Prvi domaći zadatak iz Tehnika i metoda analize podataka

Danilo Veljović, broj indeksa 1120

Opis skupa podataka

Skup podataka koji je korišćen za testiranje implementiranog modela Linearne regresije je skup podataka u kojima se nalaze informacije o procesorima (http://archive.ics.uci.edu/ml/datasets/Computer+Hardware). Zadatak modela linearne regresije je da na osnovu parametara svakog procesora proba da predvidi performanse/kvalitet procesora koji je dat ERP vrednošću u skupu podataka. ERP vrednost je kontinualna vrednost i zbog toga je pogodna za predikciju pomoću linearne regresije.

Atributi u skupu podataka su ime proizvođača procesora, model procesora, MYCT – vremen ciklusa u nanosekundama, MMIN – minimum operativne memorije u kilobajtima, MMAX – maksimum glavne memorije u kilobajtima, CACH – keš memorija u kilobajtima, CHMIN – minimum broja kanala u jedinicama, PRP – objavljena relativna performansna mera, ERP – procenjena relativna performansna mera. Sve vrednosti atributa su celobrojne. U skupu podataka nema null vrednosti.

Način implementacije linearne regresije

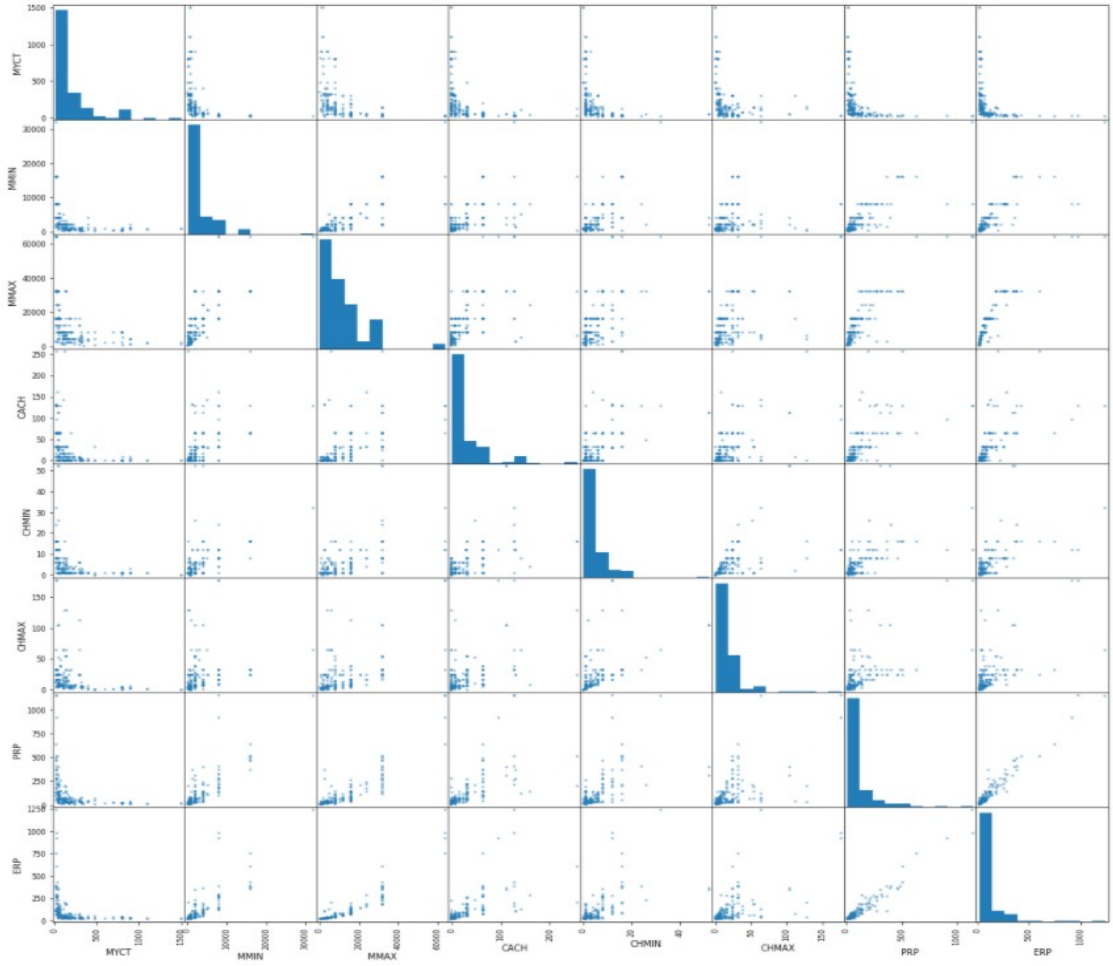
Optimizacija parametara linearne regresije se radi metodom opadajućeg gradijenta. Metod opadajućeg gradijenta treba da minimizuje funkciju gubitaka, koja će u ovom slučaju biti funkcija minimuma kvadrata. Za tačku funkcije se računa gradijent funkcije u toj tački. Nakon izračunavanja gradijenta, sledeća tačka se nalazi kao razlika trenutke tačke i gradijenta funkcije x stopom učenja. Ovim metodom se funkcija kreće ka svom globalnom minimumu (funkcija minimuma kvadrata je konveksna i ima globalni minimum). Ovom iterativnom metodu je samo potrebno zadati stopu učenja (u kodu označeno kao learning_rate) i broj iteracija, pošto je algoritam opadajućeg gradijenta iterativni algoritam.

Treba voditi računa pri izboru parametara da se ne dođe u situaciju da learning rate parametar bude preveliki, jer se u tom slučaju može desiti da iterativna metoda promaši tačku globalnog minimuma. Kada se to desi vrednosti koje se predviđaju osciluju od pozitivnih, do negativnih vrednosti. Funkcija se bukvalno približava globalnom minimumu, međutim promašuje ga i ide na drugu stranu krive. Drugi parametar koji je potrebno navesti u funkciji jeste broj iteracija. Funkcija će generalno dati bolje rezultate što je learning rate parametar manji a broj iteracija veći. U ovom konkretnom datasetu broj iteracija je bio 90000, a learning rate parametar 0.0000000001. Do ovih vrednosti se došlo eksperimentalno.

Model je u kodu predstavljen klasom LinearRegression i ima javni konstruktor gde se setuju learning rate i broj iteracija i gde se za početne težine i grešku postavlja none vrednost. Javnom metodom fit se radi treniranje modela, i optimizacija parametara algoritmom opadajućeg gradijenta. Metodom predict vrši predikcija za ulazne instance skupa podataka.

Opis analize skupa podataka

Nakon vizualizacije skupa podataka, date na slici 1, primećena je pozitivna korelacija između svih atributa i labele koju treba predvideti.



Slika 1 – Korelacija između svih atributa u skupu podataka

Nakon toga je skup podataka podeljen u trening i test skup i metodom fit je treniran model. Nakon treniranja modela, na test skupu podataka je izvršena predikcija i upoređeno je koliko predikcija odstupa u odnosu na realnu vrednost za tu instancu. Od metrika koje su korišćene u evaluaciji modela, upotrebljena je srednja kvadratna greška i dobila se vrednost 1936.43, srednja greška koja je 44, koeficijent korelacije od 0.93 i kvadratni koren koeficijenta korelacije koji je 0.96.

Srednja kvadratna greška je prosek suma kvadrata razlike između predviđene vrednosti i realne vrednosti za tu instancu. Kada se model trenira i optimizuju mu se parametri, ovaj parametar treba da se što više minimizuje. Ovde vidimo da postoji određeno odstupanje kroz srednju kvadratnu grešku, ali da je koeficijent korelacije između predviđenih i realnih vrednosti elementa koji predviđamo opet 0.93 tj dosta veliki.