3. domača naloga: Kombiniranje algoritmov strojnega učenja

Luka Krsnik (63110179)

8. april 2015

1 Uvod

Tretja naloga je razširitev druge naloge, s tem, da smo pri prvi nalogi uporabili le en algoritem strojnega učenja, tokrat pa jih uporabimo več, ter skombiniramo najboljšo rešitev.

2 Metode

LASSO LASSO (Least Absolute Shrinkage and Selection Operator) je regresijska metoda, ki se od ostalih tovrstnih metod razlikuje po tem, da poleg dejanske regresije (učenje funkcije napovedovanja na testnih podatkih) izvede tudi selekcijo parametrov. To konkretno pomeni da iz množice značilk izbere tiste, ki so najbolj uporabni in raznoliki. Zaradi te funkcionalnosti nam ni več potrebno uporabljati metod za izbor značilk (npr. PCA).

Linearna regresija Regresija je postopek, s pomočjo katerega je na podlagi učnih podatkov zgrajena "funkcija", ki iz značilk čim bolje predvidi tisto kar je naučena. Osnovni princip njenega izračuna je tak, da je izražena funkcija za izračun vsote kvadratov napak, mi pa iščemo minimum teh napak.

Ridge Ridge je nekoliko razširjena linearna regresija. Pri njegovi uporabi, lahko določimo še nekatere druge parametre, ki jih pri osnovni linearni regresiji ne moremo.

Elastic Net Elastic Net je kombinacija metod Lasso in Ridge (L1 in L2).

Random Forest Random Forest je metoda, ki zgenerira več odločitvenih dreves. Osnovni pristop je zelo primitiven, vendar večkratno postavljanje dreves in podatki, ki jih iz njih pridobimo ponavadi pripeljejo do dobrih rezultatov.

KNN Metoda KNN je metoda, uporabljena v regresiji. V osnovi deluje s prepoznavanjem povezav med najbližjimi sosedi.

Cross validation Cross validation je metoda, ki ovrednoti uspešnost našega algoritma. Deluje tako, da so učni podatki razdeljeni na k delov (ponavadi 10), nato pa se metoda izvede k-krat. V vsaki iteraciji je eden izmed k-tih delov izpuščen iz učnih podatkov in uporabljen kot testni podatek. Vsi "testni podatkišo nato postavljeni v skupno matriko in ocenjeni.

3 Rezultati

Rezultati so se pri določenih metodah izkazali za precej uspešne (po prečnem preverjanju). Žal pa so se določene metode izkazale za zelo slabe. Takšni metodi sta KNN in linearna regresija. To bi lahko bil eden izmed razlogov za slabe rezultate. (Tabela: 1). Drugi razlog bi lahko bil ta, da so bili podatki takoj po začetku razdeljeni na tiste z visoko intenziteto in razrečenostjo 1/1000. Ker so tudi testni podatki tako filtrirani je možno, da se mi prikazuje napačna slika.

Tabela 1: Tabela z rezultati metode cross validation.

metoda	intenziteta	ostalo
Lasso (5)	1.0	2.4
Ridge (20000)	1.0	2.3
Elastic Net (1)	0.7	2.4
Simple RF	1	2.7
KNN	0.7	1.2
Linear Regression	0.1	1.1
Mean	1	2.8

Tudi stacking je na strežniku dal precej slabe rezultate (0.9).

4 Izjava o izdelavi domače naloge

Domačo nalogo in pripadajoče programe sem izdelal sam.