LOGISTIK REGRESSIYA ASOSIDA TASNIFLASH MASALALARINI YECHISH

Xamdamov Rustam Xamdamovich
Raqamli texnologiyalar va sun'iy intellektni rivojlantirish ilmiy-tadqiqot instituti
elshodhaydarov1881@gmail.com

Ibrohimov Aziz Ravshanbek o'g'li
"Kiberxavfsizlik markazi" DUK, Axborot tizimlari va rusurslari bo'limi
aribrohimov@gmail.com

Haydarov Elshod Dilshod o'g'li
Muhammad al-Xorazmiy nomidagi TATU
elshodhaydarov1881@gmail.com

ANNOTATSIYA

Hozirgi kunda ko'plab masalalarni yechish obyektlarni toifalarga tasniflash muammosiga chambarchas bog'liq. Toifalarga tasniflashning ko'plab usullari mavjud. Ularni axborot xavfsizligining turli muammolarini yechishda qo'llasa bo'ladi. Shu usullardan biri logistik regressiya bo'lib, ushbu maqolada logitik regressiya usuli yordamida obyektlarni sinflashtirish masalalari yechilgan, shuningdek optimallashtirish masalasining yechimi ham ishlab chiqilgan. Ishlab chiqilgan algoritmlar yordamida toifalashtirishning muammosining turli masalalarni hal etish mumkin.

Kalit so'zlar: logistik regressiya, chiziqli regressiya, optimallashtirish, spam, ham.

SOLVING CLASSIFICATION ISSUES BASED ON LOGISTIC REGRESSION

ABSTRACT

Currently, the solution of many issues is closely related to the problem of classifying objects into categories. There are many ways to classify into categories. They can be used to solve various problems of Information Security. One of these methods is logistic regression, in this article the issues of classification of objects using the logitic regression method were solved, and a solution to the optimization issue was also developed. With the help of developed algorithms, it is possible to solve various problems of the problem of categorization.

Keywords: logistic regression, linear regression, optimization, spam, ham.

KIRISH

Logistik regressiya regressiya emas, balki tasniflashni o‘rganish algoritmidir. Bu nom statistik ma’lumotlardan kelib chiqqan va logistik regressiyaning matematik formulasi chiziqli regressiyaga o‘xshashligi bilan bog‘liq.

Logistik regressiyaning mohiyatini misol tariqasida binar tasniflash yordamida tushuntiriladi. Biroq, u tabiiy ravishda ko‘p sinfli tasnifga kengaytirilishi mumkin.

Logistik regressiyaning maqsadi: y_i ni x_i ning chiziqli funksiyasi sifatida modellash, ammo y_i ning ikkilik qiymatlari uchun bu unchalik oson emas. $wx_i + b$ kabi xususiyatlarning chiziqli birikmasi minus cheksizlikdan plus cheksizlikka o‘tadigan funktsiyadir, y_i esa faqat ikkita mumkin bo‘lgan qiymatga ega.

ADABIYOTLAR TAHLILI VA METODOLOGIYA

Kompyuterlar bo‘lmagan va barcha hisob-kitoblarni qo‘lda bajarish kerak bo‘lgan bir paytda, olimlar chiziqli tasniflash modellarini afzal ko‘rdilar. O‘shanda ular agar manfiy belgini 0, musbat belgini 1 deb belgilasangiz, (0, 1) diapazonli oddiy uzluksiz funksiyani topish kifoya ekanligini payqashdi. Bunday holda, agar x namunasi uchun model tomonidan qaytarilgan qiymat 0 ga yaqinroq bo‘lsa, unga salbiy belgi beriladi; aks holda, namuna ijobiy deb belgilanadi. Ushbu xususiyatga ega bo‘lgan funksiyalardan biri standart logistik funktsiyadir (logistik sigmasimon deb ham ataladi):

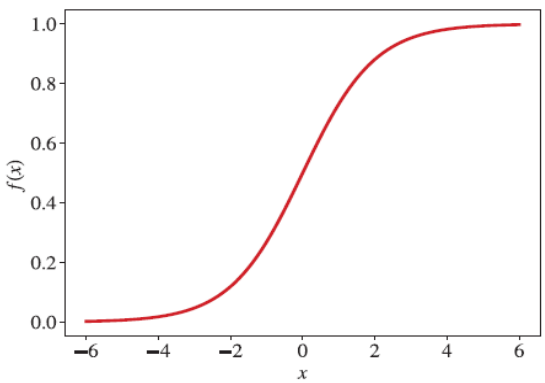
$$f(x) = \frac{1}{1+e^{-x}} \qquad (1)$$

bu yerda e - natural logarifmning asosi (eyler soni deb ham ataladi; e^{-x} qiymati dasturlash tillarida exp(x) funksiyasi sifatida ham tanilgan). Uning grafigi 1-rasmda ko‘rsatilgan.

Logistik regressiya modeli quyidagi ko‘rinishga ega:

$$f_{w,b}(x) \stackrel{\text{def}}{=} \frac{1}{1+e^{-(wx+b)}} \qquad (2)$$

Ko‘rib turganingizdek, chiziqli regressiyadan tanish wx + b atamasi bu yerda ishlatilgan.



1-rasm. Standart logistika funksiyasi

Standart logistik funksiyaning syujetini ko‘rib chiqsangiz, u bizning tasniflash maqsadimizga qanchalik mos kelishini ko‘rishingiz mumkin: agar biz w va b qiymatlarini mos ravishda optimallashtirsak, f (x) natijasini y_i ijobiy qiymatga ega bo‘lish ehtimoli sifatida talqin qilish mumkin. Masalan, agar u chegara qiymati 0,5 dan katta yoki unga teng bo‘lsa, biz x sinfini ijobiy, yani musbat deb aytamiz; aks holda bu salbiy, yani manfiy sinf bo‘ladi. Amalda, hal qilinayotgan muammoga qarab, boshqa chegara qiymatlari tanlanishi mumkin.

Endi w* va b* parametrlarning optimal baholarini qanday topish mumkinligi masalasini o‘rganamiz. Chiziqli regressiyada o‘rtacha kvadrat xato (mean squared error, MSE) deb ham ataladigan o‘rtacha kvadrat yo‘qotish funksiyasi sifatida aniqlangan empirik xavf minimallashtiriladi.

Logistik regressiyada, chiziqli regressiyadan farqli o‘laroq, o‘quv to‘plamining ehtimolliligi modelga muvofiq maksimal darajaga ko‘tariladi. Statistikada, ehtimollilik funksiyasi bizning modelimizga ko‘ra kuzatuv (namuna) qanchalik ehtimoli borligini aniqlaydi.

Misol uchun, o‘quv majmuamizda yorliqli namuna (y_i x_i) bor deylik. Bunda parametrlarimiz uchun ham ba’zi bir aniq qiymatlarni topdik (tanladik) deb faraz qilaylik. Endi biz x_i ga (2) tenglama yordamida modelni qo‘llasak, qandaydir qiymatga ega bo‘lamiz 0 < p < 1. Agar y_i musbat sinf bo‘lsa, bizning modelimiz bo‘yicha $y_i$musbat sinf bo‘lish ehtimoli p bilan beriladi. Xuddi shunday, agar y_i manfiy sinf bo‘lsa, uning manfiy sinf bo‘lish ehtimoli 1 - p bilan beriladi.

NATIJALAR

Logistik regressiyadagi optimallashtirish mezoni maksimal ehtimollilik deb ataladi. Chiziqli regressiyada bo‘lgani kabi o‘rtacha yo‘qotishni minimallashtirish o‘rniga, modelga muvofiq o‘quv ma’lumotlarining ehtimolini maksimal darajada oshiramiz:

$$L_{w,b} \overset{\text{def}}{=} \prod_{i=1.....N} f_{w,b}\,(x_i)^{y_i}(1 - f_{w,b}(x_i))^{1-y_i} \quad (3)$$

Bu erdagi $f_{w,b}(x_i)^{y_i}(1 - f_{w,b}(x_i))^{1-y_i}$ ifoda qo‘rqinchli tuyulishi mumkin, lekin bu shunchaki matematikadan "y_i=1 bo‘lganda $f_{w,b}(x_i)$ va aks holda $(1 - f_{w,b}(x_i))$" deb o‘zgartirsa bo‘ladi. Darhaqiqat, agar y_i = 1 bo‘lsa, $1 - f_{w,b}(x_i))$ 1 ga teng, chunki (1 - y_i) = 0 va bizga ma’lumki, 0 ning darajasiga teng bo‘lgan har qanday son 1 ga teng. Boshqa tomondan, y_i= 0 bo‘lsa, u holda xuddi shu sababga ko‘ra $f_{w,b}(x_i)^{y_i}$ 1 ga teng bo‘ladi.

Maqsad funksiyasida chiziqli regressiyada qo‘llanilgan yig‘indi operatori o‘rniga ko‘paytma operatoridan foydalanilgan. Buning sababi, N ta namunadagi N yorliqlarni kuzatish ehtimoli har bir kuzatuvning ehtimolliklarining mahsulotidir

(barcha kuzatuvlar bir-biridan mustaqil bo‘lsa, bizning holatimizda haqiqatan ham shunday). Ehtimollar nazariyasi bo‘yicha bir qator mustaqil eksperimentlarda natija ehtimolini ko‘paytirish bilan parallellik o‘tkazish mumkin.

Model *exp* funksiyasidan foydalanganligi sababli, amalda ehtimollikdan ko‘ra log-ehtimollikni maksimallashtirish qulayroqdir. Logarifm ehtimoli quyidagicha aniqlanadi:

$$LogL_{w,b} \overset{\text{def}}{=} \ln\left(L_{w,b}(x)\right) = \sum_{i=1}^{N}\left[y_i ln f_{w,b}(x) + (1-y_i)\ln(1-f_{w,b}(x))\right] \quad (4)$$

ln qat’iy ortib boruvchi funksiya bo‘lgani uchun uni maksimallashtirish uning argumentini maksimallashtirishga teng va bu yangi optimallashtirish masalasini hal qilish dastlabki masalani yechish bilan tengdir.

Chiziqli regressiyadan farqli o‘laroq, yuqoridagi optimallashtirish muammosi analitik yechimga ega emas. Shuning uchun bunday hollarda odatda sonli optimallashtirish masalasini echish uchun gradient tushish usulidan foydalaniladi.

Logistik regressiya asosida klassifikatsiyalash algoritmi. Logistik regressiya (LR) modeli:

$$f_{w,b}(x) = \frac{1}{1+e^{(wx-b)}} \quad (1)$$

uchun optimizatsiyalash, yani maksimallashtirish kriteriyasini quyidagi ko‘rinishda yozib olamiz:

$$q(w,b) = \sum_{i=1}^{N}[y_i \ln f_{wb}(x_i) + (1-y_i)\ln(1-f_{wb}(x_i))] \rightarrow \underset{w,b}{max} => (w*, b*) \quad (2)$$

Bu yerda

$$y_i = \begin{cases} 1, & i = 1,2,\dots,N_1 \\ 0, & i = N_1+1,\dots,N \end{cases} \quad (3)$$

(1),(3) shartlardan foydalanib, (2) optimallashtirish kriteriysini quyidagi ko‘rinishda yozib olish mumkin:

$$q(w) = \sum_{i=1}^{N} \ln\left(1 + e^{\sum_{j=0}^{n} w_j x_i^j}\right) - \sum_{i=N_1+1}^{N}\sum_{j=0}^{n} w_j x_i^j \quad (4)$$

Bu yerda

$$\sum_{j=0}^{n} w_j\, x_i^j = wx_i + b,\ i = 1,2,\dots.,N;\ x_i^0 = 1,\ w_0 = b.$$

Optimallashtirish masalasi

$$q(w) \rightarrow \underset{w}{max} => w^*, \quad (5)$$

$$f_w(x) = \frac{1}{1+e^{-\sum_{j=0}^{n} w_j x_i^j}} \geq \frac{1}{2},\ i = 1,2,\dots,N_1 \quad (6)$$

$$f_w(x) = \frac{1}{1+e^{-\sum_{j=0}^{n} w_j x_i^j}} < 1/2, \quad i = N_1+1,\dots,N \quad (7)$$

Bu yerda $w = (w_0, w_1, w_2, \dots., w_n)$.

Hosil bo‘lgan (5)-(7) optimallashtirish masalasini stoxastik tasodifiy qidiruv usullarining asoschisi L.A.Rastrigin tomonidan taklif etilgan moslashuvchan(adaptiv) tasodifiy qidiruv usulidan foydalanib yechamiz [1].

Moslashuvchan tasodifiy qidiruv usuli. Umumiy holda quyidagi optimallashtirish masalasi qo‘yilgan bo‘lsin

$$q(\omega) \to min => \omega^* \qquad (8)$$
$$\omega \in D$$

bu yerda $q(\omega)$ – umumiy holda nochiziqli ko‘p o‘zgaruvchili funksiya

$\omega = (\omega_1, \omega_2, \dots, \omega_n)$,

$\omega^* = (\omega_1^*, \omega_2^*, \dots, \omega_n^*)$ – (8) masalaning yechimi.

D – optimallashtirish masalasining aniqlanish sohasi, odatda tenglik va tengsizliklar ko‘rinishida berilishi mumkin.

Moslashuvchan tasodifiy qidiruv usulida quyidagi rekkurent formuladan foydalaniladi:

$$\omega^{k+1} = \omega^k + \Delta\omega^{k+1}, \qquad (9)$$

$$\Delta\omega^{k+1} = \begin{cases} a^{k+1}\Delta\omega^k, & agarda\ \ q(\omega^k) < q(\omega^{k-1}) \\ a^{k+1} \cdot \xi^{k+1}, & agarda\ \ q(\omega^k) \geq q(\omega^{k-1}) \end{cases} \qquad (10)$$

a^{k+1} – (k+1) qadamning uzunligini xarakterlovchi parametr. U optimallashtirish jarayonining qay darajada ketishiga qarab hozirgi holatga moslashtiriladi – agarda bundan oldingi qadam muvofaqqiyatli bo‘lsa, u holda bu parametr kattalashtiriladi, aks holda kichiklashtiriladi:

$$a^{k+1} = \begin{cases} \delta_1 a^k, agarda\ \ q(\omega^k) < q(\omega^{k-1}) \\ \delta_2 a^k, agarda\ \ q(\omega^k) \geq q(\omega x^{k-1}) \end{cases} \qquad (11)$$

Bu yerda $\delta_1 > 1$, $\delta_2 < 1$ koeffisentlar hisoblash jarayonining ijobiy natijadorligini taminlash maqsadidan kelib chiqib tanlanadi.

Bu parametrlarni to‘g‘ri tanlash masalaning ijobiy yechimini tanlashda alohida o‘rin tutadi.

ξ^{k+1} – (k+1) qadamdagi n-o‘lchovli birlik sferada barcha yo‘nalishlar bo‘yicha tekis taqsimlangan birlik tasodifiy vektordir.

U quyidagi algoritm yordamida hosil qilinadi:

$\xi = (\xi_1, \xi_2, \dots, \xi_n)$

$$\xi_i = \frac{\gamma_i}{\sqrt{\sum_{j=1}^{n} \gamma_j^2}}; \qquad i=1,2,\dots,n \qquad (12)$$

Bu yerda γ_i $(i=1,2,\dots,n)$ $\gamma_i \in [-1,1]$ intervalda tekis taqsimlangan tasodifiy sonlar ketma-ketligi.

U quyidagi

$$\theta: \quad \sqrt{2};\ \frac{\sqrt{2}}{2};\ \sqrt{3};\ \frac{\sqrt{3}}{2}; \qquad (13)$$

irratsional sonlar qatnashgan ifodalardan ixtiyoriy birortasidan foydalangan xolda quyidagi hisoblash formulasi asosida hosil qilinadi.

$$\eta_j =]j \times \theta[, \quad j = 1,2, \dots\dots., \dots. \qquad (14)$$

Bu yerda $]\cdot[- j \times \theta$ ifoda hisoblangandan so‘ng, uning kasr qismi olinishi funksiyasini bildiradi. Bu η_j sonlar [0,1] intervalda tekis taqsimlangan tasodifiy sonlar ketma-ketligi hisoblanadi.

$$\gamma_j = (c - d)\eta_j + d, \quad j=1,2,\dots\dots \qquad (15)$$

bu yerda d=*-1, c=1*.

Hosil bo‘lgan γ_j sonlar [-1,1] intervalda tekis taqsimlangan tasodifiy sonlar ketma-ketligi hisoblanadi.

Moslashuvchan (adaptiv) tasodifiy qidiruv usuli yordamida quyida keltirilgan 3ta test ko‘p o‘zgaruvchili optimallashtirish masalalarning yechimlarini topishda foydalanish mumkin.

1.”волна” funksiyasini hisoblash.

Masalaning berilishi:

$$f(x) = -2xe^{ix^2} - 2\sin(200x) \to max$$
$$x^* = 0 \quad f^* = -0.7$$
$$[-\alpha; \alpha]^| \ \alpha = 2$$

Masalaning yechimi:

-1.6285761637234164
-1.0108357120851046
-0.7
-1.0108357120851046
-1.6285761637234164

2. Ikki o‘lchamli funksiya. Levy funksiyasini hisoblash.

Masalaning berilishi:

$$f(x) = sin^2\left(\pi\left(1 + \frac{x_1 - 1}{4}\right)\right) + \sum_{i=2}^{9}\left(\frac{x_{i-1} - 1}{4}\right)^2 \left[1 + 10sin^2\left(\pi\left(1 + \frac{x_1 - 1}{4}\right)\right)\right]$$
$$+ \left(\frac{x_{10} - 1}{4}\right)^2 \to min$$
$$x^* = (0; 0; \dots.; 0) \quad f^* = 0$$
$$[-\alpha; \alpha]^x \ \alpha = 10$$

Masalaning yechimi:

[[50. 50.94739946 55.05140639 ... 55.05140639 50.94739946
50.]
[50.94739946 51.89479891 55.99880584 ... 55.99880584 51.89479891
50.94739946]
[55.05140639 55.99880584 60.10281278 ..0... 60.10281278 55.99880584
55.05140639]
[55.05140639 55.99880584 60.10281278 ... 60.10281278 55.99880584
55.05140639]
[50.94739946 51.89479891 55.99880584 ... 55.99880584 51.89479891
50.94739946]
[50. 50.94739946 55.05140639 ... 55.05140639 50.94739946
50.]]

3. Rastrigin funksiyasini hisoblash.

Masalaning berilishi:

$$f(x) = \sum_{i=1}^{|x|}\left(z_i^2 - 10\cos(2\pi z_i) + 10\right) + f^0 \to min$$

$$x^* = x_1^0; \dots\dots; x_{|x|}^0 \qquad f^* = f^0 = -330$$

$$[-\alpha; \alpha]^{|x|} \quad \alpha = 5$$

-300
-285
-242.0
-200.0
-162.0
-128.0
-98.0
-72.0
-50.0
-32.0
-18.0
-8.0
-2.0
1.3497838043956716e-31
2.0
8.0
18.0
32.0

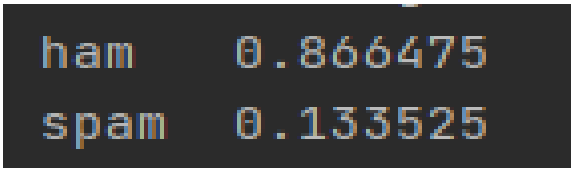
50.0
72.0
98.0
128.0
162.0

Endi **Logistik regressiya asosida klassifikatsiyalash algoritmi** asosida hosil bo‘lgan ko‘p o‘zgaruvchili optimallashtirish masalasining yechimlarini topishda **moslashuvchan (adaptiv) tasodifiy qidiruv usuli yordamida** yaratilgan dastur orqali elektron pochtadagi xabarlarni toifalarga ajratish masalasining yechimini ko‘ramiz.

1. Jami 5575 ta spam va ham xabarlar olingan birinchi navbata 3 ustunga bo‘lib olingan ichidan 10 xabar tanlab olinadi.

ID	Turi	Xabar
1901	spam	Sorry, I’ll call later
5528	ham	Its just the effect of irritation. Just ignore it
3581	spam	You are right. Meanwhile how’s project twins c...
3960	spam	Your dad is back in ph?
403	ham	None of that’s happening til you get here though
2196	spam	Not much, just some textin’. How bout you? Website https://mov.com
917	ham	Not much, just some textin’. How bout you?
2825	ham	When people see my msgs, They think Iam addict...
2662	spam	Hello darling how are you today? I would love ...
1274	ham	Let me know how to contact you. I’ve you settl...

2. Shundan so‘ng dasturning o‘zi bu xabarlarni ikkita guruhga ajratib oladi yani spam va ham gurihiga va ularni fozi nisbatini chiqarib beradi bizning dasturimizda u quyidagi nisbatda bo‘ladi



3. Shundan so‘ng sapam va ham habarlani guruhlarga ajratilgandan so‘ng ularni 8 parametrlar bo‘yicha tekshirib chiqamiz

	00	000	03	...	½s	½t	½ï
sms				...			
Sorry, I'll call later	0	0	0	...	0	0	0
Its just the effect of irritation. Just ignore it	0	0	0	...	0	0	0
You are right. Meanwhile how's project twins co...	0	0	0	...	0	0	0
Your dad is back in ph?	0	0	0	...	0	0	0
None of that's happening til you get here though	0	0	0	...	0	0	0
...	..	...	..	...	..	..	..
Not much, just some textin'. How bout you?	0	0	0	...	0	0	0
When people see my msgs, They think Iam addicte...	0	0	0	...	0	0	0
Ok lor...	0	0	0	...	0	0	0
Hello darling how are you today? I would love t...	0	0	0	...	0	0	0
Let me know how to contact you. I've you settle...	0	0	0	...	0	0	0

4. Natija quyidagicha chiqqanligini ko‘rishimiz mumkin.

```
Jami tekshirilgan qadamlar: 1411
SPAM va HAM habarlar yozuvlari: english
Aniqlik darajasi: 98.9
Jami SPAM va HAM habarlar soni:5575
```

Masalani nazorat tanlanmasi k=5575ta obyektlar uchun hisoblash tajribasini o‘tqazamiz. Bunda nazorat tanlanmasidagi obyektlar(xabarlar)ni ikkita sinf(ham yoki spam)ga toifalash maqsadida quyidagi produksion model yordamida ifodalangan qoidadan foydalanamiz:

$$x_i = \begin{cases} \text{spam}, & agar \quad \sum_{j=1}^{6} x_i^j \geq b_q \; yoki \; (x_i^7 > 0.06 \; yoki \; x_i^8 > 0.03) \\ \text{ham}, & boshqa \; hollarda \end{cases}$$

$i=1,2,\ldots,k$. b_q - belgilar yig‘indisi uchun bo‘sag‘aviy parametr, ko‘rilayotgan masala uchun xususiy xolda $b_q = 2$ olish tavsiya etiladi. Ushbu $k=5575$ nazorat tanlanmasi uchun to‘g‘ri toifalash ko‘rsatkichlari “spam” sinfi obektlari uchun 4828ta obektdan 4791tani, ya’ni 99,23%ni, “ham” sinfi obektlari uchun 747ta obektdan 738tani, ya’ni 98,8%ni tashkil etdi.

Ikkinchi turdagi xatolik bo‘yicha 10 ta “spam” xabarni 0.2% va 5 ta “ham” xabarni 0.66% noaniqlik bilan topdi. Umumiy natija alohida “spam” xabarlar uchun 99% va ham xabarlar uchun 97.99%ni hamda birgalikdagi “spam” va “ham” xabarlar uchun 98.9%ni tashkil etdi.

1-jadval.
Testlash natijalari.

obekt	Nazorat tanlanmasi (summ - sinov xabarlar soni)	Birinchi turdagi xatolik (FRR - False Rejection Rate) - noto‘g‘ri rad etishlar soni (% yolg‘on rad etish)	To‘g‘ri toifalash natijasi, % da	Ikkinchi turdagi xatolik (FAR - False Acceptance Rate) - noto‘g‘ri aniqlashlar soni (noto‘g‘ri aniqlash%)	Identifikatsiya (umumiy) natijasi % da
spam	4828	37 (0,8%)	99,23%	10 (0.2%)	99%
ham	747	9 (1,2%)	98,8%	5 (0.66%)	97,99%
Jami (e mail = spam + ham)	5575	46 (0.83%)	99.17%	15 (0,6%)	98,9%

XULOSA

Yuqoridagi masalalardan shuni ko‘rishimiz mumkinki logistik regressiya toifalashtirish usullari orasida eng yaxshilaridan biri hisoblanadi. Bu usulni takomillashtirish orqali axborot xavfsizligini turli masalalarini yechishda yuqori samaradorlikka erishish mumkin.

ADABIYOTLAR RO‘YXATI

1. *Л.А.Растригин.* Современние принципи управления сложними объектами. *М.*: Советское радио. 1980. - 230 с.
2. Abdulhamit Subasi, in Practical Machine Learning for Data Analysis Using Python, 2020
3. *А.П. Карпенко.*Современние алгоритми поисковий оптимизасии. М.:Издателство МГТУ им. Н.Э.Баумана, 2014. - 446 с.
4. Thomas W. Edgar, David O. Manz, in Research Methods for Cyber Security, 2017
5. *Бурков Андрей*, Машинное обучение без лишних слов. — СПб.: Питер, 2020. — 192 с.
6. Khamdamov R.Kh., Khaydarov E.D. “Pre-processing of primary spam classification data from email messages” Современное состояние и перспективи развития сифровых технологий и искусственного интеллекта Сборник докладов республиканской научно-технической конференции Самарканд, 26-27 октябрья 2022 г.