Jednostruka linearna regresija

Tim mašinski_učenjaci:

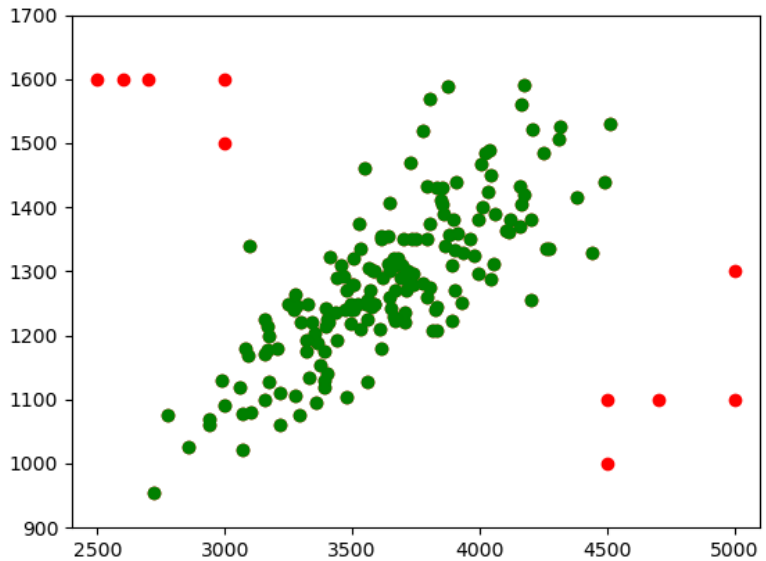
Nikola Zubić (SW-03/2016)
Filip Mladenović (SW-11/2016)
Mihajlo Kušljić (SW-53/2016)

Opis problema:

Dat je skup podataka sa dvije kolone pri čemu se u prvoj koloni nalaze vrijednosti vezane za zapreminu mozga izraženu u kubnim centimetrima (kolona *size*), dok se u drugoj koloni nalaze vrijednosti vezane za masu mozga (kolona *weight*). Koristeći jednostruku linearnu regresiju potrebno je prediktovati masu mozga (Y) u zavisnosti od zapremine mozga (X). Zadatak se smatra uspješno urađenim ukoliko se na kompletnom testnom skupu podataka dobije RMSE (*Root Mean Square Error*) manji od 87.5.

Naš pristup problemu:

Na početku smo *plotovali* podatke iz datog skupa podataka. Potom smo uočili da postoje *outlier*-i koje smo u ovom slučaju morali da uklonimo (označeni crvenim tačkicama na slici 1) usljed toga što smo ograničeni na jednostruki linearni regresioni model.



Slika: 1

U prvom *submission*-u isprobana je *Min-Max* normalizacija, Standardizacija i *Unit Vector* normalizacija.

Isprobani algoritmi:

Za određivanje (treniranje) parametara modela jednostruke linearne regresije korišćena su dva pristupa: *Batch Gradient Descent* (implementiran na dva načina) i *Closed-form solution* (*Normal Equation*).

Ostvareni rezultati:

U prvoj submisiji, uz prethodno uklanjanje *outlier*-a (koji su odstupali za 2 ili više standardnih devijacija), odrađena je standardizacija (*Z-score* normalizacija), te je iskorišćena *Normal equation* za određivanje parametara jednostruke linearne regresije u jednom koraku (analitički) pri čemu je dobijen rezultat: 68.90762.

Konačno (odabrano) rješenje:

Prilikom druge (konačne) submisije, odrađeno je otklanjanje *outlier*-a (koji su odstupali za 2.8 ili više standardnih devijacija, odabrano eksperimentalno, tako što je postojeći trening skup podijeljen *train-test split*-om u omjeru 80% - 20% i varirani raznorazni hiperparametri), nije rađen nikakav vid normalizacije/standardizacije. Za treniranje modela korišćen je *gradient descent* (odrađen na drugi način u odnosu na prvu submisiju), uz pažljiv odabir hiperparametara: broj iteracija, *learning rate*-a, početnih vrijednosti za parametre *theta*, pri čemu je dobijen rezultat: 68.64742.