Izveštaj o drugom domaćem zadatku iz predmeta Mašinsko učenje

Viktor Todosijević 3140/2021

Novembar 2022

1 Uvod

Tema drugog domaćeg zadatka glasila je "generalizovani linearni modeli i generativni algoritmi". Kako je broj indeksa studenta 3140, algoritmi čiji će rad i implementacija biti opisani su *logistička regresija* i *gausovska diskriminantna analiza*. U odeljku 3 će biti reči o implementaciji i rezultatima *logističke regresije* na datom skupu podataka, dok će u odeljku 4 biti reči o implementaciji i rezultatima *gausovske diskriminantne analize* na istom skupu podataka.

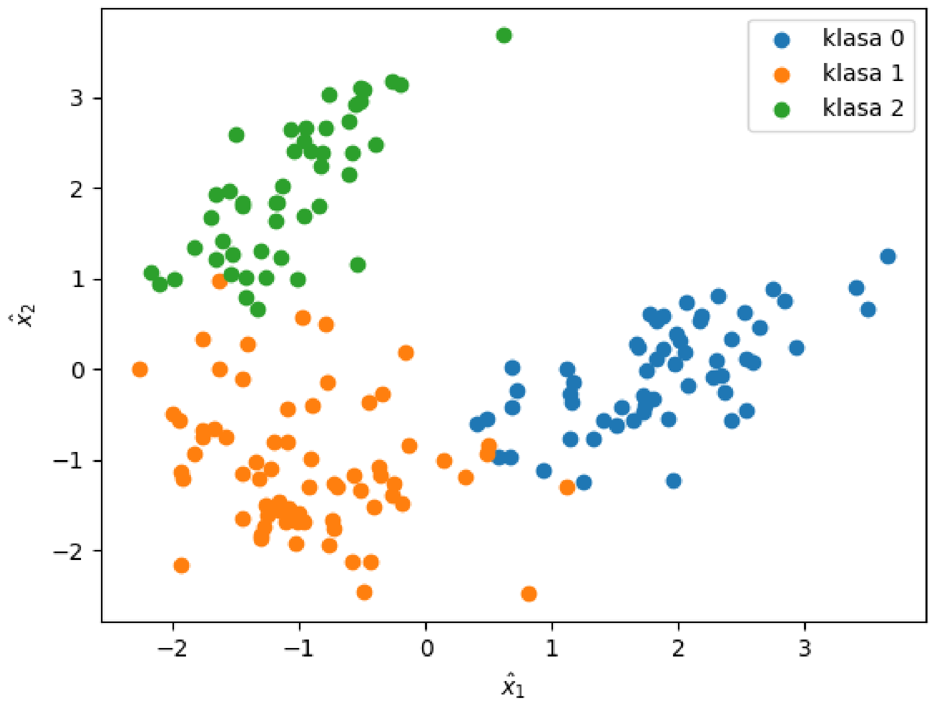
2 Transformacija podataka

Uvidom u podatke možemo videti da su prosečne vrednosti različitih atributa na različitim redovima veličine. Ovakvi podaci mogu otežati konvergenciju nekih algoritama na bazi gradijentnog spusta. Iz tog razloga je pogodno normalizovati odbirke oduzimanjem srednje vrednosti i deljenjem sa standardnom devijacijom kao što je opisano sledećom jednačinom:

$$\boldsymbol{x}_{norm} = \frac{\boldsymbol{x} - \boldsymbol{\mu}}{\boldsymbol{\sigma}} \tag{1}$$

gde je $\boldsymbol{x}$ vektor obeležaja odbirka, $\boldsymbol{\mu}$ vektor srednjih vrednosti obeležaja, a $\boldsymbol{\sigma}$ vektor standardnih devijacija obeležaja. Parametri $\boldsymbol{\mu}$ i $\boldsymbol{\sigma}$ se procenjuju iz obučavajućeg skupa.

Ako primetimo(slika 1) da se nakon normalizacije podataka i primene Karhunen-Loeve ekspanzije [1] za slučajne vektore problem može svesti na dve dimenzije, zadržavajući separabilnost klasa, onda zbog bolje vizualizacije možemo raditi sa takvim transformisanim podacima.



Slika 1: Odbirci nakon normalizacije i smanjenja dimenzije.

3 Logistička regresija

Algoritam logističke regresije vrši binarnu klasifikaciju odbiraka na sledeći način:

$$\hat{y} = \begin{cases} 1, & \sigma(\boldsymbol{\theta}^T \boldsymbol{x}) \geq 0.5 \\ 0, & \sigma(\boldsymbol{\theta}^T \boldsymbol{x}) < 0.5 \end{cases} \tag{2}$$

gde je $\hat{y}$ klasa dodeljena odbirku $\boldsymbol{x}$, $\boldsymbol{\theta}$ vektor koeficijenata klasifikatora, a σ takozvana sigmoidna funkcija definisana kao:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{3}$$

i ima osobinu da je rezultat $\sigma(x) \in [0, 1]$ što mu daje interpretaciju verovatnoće, tako da se $\sigma(\boldsymbol{\theta}^T \boldsymbol{x})$ može tumačiti kao verovatnoća pripadanja odbirka $\boldsymbol{x}$ jednoj klasi, a $1 - \sigma(\boldsymbol{\theta}^T \boldsymbol{x})$ drugoj klasi. Za dobijanje koeficijenata ne postoji analitičko rešenje, već se mora pribeći procesu numeričke optimizacije metodom gradijentnog uspona. Uzimajući gradijent takozvane log-verodostojnosti dolazimo do sledećeg algoritma za ažuriranje koeficijenata $\boldsymbol{\theta}$:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \alpha(y^{(i)} - \sigma(\boldsymbol{\theta}^T \boldsymbol{x}^{(i)}))\boldsymbol{x}^{(i)} \tag{4}$$

gde se α naziva stopom učenja, i je redni broj odbirka, a y tačna vrednost ciljne promenljive. Ovo pravilo ažuriranja ponavljamo za svaki odbirak u obučavajućem skupu(mešajući njihov redosled) dok ne dodje do konvergencije. U praksi koeficijente ne ažuriramo za svaki odbirak, već propagiramo m_{mb} odbiraka i usrednjimo vrednost gradijenta izmedju njih. Veličina m_{mb} se naziva veličinom mini-šarže. Na ovaj način dolazimo do ubrzanja na računaru jer smo vektorizovali izraz 4, a takodje i eliminišemo odredjenu dozu šuma usrednjavanjem više odbiraka.

3.1 Metodologija

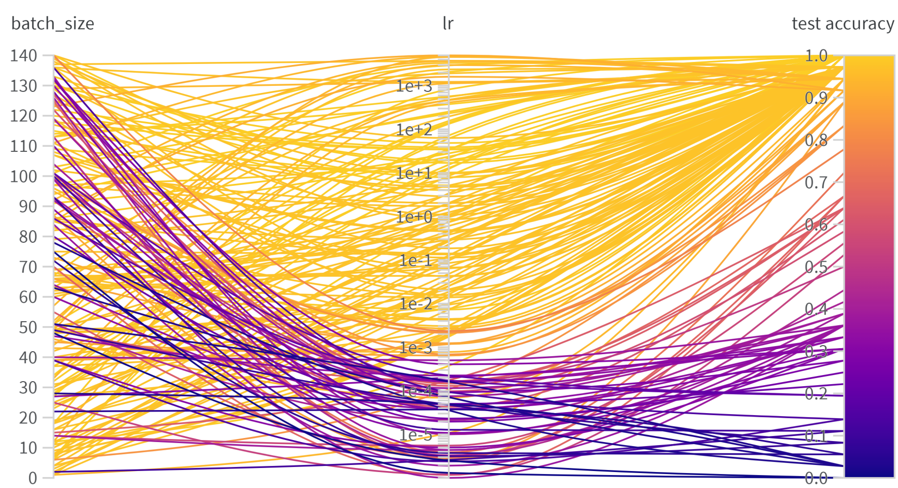
Kako je u domaćem zadatku dat skup podataka čiji odbirci pripadaju jednoj od tri klase, algoritam logističke regresije mora biti proširen i na slučaj više klasa. Ovo radimo na sledeći način: za svaku od klasa konstruišemo novi obučavajuči skup gde obeležjima iz ciljane klase dodelimo oznaku 1, a odbircima iz ostalih klasa oznaku 0. Na ovakve obučavajuće skupove možemo primeniti logističku regresiju i dobiti koeficijente klasifikatora $\boldsymbol{\theta}_i, \forall_{i \in [1, N_c]}$ gde je N_c ukupan broj klasa. Konačno, klasifikaciju vršimo tražeći najveću vrednost izlaza svih klasifikatora:

$$\hat{y} = \arg\max_{i} \sigma(\boldsymbol{\theta_i}^T \boldsymbol{x}) \tag{5}$$

Operator $\sigma(\cdot)$ je monotona funkcija, stoga ne utiče na maksimum pa se može i izostaviti. Skup podataka je podeljen na obučavajući i testirajući u odnosu 80:20.

3.2 Rezultati

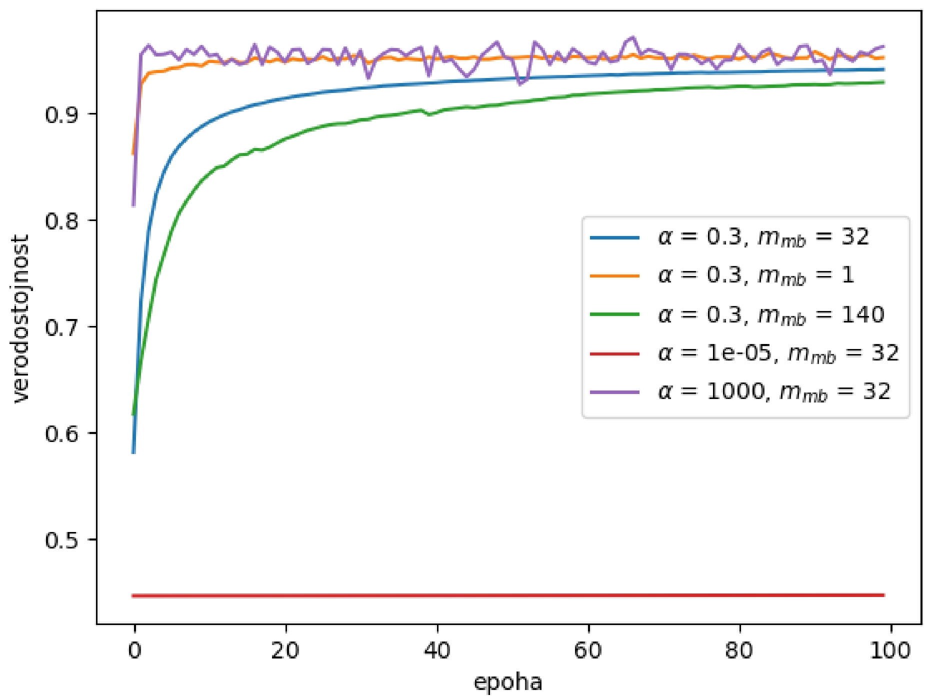
Slučajnom pretragom po prostoru hiperparametara(slika 2) možemo zaključiti da je za stopu učenja α loše uzeti vrednost manju od 0.001, dok je problem isuviše jednostavan da bi došlo do problema pri prevelikoj stopi učenja. Isto važi i za prevelike i premale vrednosti veličine mini-šarže m_{mb}. Za svrhu dalje analize, kao dobru vrednost stope učenja α^* uzimamo 0.3, a kao premalu odnosno preveliku uzimamo 10^{-5} i 1000, respektivno. Kao dobru vrednost veličine mini-šarže m^*_{mb} uzimamo 32, a kao premalu i preveliku 1 i 140, respektivno.



Slika 2: Pretraga po prostoru hiperpatametara α i m_{mb} gde je ciljna promenljiva tačnost klasifikacije na testirajućem skupu(desno). Žutom bojom su obeležene najbolje kombinacije (α, m_{mb}), a ljubičastom najgore.

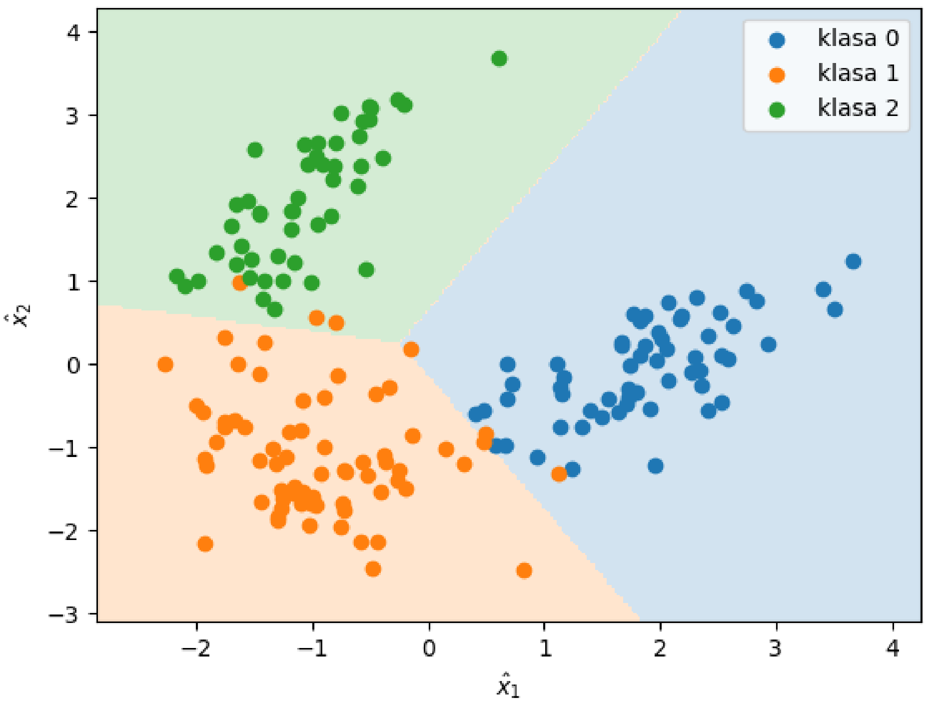
$$p(y|\boldsymbol{x}; \boldsymbol{\theta}) = (\sigma(\boldsymbol{\theta}^T \boldsymbol{x}))^y (1 - \sigma(\boldsymbol{\theta}^T \boldsymbol{x}))^{1-y} \tag{6}$$

Na slici 3 se može videti verodostojnost(data izrazom 6) usrednjena izmedju tri klasifikatora u funkciji *epohe* tj. svakog prolaska kroz obučavajući skup. Kao što smo ranije videli, proces obučavanja se uspešno završava za sve kombinacije parametara, osim one sa premalom vrednošću stope učenja. Za preveliku stopu učenja vidimo da verodostojnost ”skače” oko optimuma, jer se u prostoru parametra ne možemo pozicionirati dovoljno precizno, stoga skačemo oko optimuma. Za različite vrednosti veličine mini-šarže vidimo se menja brzina konvergencije. Naime za manje vrednosti m_{mb} je konvergencija brža jer vršimo više ažuriranja koeficijenata za isti broj epoha. Iznenadjujeće je da i za $m_{mb} = 1$ algoritam funkcioniše dobro. Ovo govori jednostavnosti problema. Treba napomenuti da za male vrednosti veličine mini-šarže obučavanje modela traje duže jer ne koristimo računarske mogućnosti vektorizacije i paralelizacije obračuna parametara.



Slika 3: Srednja verodostojnost sva tri klasifikatora za različite kombinacije hiperparametara u funkciji epohe.

Postignuta tačnost na testirajućem skupu je 97.22 %. Na slici 4 se vidi na koji način klasifikatori dele prostor; naime, diskriminacione krive su poluprave.



Slika 4: Oblasti klasifikacije i odbirci.

4 Gausovska diskriminantna analiza

Gausovska diskriminantna analiza je generativni algoritam za klasifikaciju[2], što znači da, umesto direktnog učenja uslovne raspodele pripadnosti klasi na osnovu odbiraka $p(y|\boldsymbol{x})$, estimiramo apriori verovatnoću pojavljivanja odbirka iz odredjene klase $p(y)$ i uslovnu raspodelu odbiraka na osnovu pripadnosti klasi $p(\boldsymbol{x}|y)$. Uz pomoć Bajesove teoreme, imajući u vidu da $p(\boldsymbol{x})$ ne utiče na maksimum, klasifikaciju vršimo na sledeći način:

$$\hat{y} = \arg\max_{y_i} p(y_i|\boldsymbol{x}) = \arg\max_{y_i} \frac{p(\boldsymbol{x}|y_i)p(y_i)}{p(x)} = \arg\max_{y_i} p(\boldsymbol{x}|y_i)p(y_i) \tag{7}$$

Metodom gausovske diskriminantne analize usvajamo $y_i \sim Bernoulli(\phi_i)$ i $\boldsymbol{x}|y_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$.

4.1 Metodologija

Parametre raspodela estimiramo na sledeći način:

$$\phi_i = p(y_i) = \frac{n_i}{N} \tag{8}$$

$$\boldsymbol{\mu}_i = \frac{1}{n_i} \sum_{j=0}^{n_i-1} \boldsymbol{x}_i^{(j)} \tag{9}$$

$$\boldsymbol{\Sigma}_i = \frac{1}{n_i - 1} \sum_{j=0}^{n_i - 1} (\boldsymbol{x}_i^{(j)} - \boldsymbol{\mu}_i)(\boldsymbol{x}_i^{(j)} - \boldsymbol{\mu}_i)^T \tag{10}$$

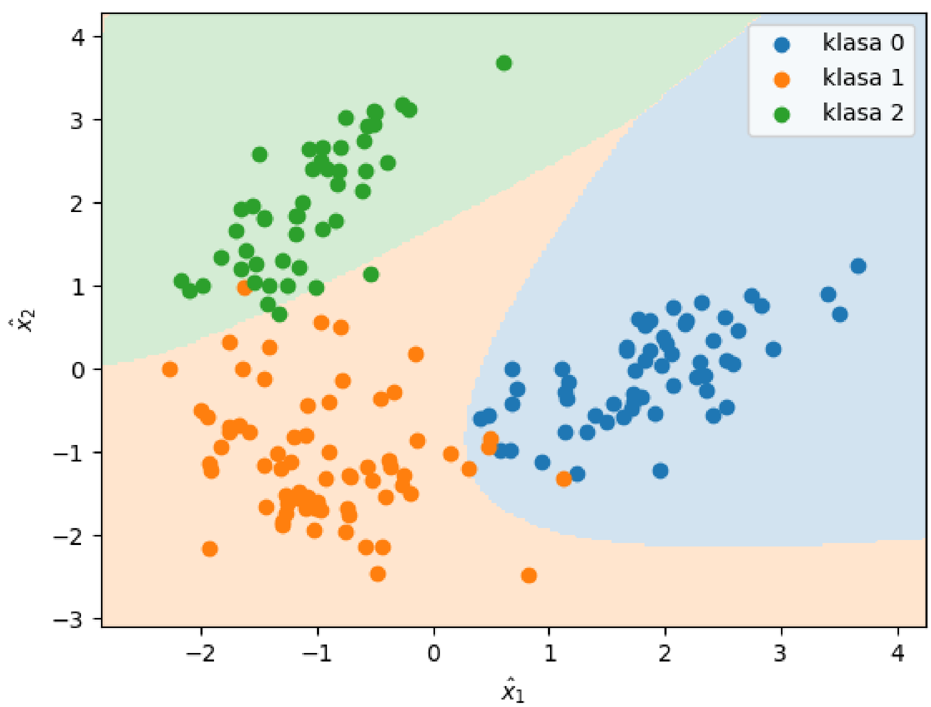
gde je N ukupan broj odbiraka u obučavajućem skupu, n_i broj odbiraka iz obučavajućeg skupa koji pripadaju klasi y_i, a $\boldsymbol{x}_i$ odbirci iz obučavajućeg skupa koji pripadaju klasi y_i. Odbirke klasifikujemo uvrštavanjem izraza

$$p(\boldsymbol{x}|y_i) = \frac{1}{(2\pi)^{\frac{dim(\boldsymbol{x})}{2}} |\boldsymbol{\Sigma}_i|^{\frac{1}{2}}} \, e^{\dfrac{(\boldsymbol{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i)}{2}} \tag{11}$$

i $p(y_i)$ u izraz 7.

4.2 Rezultati

Postignuta tačnost na testirajućem skupu je 100 %. Na slici 5 se vidi na koji način klasifikatori dele prostor; naime, diskriminacione krive su krive drugog reda.



Slika 5: Oblasti klasifikacije i odbirci.

Reference

[1] Željko Djurović. "Predavanje 9. Redukcija dimenzija oblika u statističkom prepoznavanju oblika". Predavanja iz predmeta *Prepoznavanje oblika (13E054PO)* 2022. https://automatika.etf.bg.ac.rs/images/FAJLOVI_srpski/predmeti/izborni_kursevi_os/obrada_signala/OS4PO/materijali/Predavanje%206.pdf

[2] Predrag Tadić. "Generativni algoritmi". Predavanja iz predmeta *Mašinsko učenje (13M051MU)* 2022. https://automatika.etf.bg.ac.rs/images/FAJLOVI_srpski/predmeti/master_studije/MU/04%20Generativni%20algoritmi.pdf