տեսակը Category	Քաշը Weight(kg)	Գումարը amount (USD)	Շեղումներ residuals
25	Էլեկտրոնիկա	2.5	123.45	-32.55
34	հագուստ	1.3	56.78	-99.22
42	Էլեկտրոնիկա	5.0	345.67	189.45
19	կենցաղ. ապրանք	3.2	98.01	-58.01



Նկ. 14.8: Որոշիչ ծառ



Նկ. 14.9: Նոր պրեդիկտորի կառուցման սխեման

Այնուհետև կառուցում ենք որոշիչ ծառ (կաահմանափակվենք չորս տերևանի ծառերով, (նկ.14.8):

Դրանից հետո հաշվում ենք գրադիենտային բուստինգի քայլը՝ ընտրելով այն հավասար 0.1:

Այնուհետև կառուցում ենք նոր պրեդիկտոր հետևյալ բանաձևով՝

$$a_1(x) = a_0(x) + 0.1h_0(x) \quad (\text{տե՛ս նկ.14.9}):$$

Յուրաքանչյուր օբյեկտի համար նոր կանխատեսումները և շեղումները տրված են աղյուսակ 14.4-ում:

Աղյուսակ 14.4, նոր կանխատեսում					
Տարիքը age	Տեսակը Category	Քաշը Weight(kg)	Գումարը amount (USD)	Ծեղումներ residuals	Նոր կանխա.
25	Էլեկտրոնիկա	2.5	123.45	-32.55	152.745
34	հագուստ	1.3	56.78	-99.22	146.078
42	Էլեկտրոնիկա	5.0	345.67	189.45	174.945
19	կենցաղ. ապրանք	3.2	98.01	-58.01	150.199

Ինչպես նկատելի է, շեղումները նվազում են:

Այժմ դիտարկենք գրադիենտային բուստինգի մի մոդիֆիկացիա և քննարկենք դրա զուգամիտության խնդիրը:

Դիտարկենք հետևյալ ալգորիթմը.

- 1 Input $S_n = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, ψ , $L(\cdot, \cdot)$, γ .
- 2 Output $h^{(T)}(x)$.
- 3 $h^{(0)}(x) = 0$. for $t = 1, 2, \dots, T$ do
- 4 $g_i = L'_1(h^{(t-1)}(x_i), y_i)$, $i = 1, 2, \dots, n$, (L'_1 -ը $L(h, y)$ ֆունկցիայի գրադիենտն է ըստ առաջին արգումենտի),
- 5 $\theta_t = \arg \min_{\theta \in \Theta} \sum_{i=1}^n (\gamma \eta_t \psi(\theta, x_i) + g_i)^2$,
- 6 $h^t(x) = (1 - \eta_t)h^{(t-1)}(x) + \eta_t \psi(\theta_t, x)$:

Տեղի ունի հետևյալ թեորեմը:

Թեորեմ 14.3. Դիցուք

- $\sup_{\theta \in \Theta} \|\psi(\theta, \cdot)\|_{L_2(S_n)} \leq B$,
- $L(h, y)$ ֆունկցիան ըստ առաջին արգումենտի γ -նդորկ է և ուռուցիկ,
- $\eta_t = 2/(t + 1)$:

Այդ դեպքում տեղի ունի հետևյալ գնահատականը՝

$$\frac{1}{n} \sum_{i=1}^n L(h^{(T)}(x_i), y_i) \leq \inf_{h \in \text{conv}(\psi)} \frac{1}{n} \sum_{i=1}^n L(h(x_i), y_i) + \frac{10\gamma B}{T + 1},$$

որտեղ

$$\text{conv}(\psi) = \left\{ \sum_{j=1}^m u_j \psi(\theta_j, x) : m > 0, \|u\|_1 = 1, u_j \geq 0, \theta_j \in \Theta \right\}:$$

Դիտարկենք

$$h(x) = \sum_{i=1}^m u_i \psi(\tilde{\theta}_i, x), \sum_{j=1}^m u_j = 1, u_j \geq 0, j = 1, 2, \dots, m$$

Ֆունկցիան: Ըստ θ_t -ի սահմանման՝ ունենք

$$2\eta_t \sum_{i=1}^n g_i \psi(\theta_t, x_i) + \gamma \eta_t^2 \sum_{i=1}^n \psi(\theta_t, x_i)^2 \leq$$

$$2\eta_t \sum_{i=1}^n g_i \psi(\tilde{\theta}_j, x_i) + \gamma \eta_t^2 \sum_{i=1}^n \psi(\tilde{\theta}_j, x_i)^2:$$

Այս անհավասարության երկու կողմերը բազմապատկելով u_j -ով և գումարելով դրանք ըստ j -ի՝ կստանանք

$$2\eta_t \sum_{i=1}^n g_i \psi(\theta_t, x_i) + \gamma \eta_t^2 \sum_{i=1}^n \psi(\theta_t, x_i)^2 \leq$$

$$2\eta_t \sum_{i=1}^n g_i \sum_{j=1}^m u_j \psi(\tilde{\theta}_j, x_i) + \gamma \eta_t^2 \sum_{i=1}^n \sum_{j=1}^m u_j \psi(\tilde{\theta}_j, x_i)^2:$$

Հետևաբար՝

$$\eta_t \sum_{i=1}^n g_i \psi(\theta_t, x_i) \leq \eta_t \sum_{i=1}^n h(x_i) + \frac{n\gamma}{2} \eta_t^2 B^2: \quad (14.1)$$

Այստեղից կստանանք

$$\sum_{i=1}^b L(h^{(t)}(x_i), y_i) = \sum_{i=1}^n L((1 - \eta_t)h^{(t-1)}(x_i) + \eta_t \psi(\theta_t, x_i), y_i) \leq$$

$$\sum_{i=1}^n [L(h^{(t-1)}(x_i), y_i) + \eta_t g_i (\psi(\theta_t, x_i) - h^{(t-1)}(x_i))] + \frac{\eta_t^2 \gamma}{2} \sum_{i=1}^n (\psi(\theta_t, x_i) - h^{(t-1)}(x_i))^2 \leq$$

$$\leq \sum_{i=1}^n [L(h^{(t-1)}(x_i), y_i) + \eta_t g_i (h(x_i) - h^{(t-1)}(x_i))] + \frac{n\gamma \eta_t^2 B}{2} + \frac{n\gamma \eta_t^2}{2} 4B^2 \leq$$

$$\sum_{i=1}^n L(h^{(t-1)}(x_i), y_i) + \eta_t (L(h(x_i), y_i) - L(h^{(t-1)}(x_i), y_i)) + \frac{5n\gamma \eta_t^2 B^2}{2}: \quad (14.2)$$

Այստեղ առաջին անհավասարության մեջ օգտագործվել է այն փաստը, որ $L(h, y)$ ֆունկցիան ողորկ է ըստ առաջին արգումենտի՝

$$L(h', y) \leq L(h, y) + L'_1(h, y)(h' - h) + 0.5\gamma(h - h')^2:$$

Երկրորդ անհավասարության մեջ հաշվի է առնված (14.1) անհավասարությունը: Իսկ երրորդ անհավասարությունը բխում է ուռուցիկ ֆունկցիայի հինական հետևյալ անհավասարությունից՝

$$L(h, y) + L'_1(h, y)(h' - h) \leq L(h', y):$$

Նշանակենք

$$\beta_t = \frac{1}{n} \sum_{i=1}^n L(h^t(x_i), y_i) - \frac{1}{n} \sum_{i=1}^n L(h(x_i), y_i):$$

Նկատի ունենալով, որ $\eta_t = 2/(t+1)$, (14.2) անհավասարությունից՝ կստանանք

$$\beta_t \leq \frac{t-1}{t+1} \beta_{t-1} + \frac{10\gamma B^2}{(t+1)^2}:$$

Բազմապատկելով այս անհավասարության երկու մասերը $t(t+1)$ -ով՝ կստանանք՝

$$t(t+1)\beta_t \leq (t-1)t\beta_{t-1} + 10\gamma B^2:$$

Գումարելով այս անհավասարությունները $t = 1, 2, \dots, T$ արժեքների համար՝ կստանանք պահանջվող անհավասարությունը: Վերջում նշենք, որ [21] և [22] աշխատանքներում քննարկվում են նաև գրադիենտային բուստինգի ուրիշ մոդիֆիկացիաներ և դրանց զուգամիտության պայմանները:

15

Առաջարկությունների համակարգեր

15.1 Մատրիցների սինգուլյար վերլուծություն (SVD-singular value decomposition)

Դիցուք $A(m \times n)$ ուղղանկյուն մատրից է: Մատրիցային անալիզից հայտնի են հետևյալ փաստերը.

- $AA^T(m \times m)$, $A^T A(n \times n)$ մատրիցները դրական կիսաորոշյալ սիմետրիկ մատրիցներ են,
- $\text{rank}(AA^T) = \text{rank}(A^T A) = r$, որտեղ $r = \text{rank}(A)$.
- Այս մատրիցների սեփական արժեքները ոչ բացասական թվեր են: Ընդ որում ոչ զրոյական սեփական արժեքների թիվը հավասար է $\text{rank} A = r$,
- այս քառակուսի մատրիցների սպեկտրը ներկայացված է 15.1 նկարում:

Դիցուք $A^{m,n}$ մատրիցի ռանգը հավասար է r -ի: $v \in \mathbb{R}^n$ և $u \in \mathbb{R}^m$ միավոր վեկտորները կոչվում են A մատրիցի աջակողմյան և ձախակողմյան սինգուլյար վեկտորներ (համապատասխանաբար), եթե գոյություն ունի այնպիսի $\sigma > 0$ թիվ, որ

$$Av = \sigma u, \quad A^T u = \sigma v:$$

$A \cdot A^T$	$A^T A$
$\{\lambda_1, \dots, \lambda_r, \underbrace{0, \dots, 0}_{m-r}\}$	$\{\lambda_1, \dots, \lambda_r, \underbrace{0, \dots, 0}_{n-r}\}$

Նկ. 15.1: Մատրիցների սպեկտր

$$\Sigma = \left(\begin{array}{ccc|ccc} \sigma_1 & & & & & \\ & \sigma_2 & & & & \\ & & \ddots & & & \\ & & & \sigma_r & & \\ \hline & & & & \mathbb{O}_{(m-r) \times (n-r)} & \end{array} \right)$$

Նկ. 15.2: Մատրիցի սինգուլյար ներկայացում

Թեորեմ 15.1. Դիցուք $\text{rank}(A) = r$: Այդ դեպքում A մատրիցը կարելի է ներկայացնել հետևյալ տեսքով՝

$$A = U \Sigma V^T,$$

որտեղ

ա) $U(m \times m)$, $V(n \times n)$ օրթոգոնալ մատրիցներ են՝ $U \times U^T = E$, $V \times V^T = E$:

բ) Σ մատրիցի կառուցվածքը տրված է նկար 15.2-ում: Այս մատրիցի ոչ զրոյական էլեմենտներն են $\sigma_{jj} = \sigma_j$, $j = 1, 2, \dots, r$, որտեղ

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0, \quad \sigma_i = \sqrt{\lambda_i}, \quad i = 1, 2, \dots, r:$$

λ_i , $i = 1, 2, \dots, r$ թվերը AA^T մատրիցի սեփական արժեքներն են:

գ) U մատրիցի սյուներն AA^T մատրիցի սեփական վեկտորներն են: Իսկ V մատրիցի սյուները $A^T A$ մատրիցի սեփական վեկտորներն են:

Թեորեմի ապացույցը հենվում է մի քանի լեմմաների վրա: Նախ մենք ցույց կտանք, որ եթե կարելի է գտնել r օրթոնորմալ սինգուլյար վեկտորներ, ապա A մատրիցը կարելի է ներկայացնել $A = U \Sigma V^T$ տեսքով, որտեղ U և V մատրիցների սյուները բաղկացած են համապատասխանաբար մատրիցի աջակողմյան և ձախակողմյան սինգուլյար վեկտորներից: Իսկ $\Sigma^{r,r}$ -ն անկյունագծային մատրից է, որի անկյունագծի վրա դասավորված են սինգուլյար արժեքները:

Լեմմա 15.1. Դիցուք մատրիցի ռանգը հավասար է r -ի: Ենթադրենք $v_1, v_2, \dots, v_r$ վեկտորները A մատրիցի աջակողմյան սինգուլյար օրթոնորմալ վեկտորների համախումբ է, $u_1, u_2, \dots, u_r$ ՝ մատրիցի ձախակողմյան սինգուլյար օրթոնորմալ վեկտորների բազմություն, իսկ $\sigma_1, \sigma_2, \dots, \sigma_r$ արժեքները համապատասխան սինգուլյար արժեքներն են: Այդ դեպքում մատրիցը կարելի է ներկայացնել հետևյալ տեսքով՝

$$A = U \Sigma V^T, \quad \Sigma_{i,i} = \sigma_i:$$

Ապացույց: Ամեն մի աջակողմյան սինգուլյար վեկտոր պատկանում է A^T մատրիցի պատկերին (հակառակ դեպքում սինգուլյար թիվը հավասար կլինի զրոյի): Հետևաբար՝ $v_1, v_2, \dots, v_r$ վեկտորների համախումբը օրթոնորմալ բազիս է կազմում $\text{Im} A^T$ -ում: Լրացնենք այն $v_{r+1}, \dots, v_n$ վեկտորներով այնպես, որ $v_1, v_2, \dots, v_n$

համախումբը լինի օրթոնորմալ բազիս $\mathbb{R}^n$ -ում: Դիցուք $B = \sum_{i=1}^r \sigma_i u_i v_i^T$: Այժմ ցույց տանք, որ $Av_i = Bv_i$, $i = 1, 2, \dots, n$: Ակնհայտ է, որ եթե $i > r$, ապա $Av_i = Bv_i = 0$: Իսկ եթե $i \leq r$, ապա

$$Bv_i = \sum_{j=1}^r \sigma_j u_j v_j^T v_i = \sigma_i u_i = Av_i:$$

Լեմման ապացուցվեց:

Հաջորդ լեմման կապ է հաստատում A մատրիցի սինգուլյար թվերի և $A^T A$ ու AA^T մատրիցի սեփական արժեքների միջև:

Լեմմա 15.2. v, u վեկտորները A մատրիցի աջակողմյան և ձախակողմյան սինգուլյար վեկտարներն են σ սինգուլյար արժեքով այն և միայն այն ժամանակ, երբ v -ն $A^T A$ մատրիցի սեփական վեկտորն է σ^2 սեփական արժեքով, իսկ $u = \sigma^{-1} Av$ վեկտորը AA^T մատրիցի սեփական վեկտորն է σ^2 սեփական արժեքով:

Ապացույց: Դիցուք v -ն աջակողմյան սինգուլյար վեկտոր է σ սինգուլյար արժեքով: Այդ դեպքում

$$A^T Av = \sigma A^T u = \sigma^2 v:$$

Նմանապես կունենանք

$$AA^T u = \sigma Av = \sigma^2 v:$$

Եվ ընդհակառակը, եթե $\lambda \neq 0$ թիվը $A^T A$ մատրիցի v սեփական վեկտորին համապատասխան սեփական արժեքն է, ապա $\sigma = \sqrt{\lambda}$, $u = \sigma^{-1} Av$ -ի դեպքում կունենանք՝

$$\sigma u = \sqrt{\lambda} \frac{Av}{\sqrt{\lambda}} = Av$$

և

$$A^T u = \frac{1}{\sigma} A^T Av = \frac{\lambda}{\sigma} v = \sigma v:$$

Վերջապես առանց ապացույցի ընդունենք այն փաստը, որ եթե մատրիցի ռանգը հավասար է r -ի, ապա գոյություն ունեն r հատ օրթոնորմալ սինգուլյար վեկտորներ:

Այս երեք պնդումներից անմիջապես հետևում են թեորեմ 15.1-ի եզրակացությունները: Դիտարկենք թվային օրինակ:

Կառուցենք հետևյալ մատրիցի սինգուլյար վերլուծությունը՝

$$\begin{pmatrix} 3 & 0 \\ 4 & 5 \end{pmatrix}:$$

Ունենք

$$AA^T = \begin{pmatrix} 9 & 12 \\ 12 & 41 \end{pmatrix}, \quad A^T A = \begin{pmatrix} 25 & 20 \\ 20 & 25 \end{pmatrix}:$$

Գտնենք դրանց սեփական արժեքները՝ $\lambda_1 = 45$, $\lambda_2 = 5$:

Հետևաբար՝

$$\begin{pmatrix} 25 & 20 \\ 20 & 25 \end{pmatrix} \begin{pmatrix} -11 \\ 1 \end{pmatrix} = 45 \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

$$\begin{pmatrix} 25 & 20 \\ 20 & 25 \end{pmatrix} \begin{pmatrix} -1 \\ 1 \end{pmatrix} = 5 \begin{pmatrix} -1 \\ 1 \end{pmatrix}:$$

Ուրեմն այդ սեփական արժեքներին համապատասխան միավոր սեփական վեկտորներն են՝

$$v_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad v_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 1 \end{pmatrix}:$$

Այս վեկտորները A մատրիցի ձախակողմյան սինգուլյար վեկտորներն են: Հետևաբար՝

$$V = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}:$$

Այժմ գտնենք A մատրիցի աջակողմյան սինգուլյար վեկտորները: Ունենք

$$\sigma_1 = \sqrt{45}, \quad \sigma_2 = \sqrt{5}, \quad Av_1 = \frac{3}{\sqrt{2}} \begin{pmatrix} 1 \\ 3 \end{pmatrix} = \sqrt{45} \frac{1}{\sqrt{10}} \begin{pmatrix} 1 \\ 3 \end{pmatrix} = \sigma_1 u_1,$$

$$Av_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} -3 \\ 1 \end{pmatrix} = \sqrt{5} \frac{1}{\sqrt{10}} \begin{pmatrix} -3 \\ 1 \end{pmatrix} = \sigma_2 u_2:$$

Այստեղից՝

$$U = \frac{1}{\sqrt{10}} \begin{pmatrix} 1 & -3 \\ 3 & 1 \end{pmatrix}:$$

Վերջնականում կունենանք

$$A = U \Sigma V^T,$$

որտեղ

$$\Sigma = \begin{pmatrix} \sqrt{45} & 0 \\ 0 & \sqrt{5} \end{pmatrix}:$$

Այժմ առանց ապացույցի բերենք մատրիցային անալիզի Էքարդ-Յանգ-Միրսկի թեորեմը, որում r ռանգի մատրիցը մոտարկվում է k ռանգ ունեցող մատրիցներով ($k < r$):

$$\Sigma_k = \left(\begin{array}{ccc|c} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_k \\ \hline & & & \end{array} \begin{array}{c} \mathbb{O}_{k \times (n-k)} \\ \\ \\ \mathbb{O}_{(m-k) \times (n-k)} \end{array} \right)$$

Նկ. 15.3: Σ_k մատրիցի կառուցվածքը:

Տեղի ունի հետևյալ պնդումը:

Թեորեմ 15.2. (Էքսթրեմ-Յանգ-Միրսկի): Դիցուք $\text{rank}(A(m \times n)) = r$: Դիտարկենք հետևյալ խնդիրը՝

$$\min_{\tilde{A}(m \times n)} \|A - \tilde{A}\|, \text{rank}(\tilde{A}) = k < r:$$

Այս խնդիրը ունի լուծում և այն հետևյալ մատրիցն է՝

$$A_k = U \Sigma_k V,$$

որտեղ Σ_k մատրիցի կառուցվածքը տրված է նկար 15.3-ում:

Անկյունագծի վրա դասավորված են A մատրիցի առաջին k սինգուլյար թվերը: Իսկ ֆունկցիոնալի մինիմալ արժեքը հավասար է

$$\sqrt{\sum_{j=k+1}^r \sigma_j^2}:$$

Այսինքն՝ տեղի ունի հետևյալ հավասարությունը՝

$$A = \sum_{j=1}^r u_j v_j^T,$$

որտեղ u_j և v_j վեկտորները համապատասխանաբար U և V մատրիցի սյուներ են:

Օրինակի միջոցով մեկնաբանենք այս պնդումը: Դիտարկենք նախորդ օրինակը: Համաձայն թեորեմ 15.2-ի

$$A = \begin{pmatrix} 3 & 0 \\ 4 & 5 \end{pmatrix}$$

մատրիցը, որի ռանգը հավասար է երկուսի, կարող ենք ներկայացնել մեկ ռանգ ունեցող երկու մատրիցի գծային կոմբինացիայի տեսքով հետևյալ կերպ՝

$$A = \begin{pmatrix} 3 & 0 \\ 4 & 5 \end{pmatrix} = \frac{\sqrt{45}}{20} \begin{pmatrix} 1 & 1 \\ 3 & 3 \end{pmatrix} + \frac{\sqrt{5}}{20} \begin{pmatrix} 3 & -3 \\ -1 & 1 \end{pmatrix}:$$

Քանի որ այստեղ փոքր սինգուլյար թիվը $\sqrt{5}$ -ն է, ապա այն վերևի վերլուծության

մեջ հաշվի չառնելով՝ կատանանք

$$A_1 = \frac{\sqrt{45}}{20} \begin{pmatrix} 1 & 1 \\ 3 & 3 \end{pmatrix},$$

մատրիցը, որը ըստ թեորեմ 15.2-ի մեկ ռանգ ունեցող երկչափ մատրիցների մեջ ամենամոտիկն է A մատրիցին:

15.2 SVD մոդելի կիրառությունը առաջարկությունների համակարգերում

Առաջարկությունների համակարգերը այգորիթմներ են, որոնք հաճախորդներին առաջարկվում են տարբեր տիպի ապրանքներ, ֆիլմեր, երաժշտություններ և այլն՝ նպատակ ունենալով ավելացնել եկամուտը:

Այդպիսի համակարգերի ստեղծման տարբեր մոտեցումներ կան:

Ամենալայն մոտեցումներից է SVD մեթոդը:

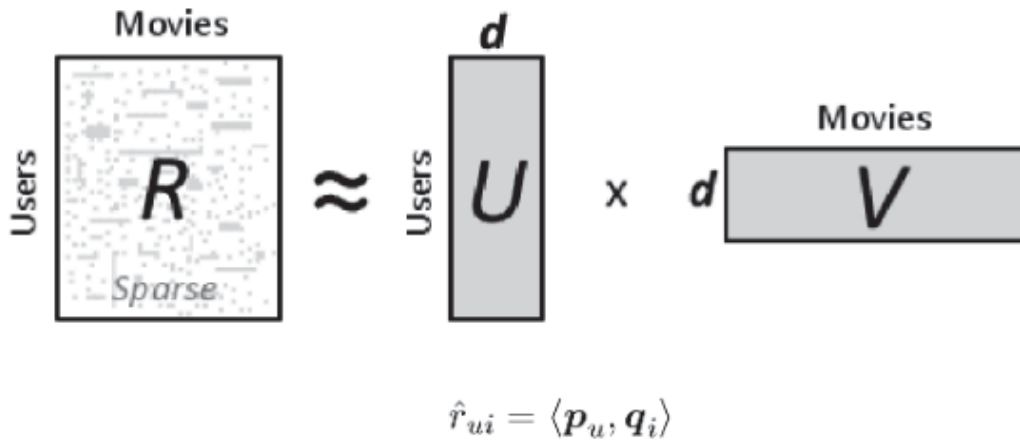
Ենթադրենք՝ կա մի օնլայն կինոթատրոն, և ուզում ենք, որ օգտատերերը լինեն գոհ և նայեն ֆիլմեր այդտեղ: Ստեղծենք ֆիլմերի ռեյտինգների մատրից, օգտատերի գնահատականների հիման վրա (տես աղյուսակ 15.1):

Աղյուսակ 15.1, կինոնկար										
Տերմինատոր										
Ռոբոկոպ										
Ռեմբո										
Չեյմյս Բոնդ										
Մատանիների տիրակալը										
Հոբիտ										
Վրիժառուն										
Տղան Նյու-Յորքում										
Տրոյա										
Ալեքսանդր Երկրի տիրակալը										

Օգտատերեր										
Արմեն	Հակոբ				Տիգրան			Արթուր		

Աղյուսակ 15.2, ֆիլմերի ռեյտինգներ										
9	0	0	6	4	3	5	5	8	6	3
0	0	0	8	0	0	8	5	6	2	0
8	0	5	7	4	0	5	7	8	8	5
5	2	4	10	0	9	0	4	0	4	2

Հարց է առաջանում թե ի՞նչ ֆիլմ առաջարկել, որ մեր օգտատերը լավ երեկո



Նկ. 15.4: Ռեյտինգային մատրիցի սինգուլյար ներկայացումը:

անցկացնի: Համացանցում մի փոքր փնտրելուց հետո մենք գտնում ենք մի հոդված, որում ասվում է, որ մի մեծ առցանց կինոթատրոն՝ Netflix-ը, մրցույթ էր հայտարարել 2006 թվականին՝ ֆիլմերի առաջարկությունների համակարգ մշակելու համար: Հայտարարվել է 1 միլիոն դոլար պարգևատրում այն լուծման համար, որը կբարելավի առաջարկությունների որակը 10 տոկոսով: Արդյունքում 2011 թվականին ստացվեց նման լուծում, և այն աշխատեց SVD-ի հիման վրա՝ singular value decomposition կամ singular matrix decomposition: Կատարենք ռեյտինգի մատրիցի սինգուլյար վերլուծություն: (նկ. 15.4): Այստեղ $\hat{r}_{ui} = \langle p_u, q_i \rangle$ ֆիլմերի ռեյտինգն է, որը ստացվում է U և V մատրիցների բազմապատկումից: Այսպիսով, սկզբնական գնահատման մատրիցը ստացվում է օգտատերերի հետաքրքրությունների և ֆիլմի հատկությունների համար պատասխանատու երկու մատրիցների բազմապատկմամբ: Բայց խնդիրն այն է, որ մեր R մատրիցը նոսր է, և պետք է դատարկ (գրոյական) էլեմենտները լրացնել օգտվողների տված վարկանիշներով: $\hat{r}_{ui}$ գնահատականները ստանալու համար ելնում են հայտնի r_{ui} գնահատականներից, լուծելով օպտիմիզացիոն հետևյալ խնդիրը՝

$$\sum_{u,i \in R} (r_{ui} - \hat{r}_{ui})^2 \rightarrow \min_{p,q}$$

որտեղ գումարը տարածվում է այն u, i ինդեքսներով, որոնց էլեմենտը մատրիցում գրոյ է:

Ռեգրեսիայի այս խնդիրը լուծում են ստոխաստիկ գրադիենտային մեթոդով՝ կատարելով նաև L_2 ռեգուլյարիզացիա՝ $\lambda(\sum_u p_u^2 + \sum_i q_i^2)$: Այս դեպքում որոնվող պարամետրերի թարմացումը կատարվում է հետևյալ ռոկուրենտ առնչություններով՝

$$p_{u,j} = p_{u,j} + \gamma(r_{ui} - \hat{r}_{ui})q_{i,j} - \lambda p_{u,j},$$

$$q_{i,j} = q_{i,j} + \gamma(r_{ui} - \hat{r}_{ui})p_{u,j} - \lambda q_{i,j},$$

որտեղ u, i ինդեքսները համապատասխանաբար U և V մատրիցների տողերի և սյուների համարներն են, իսկ $p_{u,j}$ և $q_{i,j}$ թվերը համապատասխանաբար P_u և Q_i վեկտորների կոմպոնենտներն են:

15.3 Memory-based մոդելներ

Դիցուք

- I -ն նշանակենք ապրանքների (items) բազմությունը,
- U -ն օգտատերերի բազմությունն է,
- R -ը «օգտատեր-ապրանք» զույգերի բազմությունն է, որի համար հայտնի է գնահատականը՝ ռեյտինգը,
- եթե (u, i) զույգի համար հայտնի է գնահատական, ապա գրում են $\exists r_{ui}$,
- գնահատականը ռեյտինգ է, ապրանքների գնման ինդիկատոր է և այլն:

Սահմանվում է օգտատերերի նմանության որոշակի կանոններ;

- I_{uv} -ով նշանակենք այն ապրանքների բազմությունը, որոնք գնահատված են u և v օգտատերերի կողմից՝

$$I_{ui} = \{i \in I : \exists r_{ui}, \exists r_{vi}\}, :$$

- սահմանվում է նմանության գործակից հետևյալ բանաձևով՝

$$w_{uv} = \frac{\sum_{i \in I_{uv}} (r_{ui} - \bar{r}_u)(r_{vi} - \bar{r}_v)}{\sqrt{\sum_{i \in I_{ui}} (r_{ui} - \bar{r}_u)^2} \sqrt{\sum_{i \in I_{vi}} (r_{vi} - \bar{r}_v)^2}},$$

որտեղ $\bar{r}_u$ $\bar{r}_v$ u և v օգտատերերի գնահատականների միջինն է: Այսպիսով, w_{uv} -ն օգտատերերի գնահատականների կորելացիոն գործակիցն է: Որքան այդ գործակիցը մեծ է, այնքան այդ օգտատերերը նման են իրար:

Օգտատերերին առաջարկություններ արվում է «User-based collaborative filtering» սկզբունքով:

- Դիցուք ունենք u_0 օգտատեր:
- Գտնենք բոլոր այն օգտատերերին, որոնք նման են իրեն.

$$U(u_0) = \{v \in U : w_{u_0v} > \epsilon\}:$$

- u_0 -ին առաջարկվում են այն ապրանքները, որոնք հաճախ են գնվել $U(u_0)$ օգտատերերի կողմից:

Դիտարկենք այս մոդելը օրինակի վրա: Դիտարկենք աղյուսակ.15.3-ը:

Աղյուսակ 15.3, User-based collaborative filtering

1	1	0		1	
0	1	1			1
			1	1	0
	1	1		0	
	1				1

Այս աղյուսակի տողերը օգտատերերն են, իսկ սյուները ապրանքները: Եթե $r_{ni} = 1$, դա նշանակում է, որ u օգտատերը «like» է տվել այդ i ապրանքին: Եթե 0 է, ապա այդ ապրանքը նրան դուր չի եկել: Իսկ եթե այդ վանդակում գնահատական չկա, նշանակում է՝ օգտատերը առնչություն չի ունեցել այդ ապրանքի հետ:

Դիցուք u_0 -ն վերջին հինգերորդ օգտատերն է: Ո՞ր ապրանքները առաջին հերթին կարելի է առաջարկել այդ օգտատերին: Հեշտ է տեսնել, որ նրան նման են առաջին, երկրորդ և չորրորդ օգտատերերը: Հետևաբար հինգերորդ օգտատերին պետք է առաջարկել երրորդ ապրանքը, որովհետև այդ ապրանքը առավել շատ «like» է ստացել:

16

Հավելվածներ

16.1 Ավտոմատ դիֆերենցում և դիֆերենցիալ հավասարումների լուծումը ներդրողին ցանցերով

Համակարգչային հանրահաշվում ավտոմատ դիֆերենցումը, կամ AD (ալգորիթմական դիֆերենցում) համակարգչային ծրագրի կողմից սահմանված ֆունկցիայի մասնակի ածանցյալը հաշվելու մեթոդ է: Ավտոմատ դիֆերենցումը նույնիսկ և կենտրոնական գործիք է՝ բարդ ֆունկցիաների և դրանց ածանցյալների թվային արժեքների միաժամանակյա հաշվարկն ավտոմատացնելու համար: Ավտոմատ դիֆերենցումը ո՛չ թվային է, ո՛չ սիմվոլային, ոչ էլ երկուսի համակցություն է: Այն նաև նախընտրելի է սովորական թվային մեթոդներից: Ի տարբերություն վերջավոր տարբերությունների վրա հիմնված ավանդական թվային մեթոդների՝ ավտոմատ դիֆերենցումը «տեսականորեն» ճշգրիտ է:

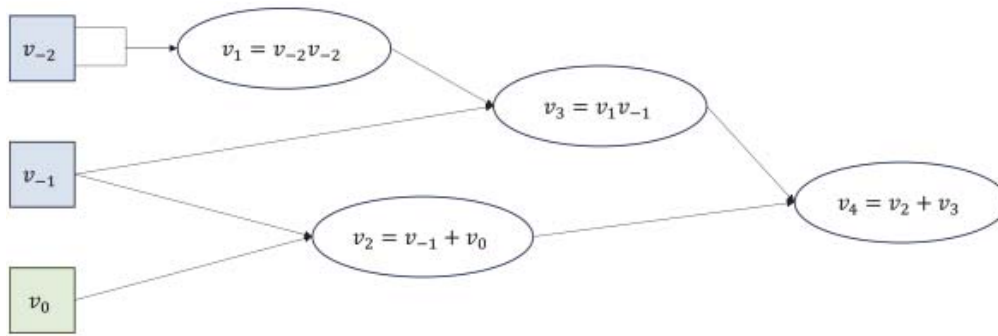
Օգտագործելով բարդ ֆունկցիայի ածանցման շղթայական կանոնը՝ ավտոմատ դիֆերենցումը հաշվարկում է այդ ֆունկցիայի կամայական կարգի մասնակի ածանցյալները: Հաշվարկային այս տեխնիկան օգտագործում են ML-ի խորը ուսուցման ալգորիթմներում:

Այս պարագրաֆում դիտարկելու ենք ֆունկցիայի գրադիենտի հաշվման եղանակ հաշվարկային գրաֆի օգնությամբ: Այդ մեթոդի բացատրությունը տանք օրինակի միջոցով:

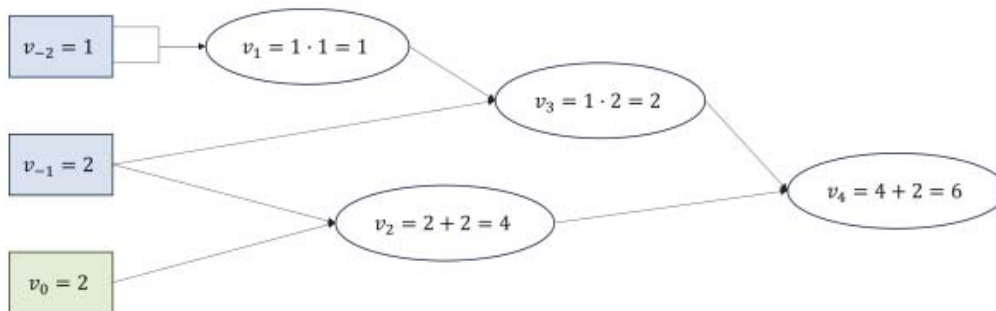
Դիցուք տրված է երկու փոփոխականի $f(x_1, x_2) = x_1^2 x_2 + x_2 + 2$ ֆունկցիան: Պարզ է, որ

$$f'(x) = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2} \right) = (2x_1 x_2, x_1^2 + 1), \quad f'(1, 2) = (4, 2):$$

Դիտարկենք սխեմա, որը նկարագրված է նկար 16.1-ում:



Նկ. 16.1: Հաշվարկային գրաֆ

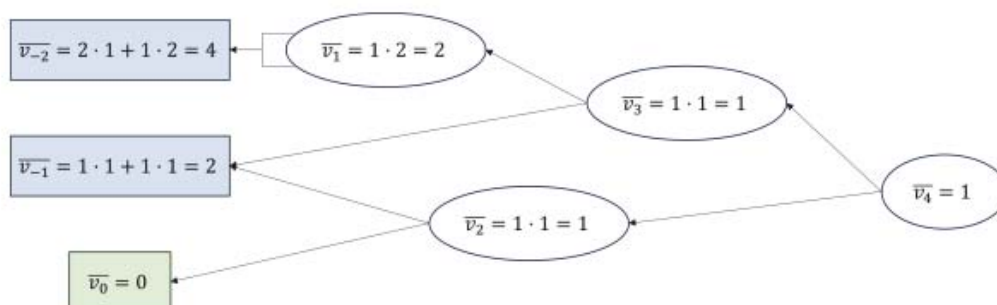


Նկ. 16.2: Ուղիղ փոխանցում

Այն կոչվում է f ֆունկցիայի հաշվարկային գրաֆ: Այստեղ $v_0 = 2$, $v_{-1} = x_2$, $v_{-2} = x_1$:

Այս գրաֆով այսպես ասած «ուղիղ փոխանցումով» (forward pass) կատարվում է ֆունկցիայի արժեքի հաշվարկ՝ արդյունքը գրանցվում է v_4 հանգույցում: նկար 16.2-ում կատարված է այդ հաշվարկը $x = 1, y = 2$ արժեքների դեպքում:

Պարզվում է, որ այդ գրաֆի միջոցով «հետընթաց փոխանցման եղանակով» (backward pass) կարելի է հաշվել ֆունկցիայի գրադիենտի արժեքը նշված կետում: Այն նկարագրված է նկար 16.3-ում: Այստեղ $\bar{v}_i$ փոփոխականները ածանցյալներն են, այսինքն՝ օրինակ, եթե ուզում ենք հաշվել $\bar{v}_{-1} = \frac{df}{dv_{-1}}$ ածանցյալը, ապա դիտարկում ենք v_2 և v_3 փոփոխականներով հանգույցները, որոնք մուտք ունեն v_{-1} հանգույցում և այդ փոփոխականների արտահայտությունները ածանցում ենք ըստ v_{-1} -ի և գումարում իրար: Այնուհետև արդյունքը գրանցում ենք v_{-1} -ի հանգույցում:



Նկ. 16.3: Հետընթաց փոխանցում

Հետևաբար կունենանք՝

1. $\overline{v_4} = \frac{df}{dv_4} = \frac{dv_4}{dv_4} = 1$;
2. $\overline{v_3} = \frac{df}{dv_3} = \frac{df}{dv_4} \frac{dv_4}{dv_3} = \overline{v_4} \frac{dv_4}{dv_3} = 1 \times \frac{d(v_2+v_3)}{dv_3} = \frac{dv_3}{dv_3} = 1$;
3. $\overline{v_2} = \frac{df}{dv_2} = \overline{v_4} \frac{dv_4}{dv_2} = 1$;
4. $\overline{v_1} = \frac{df}{dv_1} = v_3 \frac{dv_3}{dv_1} = 1 \times v_{-1} = 2$;
5. $\overline{v_0} = 0$;
6. $\overline{v_{-1}} = \frac{df}{dv_{-1}} = \overline{v_2} \frac{dv_2}{dv_{-1}} + \overline{v_3} \frac{dv_3}{dv_{-1}} = 1 \times 1 + 1 \times v_1 = 1 + 1 = 2 = \frac{\partial f}{\partial x_2}$;
7. $\overline{v_{-2}} = \frac{df}{dv_{-2}} = \overline{v_1} \frac{dv_1}{dv_{-2}} + \overline{v_1} \frac{dv_1}{dv_{-2}} = 2 + 2 = \frac{\partial f}{\partial x_1}$;

Առաջադրանք: Տրված $f(x_1, x_2) = x_1 x_2 + x_2 - \ln(x_1)$ ֆունկցիայի համար կազմել նրա հաշվարկային գրաֆը և կատարել ուղիղ և հետընթաց փոխանցումով հաշվարկներ $x_1 = 2$, $x_2 = 4$ արժեքների համար: *Python*-ով կատարել ծրագրային իրականացում:

Այժմ դիտարկենք ընդհանուր տեսքի դիֆերենցիալ հավասարում՝ $Nu(x) = 0$, $x \in \Omega$, որտեղ N -ը որոշակի դիֆերենցիալ օպերատոր է: Ընդ որում Ω տիրույթի եզրում կան $u(x_i) = y_i$, $i = 1, 2, \dots, N_b$ պայմաններ: Ենթադրենք՝ լուծումը ուզում ենք փնտրել ֆունկցիաների որոշ դասից, որը ներկայացվում է պարամետրի տեսքով: Օրինակ, դա կարող է լինել ինչ-որ աստիճանի բազմանդամների դաս՝

$$A(x, w) = \{f(x, w) : f(x, w) = \sum_{i=0}^n w_n x^n, \quad w = (w_0, w_1, \dots, w_n)\}:$$

Կազմում ենք կորստի ֆունկցիա հետևյալ բանաձևով՝

$$Loss = Loss_{border} + Loss_{equation},$$

որտեղ

$$Loss_{border} = \frac{1}{N_b} \sum_{i=1}^{N_b} \|f(x_i, w) - y_i\|^2,$$

$$Loss_{equation} = \frac{1}{N_f} \sum_{i=1}^{N_f} \|Nf(t_i, w)\|^2:$$

Այստեղ $t_1, t_2, \dots, t_f$ կետերը պատկանում են տիրույթին: Կորստի այս ֆունկցիայի միմիակցական, կատարվում է նեյրոնային ցանցերի օգնությամբ, որտեղ որպես կշիռներ համարվում են w վեկտորի կոորդինատները (տե՛ս գլուխ 7):

Առաջադրանք: Դիցուք տրված է $y' + 3y = 0$, $y(0) = 1$, $x \in [0, 1]$ խնդիրը: Որպես մոդել ընտրել հինգերորդ կարգի բազմանդամները: Կատարել նեյրոնային ցանցի ուսուցում վերը նկարագրված եղանակով՝ որպես օպտիմիզատոր ընտրել գրադիենտային իջեցման Adam մեթոդը: Գնահատել մոդելի որակը միջին քառակուսային շեղման չափանիշով՝

$$\frac{1}{N_f + N_b} \sum_{i=1}^{N_b + N_f} (y_i - \hat{y}_i)^2:$$

Ածանցյալները հաշվել ավտոմատ դիֆերենցման եղանակով:

16.2 Ստոխաստիկ գրադիենտի փուլային ալգորիթմը (դիսպերսիայի նվազեցում)

Դիտարկենք օպտիմիզացիայի հետևյալ խնդիրը՝

$$f(w) = \frac{1}{l} \sum_{i=1}^l L_i(w) \quad w \in R^n:$$

Այս պարագրաֆում ցույց է տրվում, որ եթե l -ը բավականաչափ մեծ է, ապա ավելի էֆեկտիվ է կատարել ստոխաստիկ գրադիենտային իջեցում միևնույն ճշտությունը ապահովելու համար:

Այստեղ ենթադրվում է, որ $L_i, i = 1, 2, \dots, l$ ֆունկցիաները β -ոդորկ են և α -ուժեղ ուռուցիկ: Ենթադրվում է նաև, որ $\gamma \equiv \alpha/\beta < 1$: Կատարվում է գրադիենտային իջեցում հետևյալ ռեկուրենտ առնչությամբ՝

$$w_{t+1} = w_t - \frac{1}{\beta} f'(w_t):$$

Տեղի ունի գուգամիտության հետևյալ թեորեմը (տե՛ս [7]):

Թեորեմ 16.1. *Ճշմարիտ է հետևյալ գնահատականը՝*

$$f(w_{t+1}) - f(w_*) \leq (f(w_0) - f(w_*)) \exp(-\gamma t):$$

Ենթադրենք՝ յուրաքանչյուր իտերացիայում L_i ֆունկցիայի գրադիենտի հաշվման համար պահանջվում է $O(d)$ ժամանակ: Հետևաբար f -ի գրադիենտի հաշվարկման ժամանակամիջոցը կլինի $O(ld)$: Վերևի թեորեմից հետևում է, որ ϵ -ճշտություն ապահովելու համար պետք է կատարել առնվազն

$$O\left(\frac{1}{\gamma} \ln \frac{1}{\epsilon}\right)$$

իտերացիա: Այսպիսով ընդհանուր ժամանակամիջոցը հավասար կլինի՝

$$O\left(\frac{ld}{\gamma} \ln \frac{1}{\epsilon}\right):$$

Այժմ նկարագրենք Ստոխաստիկ գրադիենտի փուլային հետևյալ ալգորիթմը:

- $w_0^1 \in R^n$, $D_1 \geq \|w_0^1 - w_*\|$, $\eta_t = 1/\beta$.
- **for** $k=1$ **to** $K = \lceil \ln \frac{1}{\epsilon} \rceil$ **do**
- Let $B_{D_k}(w_0^k)$ be the ball of radius D_k around w_0^k .
- compute full gradient $f'(w_0^k)$:
- **for** $t = 1$ **to** $T = \frac{1}{\gamma^2}$ **do**
- Sample $i_t \in [l]$ uniformly at random, let $L_t = L_{i_t}$.

- construct stochastic gradient $g_t^k = L'_t(w_t^k) - L'_t(w_0^k) + f'(w_0^k)$:
- Let $v_t^{k+1} = w_t^k - \eta_t g_t^k$,
 $w_{t+1}^k = \Pi_{B_{D_k}(w_0^k)}(v_{t+1}^k)$.
- **end for**
- Set $w_0^{k+1} = 1/T(\sum_{t=1}^T w_t^k)$, $D_k/2 \rightarrow D_{k+1}$.
- **end for**
- **return** w_0^{K+1}

Տեղի ունի հետևյալ թեորեմը ([7]):

Թեորեմ 16.2. Ստոխաստիկ գրադիենտի փուլային ալգորիթը վերադարձնում է w_0^{K+1} վեկտորը

$$O((l + \frac{1}{\gamma^2})d \ln \frac{1}{\epsilon}),$$

ժամանակահատվածում:

Այդ վեկտորը դիտարկվող օպտիմիզացիոն խնդրի ϵ -ճշտությամբ մինիմումի կետ է:

Ակնհայտ է, որ մեծ l -ի դեպքում

$$(l + \frac{1}{\gamma^2})d \ln \frac{1}{\epsilon} < \frac{ld}{\gamma} \ln \frac{1}{\epsilon},$$

ինչը նշանակում է, որ նման խնդիրներում օգտակար է կատարել ստոխաստիկ գրադիենտի փուլային ալգորիթը:

16.3 Ստոխաստիկ գրադիենտի մեթոդը ուժեղ աճի պայմանով (strong growth condition)

Դիտարկենք հետևյալ խնդիրը՝

$$f(x) = \frac{1}{l} \sum_{i=1}^m l_i(w) \rightarrow \min_w:$$

Ենթադրվում է, որ բոլոր l_i , $i = 1, 2, \dots, m$ ֆունկցիաները β -ողորկ են և ներքևից սահմանափակ են: Դիտարկենք ստոխաստիկ գրադիենտի իջեցման հետևյալ եղանակը՝

$$w_{k+1} = w_k - \alpha_k f'_{\xi_k}(w_k),$$

որտեղ ξ_k պատահական մեծություն է, որը հավասար հավանականությամբ արժեքներ է ընդունում $\{1, 2, \dots, m\}$ ինդեքսների բազմությունից, իսկ

$$\alpha_k > 0, \sum_{i=1}^{\infty} \alpha_k = +\infty, \sum_{k=1}^{\infty} \alpha_k^2 < \infty:$$

Ենթադրենք նաև $E_{\xi_k}[|f'(w_{\xi_k})|^2] \leq \sigma^2 + \|f'(w_k)\|^2$ (strong growth condition):

Դիցուք նաև $K \in N$ և $k(K)$ այնպիսի պատահական մեծություն է, որ

$$P(k(K) = k) = \frac{\alpha_k}{\sum_{k=1}^K \alpha_k}:$$

Նշված պայմանների դեպքում տեղի ունի հետևյալ պնդումը:

Թեորեմ 16.3. $\forall \epsilon > 0 \ P\left(\|f'(w_{k(K)})\| \geq \epsilon\right) \rightarrow 0, \ K \rightarrow \infty:$

Ապացույց: Քանի որ f ֆունկցիան β -ռոդրկ է, ապա

$$f(w_{k+1}) - f(w_k) \leq (f'(w_k), w_{k+1} - w_k) + \frac{\beta}{2} \|w_{k+1} - w_k\|^2:$$

Այստեղից կստանանք

$$E[f(w_{k+1}) - f(w_k)] \leq -\frac{\alpha_k}{2} E[|f'(w_k)|^2] + \frac{\alpha_k^2 \beta}{2} \sigma^2, \text{ եթե } \alpha_k \leq \frac{1}{\beta}:$$

Գումարելով այս անհավասարությունը $k = 1, 2, \dots, K$ ինդեքսների համար՝ կստանանք

$$E[f(w_K) - f(w_1)] \leq -\frac{1}{2} \sum_{k=1}^K \alpha_k E[|f'(w_k)|^2] + \frac{\sigma^2 \beta}{2} \sum_{k=1}^K \alpha_k^2:$$

Հետևաբար, քանի որ f ֆունկցիան սահմանափակ է ներքևից, և $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$, ապա $\sum_{k=1}^K \alpha_k E[|f'(w_k)|^2]$ գումարը նույնպես կլինի սահմանափակ:

Մյուս կողմից, քանի որ $\sum_{k=1}^{\infty} \alpha_k = +\infty$, ապա

$$\lim_{K \rightarrow \infty} \frac{1}{\sum_{k=1}^K \alpha_k} \sum_{k=1}^K \alpha_k E[|f'(w_k)|^2] = 0:$$

Այստեղից՝ կիրառելով Մարկովի անհավասարությունը, կստանանք՝

$$P\left(\|f'(w_{k(K)})\|^2 \geq \epsilon^2\right) \leq \frac{1}{\epsilon^2} E_{\xi_k} \left[E_{k(K)}[|f'(w_{k(K)})|^2] \right] =$$

$$\frac{1}{\epsilon^2} E_{\xi_k} \left[\sum_{k=1}^K P(k(K) = k) \|f'(w_k)\|^2 \right] =$$

$$\frac{1}{\epsilon^2} E_{\xi_k} \left[\sum_{k=1}^K \frac{\alpha_k}{\sum_{k=1}^K \alpha_k} \|f'(w_k)\|^2 \right] \rightarrow 0, \text{ երբ } K \rightarrow \infty:$$

Թեորմն ապացուցվեց:

16.4 Օպտիմալ գրադիենտային մեթոդների մասին

Ինչպես արդեն գիտենք գրադիենտային մեթոդի զուգամիտության արագությունը կախված է քայլի ընտրության եղանակներից: Հարց է առաջանում, թե քայլի ընտրության որ եղանակի դեպքում կլինի զուգամիտության բարձր արագություն: Քննարկենք այս խնդիրը ստոխաստիկ գրադիենտի մեթոդում:

Դիտարկենք $f(x)$ ֆունկցիայի մինիմիզացիայի ստոխաստիկ գրադիենտի հետևյալ մեթոդը՝

$$x^{k+1} = x^k - \alpha_k g^k, \quad g^k = f'(x^k) + \xi^k,$$

որտեղ ξ^k -ն այնպիսի պատահական մեծություն է, որ $E[\xi^k] = 0$, $E[||\xi^k||^2] \geq \sigma^2$: Ճշմարիտ է հետևյալ պնդումը (տե՛ս թեորեմ 1 [40]):

Թեորեմ 16.4. Դիցուք f ֆունկցիայի գրադիենտը բավարարում է L լիպշիցի պայմանը L հաստատունով, և x^* -ը նրա մինիմումի կետն է: Այդ դեպքում տեղի ունի հետևյալ գնահատականը՝

$$E[||x^k - x^*||^2] \geq \left(\frac{1}{||x^0 - x^*||^2} + \frac{kL^2}{\sigma^2} \right)^{-1} :$$

Այս թեորեմից հետևում է, որ ոչ մի գրադիենտային մեթոդ (նշված պայմանների դեպքում) չի կարող ունենալ $O(1/k)$ կարգից ավելի բարձր զուգամիտության արագություն: Մյուս կողմից, արդեն ապացուցել ենք, որ եթե $\alpha_k = \alpha/k$, ապա

$$E[||x^k - x^*||^2] = O(1/k),$$

այսինքն՝ քայլի ընտրության նման եղանակը ասիմպտոտիկ օպտիմալ է:

Резюме
Методы оптимизации машинного обучения
Хачатрян Р. А.

В пособии изложены современные методы оптимизации и множество их приложений в задачах машинного обучения.

Сначала представлены основные концепции машинного обучения: коллекция объектов, модели, функции потерь и т. д.

Даны вероятностная и статистическая интерпретации задач машинного обучения. Введено понятие оптимальных предикторов. Обсуждаются алгоритмы их существования и построения. Изучаются условия сходимости, устойчивости и оптимальности этих алгоритмов. Приводятся различные оценки разницы между эмпирическими и реальными рисками, выраженные в фундаментальной теореме машинного обучения.

Описаны методы градиентного спуска и их модификации, особенно стохастические и ускоренные градиентные методы, которые имеют широкое применение в алгоритмах машинного обучения.

Пособие предназначено для студентов и магистров, изучающих информатику. Оно может быть полезно для исследователей, желающих углубить свои теоретические знания в области машинного обучения.

Предполагается, что читатель знаком с основами теории вероятностей, линейной алгебры, математического анализа, методов оптимизации и теории алгоритмов.

Abstract
Optimization Methods in Machine Learning
Khachatryan R.

The manual presents modern optimization methods and many of their applications in machine learning problems. First, the basic concepts of machine learning are presented: collection of objects, models, loss functions, etc. Probabilistic and statistical interpretations of machine learning problems are given. The concept of optimal predictors is introduced. Algorithms for their existence and construction are discussed. The conditions for convergence, stability, and optimality of these algorithms are studied. Various estimates of the difference between empirical and real risks, expressed in the fundamental theorem of machine learning, are given. Gradient descent methods and their modifications are described, especially stochastic and accelerated gradient methods, which are widely used in machine learning algorithms. The manual is intended for undergraduate and graduate students studying computer science. This may be useful for researchers wishing to deepen their theoretical knowledge in the field of machine learning. It is assumed that the reader is familiar with the basics of probability theory, linear algebra, mathematical analysis, optimization methods, and the theory of algorithms.

Գրականություն

- [1] Pau Baldillou Salse, An Introduction to Neural Ordinary Differential Equations, Barselona, 17 de gener de 2024.
- [2] Alexandre Defosses, Leon Bottou, Francis Banch, Nicolas Usunier, A Simple Convergence Proof of Adam and Adagrad, [https : //arxiv.org/abs/2003.02395](https://arxiv.org/abs/2003.02395)
- [3] Gerard Biau and Benoit Cadre, Optimization by gradient boosting, <https://arxiv.org/abs/1707.05023>, [Submitted on 17 Jul 2017].
- [4] Leon Bottou, Frank E.Curtis, Jorge Nuedal, Optimization Methods for Large-Scale Machin Learning, SIAM Review, Vol. 60, No 2, pp. 223-311.
- [5] Gabriel Peyre, Course Notes on Optimization for Machine Learning, www.numerical-tours.com, 2024.
- [6] E. Hazan, Introduction to Online Convex Optimization, Princeton Univercity, Boston -Delft, 2016.
- [7] E.Hazan, Lecture Notes. Optimization for Machine Learning, [https : //arxiv.org/abs/1909.03550](https://arxiv.org/abs/1909.03550), 2019.
- [8] Hastie T., Tibshirani R., Friedman J., The Elements of Statistical Learning, Data Mining, Inference, and prediction, Springer, Stanford, California, 2008.
- [9] Kearns M., Vazirani U., An Introduction to Computational Learning Theory, MIT Press, 1994.
- [10] Ester, Martin, Hans P Kriegel, Jorg Sander, and Xiaowei Xu. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, 226-31. doi:10.1.1.71.1980.
- [11] Krank M., Wolfe p., An algorithm for quadratic proframming, Naval Rs., Logist. Quart, 3, pp. 95-110, 1956.
- [12] Francisco S., Melo., Convergence of Q-Learninig: A simple proof, Institut of Systems and Robotics, Lisboa, Portugal, [http : //users.isr.ist.utl.pt/ mtjspaam/readingGroup/ProofQlearning.pdf](http://users.isr.ist.utl.pt/~mtjspaam/readingGroup/ProofQlearning.pdf)
- [13] Laurens van der Maaten, Geoffrey Hinton, Visualizing using t-SNE, Journal of Machine Learning Reasearch,9.2008, pp. 2579-2605.

- [14] L. Mclellenes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, Tutte Institute for Mathematics and Computing, Arxiv: 1802.03426v3[State ML] 18 Sep 2020.
- [15] Nesterov Y, Introductory lectures on convex optimization, Springer Science and Business Media, v.87(2003).
- [16] B. T. Polyak, Some methode of speeding up the convergence of iteration methods, Ussr computational mathematics and mathematical physics, v. 4(5), 1964, pp. 1-17.
- [17] Richard S. Sutton and Andrew G. Barto, Reinforcement Learninig: An Introduction, Second edition, in progres, A Bradford Book, The MIT Press Cambridge, Massachusetts London, England, 2015.
- [18] Rakesh Agrawal and Ramakrishnan Srikant, Fast algorithms for mining association rules, Preceedings of the 20th international Conference on Very Large Data Bases, VLDB, pages 487-499, Santiago, Chile, September 1994.
- [19] Suvrit Sra, Sebastian Nowozin, Stephen J. Wright, Optimization for Machine Learning, Cambridge, Massachusetts, London, England, 2012.
- [20] Shai Shalev-Shwartz and Shai Ben- David, Understanding Machine Learning, Cambridge University Press, 2014.
- [21] Schapire R. The boosting approach to maching learninig, An overview// MSRI Workshop on Nonlinear Estimation and Classification, Berkeley, CA.- 2001.
- [22] Tong Zhang, Mathematical Analysis of Machine Learning Algorithm, Cambridge University Press, July 2023.
- [23] Tengyu Ma and Andrew Ng, CS229 Lecture Nots, Parts 11 The EM algorithm, June 11, 2023.
- [24] Vladimir N. Vapnik, The Nature of Statistical Learninig Theory, Statistics for Engineering and information Science, Springer, 2000.
- [25] Н. В.Артамонов, Е. А., Ивин, А. Н. Курбатцкий, Д. Фантаццини, Введение в анализ временных рядов, Вологда, Вол НЦ РАН, 2021.
- [26] Андрей Б. Машинное обучение без лишних слов, « ПИТЕР», 2020.
- [27] Вьюгин В. В. Математические основы машинного обучения и прогнозирования. Москва, МЦНМО, 2018.
- [28] Дж. Вандер Плас. Python для сложных задач, наука о данных и машинное обучение, изд. Питер, 2018.
- [29] К. В. Воронцов, Лекции по методу опорных векторов, МГУ, 2007.

- [30] К. В. Воронцов, курс "Машинное обучение МГУ, 2014.
- [31] Володин И. Н. Лекции по теории вероятностей и математической статистике. Казань, 2006.
- [32] Б. А. Гладких, Методы оптимизации и исследование операций для бакалавров информатики, часть 2, изд. НТЛ. Томск, 2011.
- [33] Р. Дуда, П. Харт, Распознавание образов и анализ сцен, Мир, Москва, 1976.
- [34] Х. Ю. Золотых, Задачи по машинному обучению, Нижний Новгород, 2020, [http : //www.uic.unn.ru/ zny/ml](http://www.uic.unn.ru/zny/ml).
- [35] В. Г. Карманов, Математическое программирование, М. :, Физматлит, 2004.
- [36] А. Е. Лепский, А. Г. Броневич, Математические основы распознавания образов (Курс лекции), Изд-во ТТИ ЮФУ, Таганрог, 2009.
- [37] Лимановская О. В., Алферьева Т. И., Основы машинного обучения, Екатеринбург, Изд. Уральского университета, учебное пособие, 2020.
- [38] А. М. Миронов, Машинное обучение, часть 1, МГУ, 2018.
- [39] Нестеров Ю. М. Методы выпуклой оптимизации, МЦМНО, М.:, 2010.
- [40] Поляк Б. Т. Введение в оптимизацию, М.:, Наука, 1983
- [41] Поляк Б. Т. Метод Ньютона и ее роль в оптимизации и вычислительный математике, Труды ИСА РАН, 2006, т.28, с. 48-66.
- [42] Б.Н. Пшеничный, Выпуклый анализ и экстремальные задачи, М.:, Наука, 1980.
- [43] Б. Н. Пшеничный, Метод линеаризации, М.: Наука,1983.
- [44] Е. Соколов. Семинары по линейным классификатором. 2016. [https : //github.com/esokolov/ml – course – msu/blob/master/ML16/lecture – notes/Sem12_linear.pdf](https://github.com/esokolov/ml-course-msu/blob/master/ML16/lecture-notes/Sem12_linear.pdf).
- [45] Е. Соколов. Семинары по байесовским методом, 2016, декабря, [https : //github.com/esokolov/ml – course – msu/blob/master/ML16/lecture – notes/Sem13_bayes.pdf](https://github.com/esokolov/ml-course-msu/blob/master/ML16/lecture-notes/Sem13_bayes.pdf).
- [46] А. Г. Сухарев, А. Г. Тимохов, В.В. Федоров. Курс методов оптимизации, Наука, М., 1986.
- [47] С. Филипп, Решающие деревья. Учебник по машинному обучению, Яндекс, <https://education.yandex.ru/handbook/ml>

- [48] Шалев -Шварц Ш., Бен-Давид Ш., Идеи машинного обучения: от теории к алгоритмам, перевод с английского, ДМК Пресс, 2019.
- [49] Ա.Կ.Պետրոսյան, Հավանականությունների տեսություն և կիրառական վիճակագրություն, մաս երկու, ուսումնական ձեռնարկ, « Տնտեսագետ », Երևան, 2011.
- [50] Ռ. Ա. Խաչատրյան, Օպտիմիզացիայի մեթոդներ, ԵՊՀ հրատ., 2020.
- [51] Մ. Գ. Չիլինգարյան, Մեքենայական ուսուցման հիմունքներ (Մեքենայական ուսուցման ալգորիթմներ: Ռեգրեսիա), Երևան, Ճարտարագետ, 2023.

ԵՐԵՎԱՆԻ ՊԵՏԱԿԱՆ ՀԱՄԱԼՍԱՐԱՆ

ՌԱՖԻԿ ԱՂԱՍՈՒ ԽԱՉԱՏՐՅԱՆ

ՕՊՏԻՄԻԶԱՑԻԱՅԻ ՄԵԹՈԴՆԵՐԸ
ՄԵՔԵՆԱՅԱԿԱՆ
ՈՒՍՈՒՑՄԱՆ ԽՆԴԻՐՆԵՐՈՒՄ

ՈՒՍՈՒՄՆԱԿԱՆ ՁԵՌՆԱՐԿ

Համակարգչային ձևավորումը՝ Կ. Չալաբյանի
Կազմի ձևավորումը՝ Ա. Պատվականյանի
Հրատ. խմբագրումը՝ Լ. Ավետիսյանի

Հեղինակը հաստատում է, որ ծանոթ է «ԵՊՀ գրահրատարակչական քաղաքականությանը», և գրքում առկա փաստերը, դիրքորոշումները, կարծիքները շարադրված են հեղինակային իրավունքի և էթիկայի միջազգայնորեն ընդունված սկզբունքների պահպանմամբ:

Ստորագրված է տպագրության՝ 03.06.2025:
Չափսը՝ 60x84 ¹/₁₆: Տպ. մամուլը՝ 24.125:
Տպաքանակը՝ 100:

ԵՊՀ հրատարակչություն
ք. Երևան, 0025, Ալեք Մանուկյան 1
www.publishing.ysu.am