Domača naloga - 1.del

Izbira metode in optimizacija hiperparametrov

Vito Rozman

2. april 2023

1 Izbira in evalvacija modelov

Za ročno iskanje najboljšega modela sem izbral *knn*, *svm* in *rf*. Podatke sem najprej razdelil v razmerju 1:4 na testne in učne. Na učnih sem preverjal točnost modela z AUC metriko (ploščino pod ROC krvuljo) z metodo prečnega preverjanja (ang. cross validation). Potem sem testeral model še na testnih podatkih. Iskal sem model z naboljšim rezultatom AUC *cross validation*. Opisan potopek sem najprej izvedel na neskaliranih podatkih, potem pa še na skaliranih, ter primerjal rezultate. Iskazalo se je da so skalirani podatki bolje obnesli, vendar pri izbiri najboljšega modela niso privedli do večjih razlik.

1.1 Hiperparamatri pri ročnem iskanju

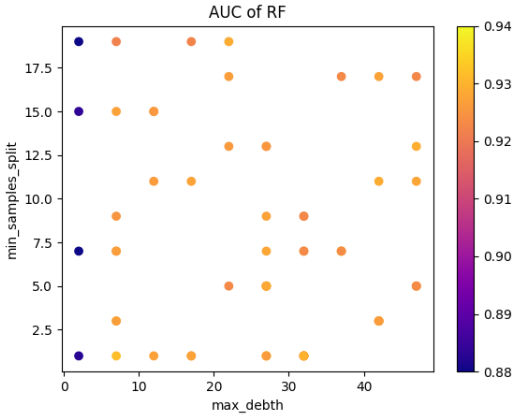
Najbližji sosede: parameter $k = 1 : 30$
Podporni vektorji: izbira *jedra* $\in \{\text{linearno}, \text{polinomsko}, \text{sigmoidno}\}$, parametr $C = \{1, 10\}$
Naključni gozdovi: parameter *največja dovoljena globina* $\in \{2, 7, 12, 17, 22, 27, 32, 37, 42, 47\}$, parameter *najmaša radelitev vzorca* $\in \{1, 3, 5, 7, 9, 11, 13, 15, 17, 19\}$

1.2 Hiperparamatri pri avtomatiziranem iskanju

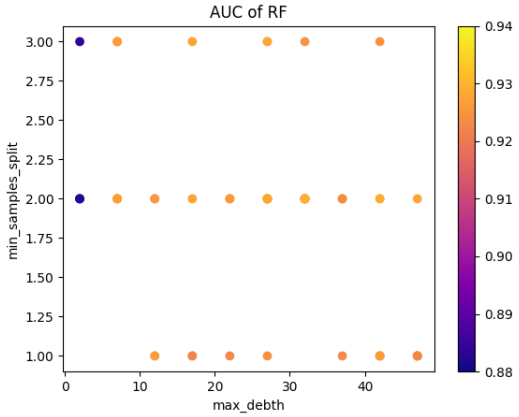
Pri avtomatiziranem iskanje pa sem izbral enake hiperparametre kot pri ročnem, razen pri naključnih gozdovih sem dodal parameter *minimalno število primerov v listu*$\in \{1, 2, 3\}$.

2 Zmogljivosti algoritmov

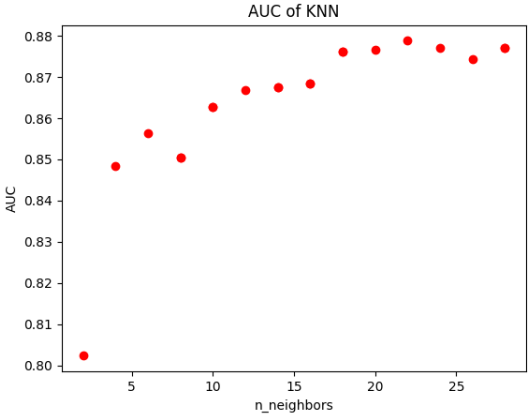
Z izbiro hiperparametrev pri obeh prezkusih sem želel testirati kako učinkovito je avtomatizirano iskanje najbolšega modela in sem predvideval, da bom pri obeh dobil ista modela. Za izbiro parametov pa sem bil malo presenečen, da nisem dobil tako podobnih rezultatov. Pri rf na sliki 1 in sliki 2 vidimo, da za majhno vrednost max-depth dobimo slab rezultat, v vseh ostalik kofiguracijah pa so rezultati primerljivo podobni. Morda bi bilo smiselno vzeti več paramatrov z večjim korakom. Na sliki 3 vidimo, da večanje število sosedov izbolšuje AUC modela. Zaradi malega števila različnih parametrov je na sliki 4 vidno, da so različne konfiguracije z večimi poizkušanji privedle podobne rezultate.



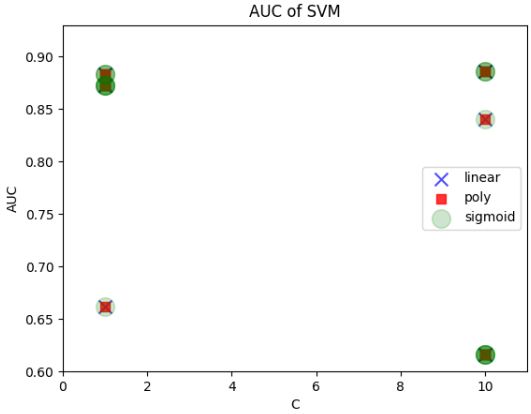
Slika 1: Zmogljivost *rf* glede na parametra max-depth in min-sample-split



Slika 2: Zmogljivost *rf* glede na parametra max-depth in min-sample-leaf



Slika 3: Zmogljivost *knn* glede na parameter n-neighbors



Slika 4: Zmogljivost *svm* glede na jerdo in parameter C

2.1 Najboljši model in njegovi hiperparametri

Ročno iskanje Najbolje se je izkazal model *rf* z izbranimi prarametri: *max-depth*= 37 in *min-sample-split*= 3.
Avtomatizirano iskanje Najbolje se je izkazal model *rf* z izbranimi prarametri: *max-depth*= 27, *min-sample-split*= 1 in *min-sample-leaf*= 2.

rf	AUC - cross validation	AUC test set
Ročno	0.928039	0.887989
Avtomatizirano	0.928101	0.847335

3 Zaključek

Avtomatizirano iskanje najboljšega modela in njegovih hiperparametriv se je izkazalo za dokaj učinkovi pristop, saj sem dobil podobne rezultate pri kot pri ročnem iskanju. Zanimivo mi je bilo, da ko sem uporabil skalirane podatke je *hiperopt* deloval veliko hiteje kot z neskaliranimi podatki.