Izveštaj o prvom domaćem zadatku iz predmeta Mašinsko učenje

Viktor Todosijević 3140/2021

Oktobar 2022

1 Sažetak izveštaja

U ovom izveštaju je opisana izrada prvog domaćeg zadatka iz predmeta Mašinsko učenje. Student je u programskom jeziku *Python* implementirao algoritam lokalno ponderisane linearne regresije na brz i efikasan način. Student je izložio način funkcionisanja ovog algoritma i primenio isti na skupu podataka priloženom uz zadatak. Uz primenu i rezultate na datom skupu podataka prikazan je i odabir hiperparametra.

2 Uvod

Tema prvog domaćeg zadatka iz predmeta Mašinsko učenje je regresija. Kako je broj indeksa studenta 3140, a

$$(3 + 1 + 4 + 0)\ mod\ 3 = 2$$

algoritam domaćeg zadatka je *lokalno ponderisana linearna regresija*. U odeljku *Metodologija* je opisan izabrani algoritam kao i njegova implementacija, dok je u odeljku *Rezultati* opisan rezultat primene pomenutog algoritma na priloženi skup podataka.

3 Metodologija

Lokalno ponderisana linearna regresija je algoritam mašinskog učenja koji za dati odbirak, na osnovu odbiraka iz skupa podataka bliskih datom, modeluje (lokalno) linearnu zavisnost izmedju tog odbirka i ciljne promenljive. Greška J koja se minimizuje data je sledećom formulom [1]:

$$J(\boldsymbol{\theta}) = \sum_{i=0}^{N-1} \omega^{(i)} (y^{(i)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(i)}) \tag{1}$$

$$\omega^{(i)} = e^{-\frac{||\boldsymbol{x}^{(i)} - \boldsymbol{x}||_2^2}{2\tau^2}} \tag{2}$$

gde je N broj odbiraka u obučavajućem skupu, $\boldsymbol{x}^{(i)}$ vektor obeležaja i-tog odbirka, $\omega^{(i)}$ težinski faktor i-tog odbirka, y ciljna promenljiva, $\boldsymbol{\theta}$ vektor koeficijenata linearne regresije, a τ hiperparametar. Optimizacijom ovog kriterijuma, odnosno primenom parcijalnog izvoda po $\boldsymbol{\theta}$ i izjednačavanjem sa nulom dobijamo sledeći izraz za vektor koeficijenata linearne regresije:

$$\hat{\boldsymbol{\theta}} = (X^T W X)^{-1} X^T W \boldsymbol{y} \tag{3}$$

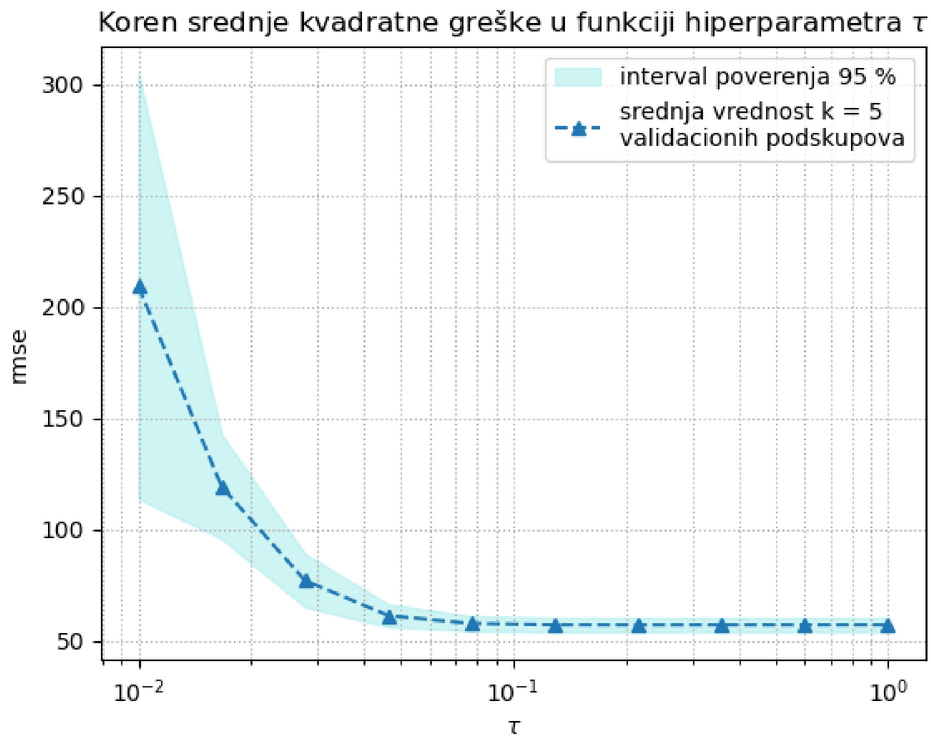
gde je X matrica obeležaja odbiraka, $\boldsymbol{y}$ vektor ciljnih promenljivih, a $W = diag(\omega^{(0)}, \omega^{(1)} \ldots \omega^{(N-1)})$ dijagonalna matrica težina iz jednačine 2.

Ono što lokalno ponderisanu linearni regresiju izdvaja od grebene ili LASSO regresije je to što se za svaki odbirak koeficijenti linearne regresije $\hat{\boldsymbol{\theta}}$ računaju iznova. Ovo znači da, pored toga što metoda zahteva da se skup podataka čuva, a ne odbaci nakon obučavanja kao kod drugih metoda, testiranje zahteva da se kroz svaki odbirak test skupa prodje u nekoj programskoj petlji. U programskom jeziku *Python* ovo može biti prilično vremenski skupo. Iz tog razloga se u priloženoj implementaciji koristi pristup koji korišćenje petlji prebacuje na ugradjene funkcije iz programskog paketa *numpy* koji će taj računski posao delegirati programskom jeziku *C*. Naime, korišćenjem koncepta Ajnštajnove notacije [2] možemo izračunati dijagonalne matrice težina W za svaki odbirak u test skupu veličine M i posložiti ih u trodimenzionalnu matricu W^* dimenzija $M \times N \times N$. Daljim korišćenjem koncepta Ajnštajnove notacije i jednačine 3 sa matricom W^* umesto W dobijamo vektor koeficijenata linearne regresije za svaki odbirak u test skupu bez korišćenja ijedne petlje u *Python*-u.

Metoda lokalno ponderisane linearne regresije sadrži i jedan hiperparametar τ čija veća ili manja vrednost odredjuje širinu Gausovog zvona iz jednačine 2, a samim tim i koliko lokalno susedstvo treba uzeti u obzir pri odredjivanju vrednosti ciljne promenljive. Razumno je da za funkcije ciljne promenljive koje imaju komponente na visokim učestanostima treba izabrati manju vrednost hiperarametra τ, dok suprotno važi za one koje nemaju takve komponente. Njegova odgovarajuća vrednost za dati skup podataka odredjena je pretragom po domenu $\tau \in [10^{-2}, 10^{0}]$ gde je n tačaka raspodeljeno na taj način da su one na logaritamskoj skali parametra τ jednako raspodeljene tj. rastojanje izmedju svake susedne tačke raste eksponencijalno. Za svaku vrednost hiperparametra uradjena je unakrsna validacija u $k = 5$ delova.

4 Rezultati

Na slici 1 se vidi da koren srednje kvadratne greške konvergira ka vrednosti od oko 57.1 za $\tau > 0.07$. Kako se Gausova funkcija sa većim vrednostima τ sve više i više širi tako sve više i više tačaka u skupu podataka ulazi u račun koeficijenata linearne regresije. To što za veće vrednosti τ greška ostaje konstantna sugeriše da je ciljna funkcija linearna funkcija odbiraka i lokalno ponderisana linearna regresija se za te vrednosti pretvara u grebenu regresiju jer matrica W teži jediničnoj matrici .



Slika 1: Koren srednje kvadratne greške u funkciji hiperparametra τ sa intervalom poverenja 95%.

Reference

[1] Predrag Tadić. ”Linearna regresija”. Predavanja iz predmeta *Mašinsko učenje (13M051MU)* 2022. https://automatika.etf.bg.ac.rs/images/FAJLOVI_srpski/predmeti/master_studije/MU/01%20Linearna%20regresija.pdf

[2] https://numpy.org/doc/stable/reference/generated/numpy.einsum.html