Mašinsko učenje – Domaći 4

Random forest

S obzirom da *random forest* ima tedenciju da *overfitt*-uje podatke, potrebno je izvrsiti regularizaciju hiperparametara modela. Najbolji nacin da se ovo uradi je pomocu *grid search*-a.

Pomocu *random grid search*-a mogu se zadati opsezi vrednosti hiperparametara od kojih ce on na slucajan nacin birati kombinacije i racunati K-fold kros validaciju za svaku kombinaciju. Na kraju ce biti sacuvan najbolji model i njegovi parametri. Velika prednost *random grid search*-a je to sto se na brz nacin mogu pronaci solidni parametri jer ne razmatra sve moguce kombinacije vec samo one koje je slucajno odabrao.

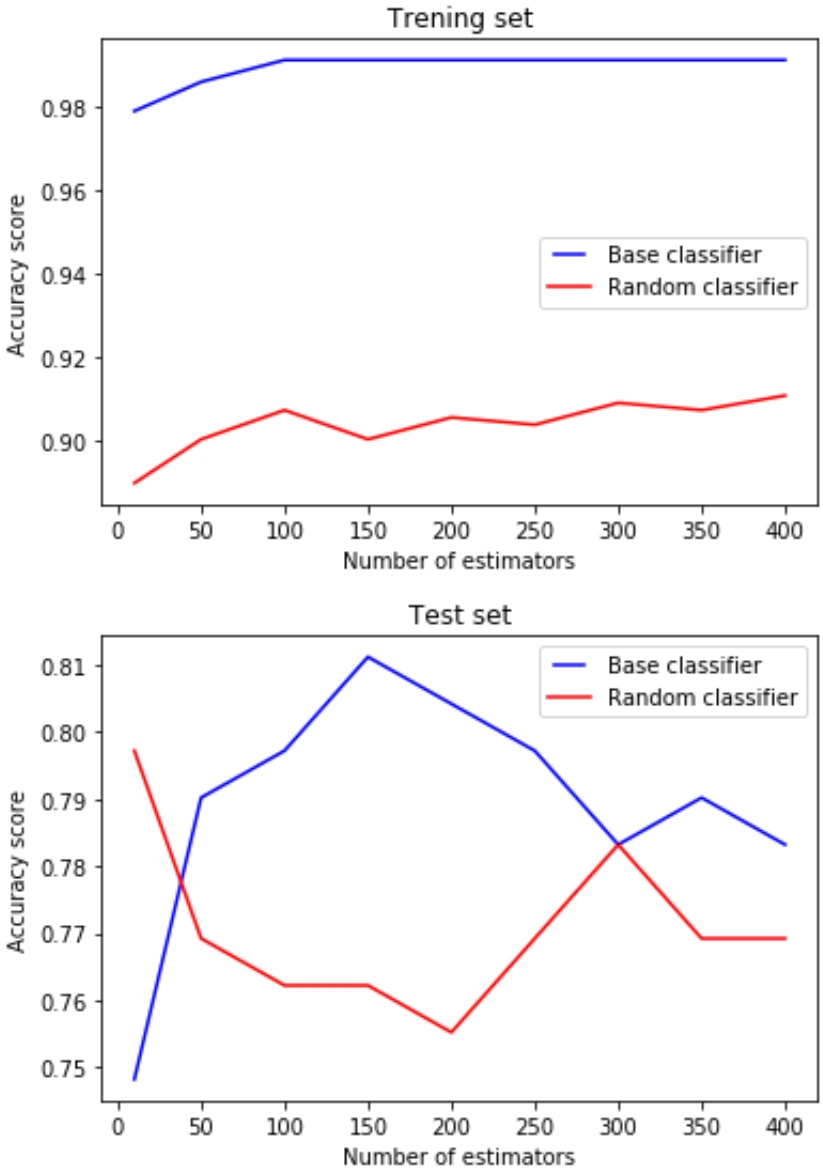
Nakon primene *random grid search*-a moze se primeniti *grid search* sa vrednostima parametara blizu onih koje je pronasao *random grid search*. Na taj nacin se ne ispituju samo slucajne kombinacije hiperparametara vec sve moguce kombinacije, i moguce je pronaci najbolji moguci model.

Parametri koji su isprobani za *random grid search*:

```
{'bootstrap': [True, False],
 'max_depth': [10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, None],
 'max_features': [2, 4, 6, 'auto'],
 'min_samples_leaf': [1, 2, 5],
 'min_samples_split': [2, 5, 10],
 'n_estimators': [10, 50, 100, 150, 200, 250, 300, 350, 400]}
```

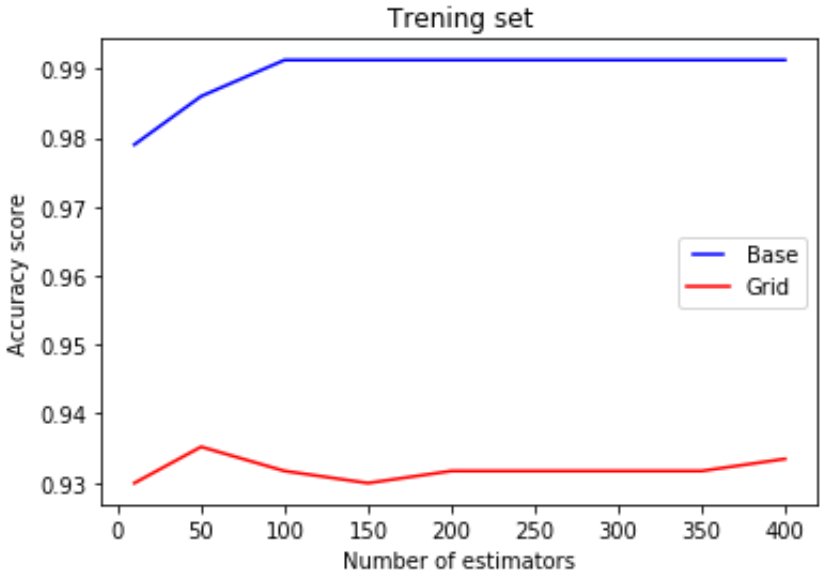
Svaki put kada se pozove ova metoda mogu se dobiti razliciti parametri modela, s obzirom da se uvek na slucajan nacin biraju kombinacije hiperparametara.

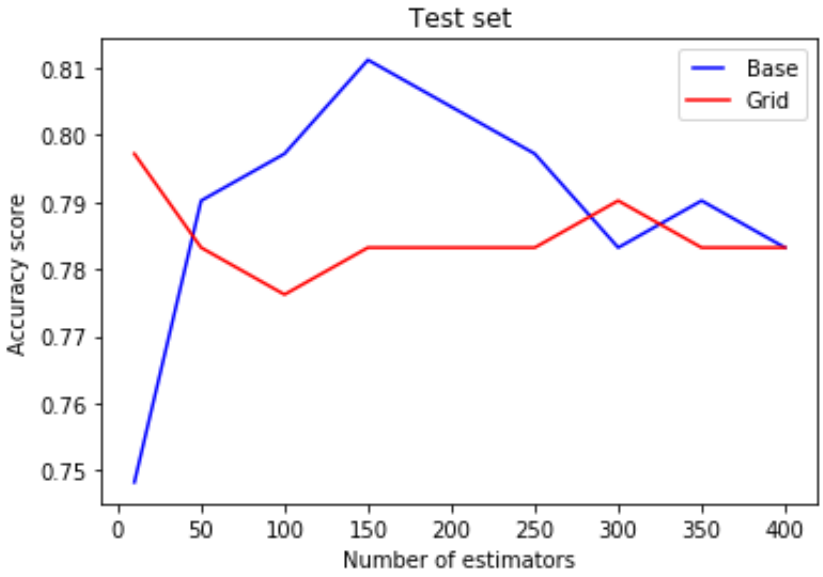
Na sledecoj slici mogu se videti tacnosti za *base* klasifikator bez *tunovanih* hiperparametara i za klasifikator pronadjen pomocu *random grid search*-a. Iako deluje da *base* klasifikator ima bolju tacnost, ako se poredi tacnost na trening setu, moze se uociti da *base* klasifikator *overfitt*-uje podatke.



Nakon analize performansi modela pronadjenog *random grid search*-om, nije lose pokusati i dalje tunovanje hiperparametara pomocu standardnog *grid search*-a.

Vrednosti parametara razmatrane u *grid search*-u su u okolini vrednosti parametara najboljeg moguceg modela koji je pronadjen *random grid search*-om.



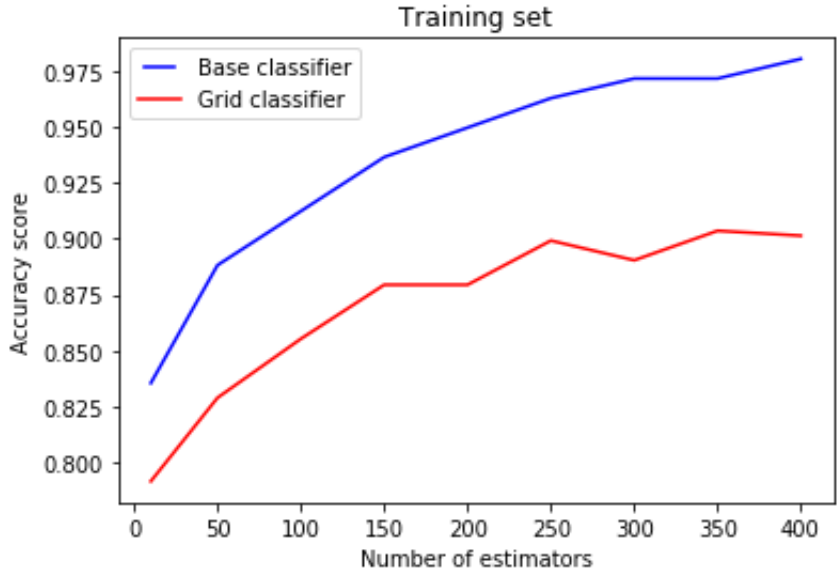


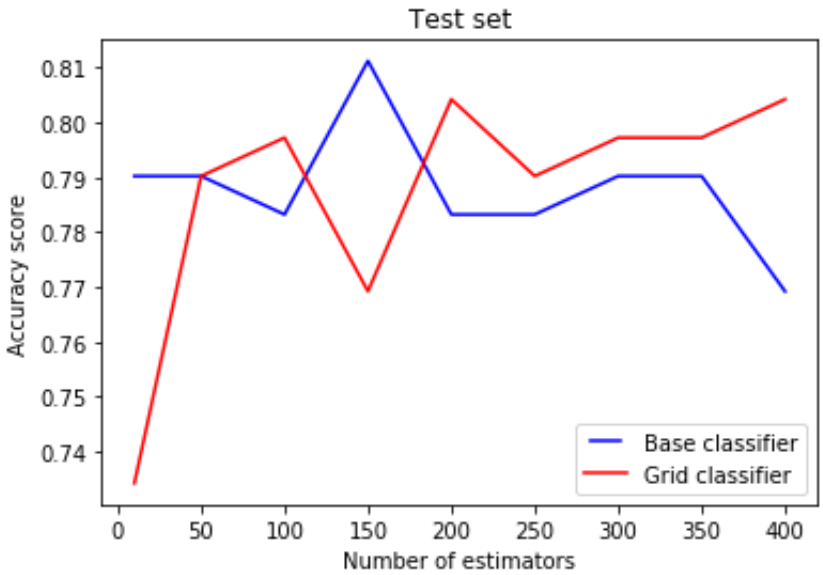
Gradient Boosting Classifier

Kod ovog klasifikatora bitan paramtar koji se pojavljuje, koga nema *Random Forest* klasifikator, je *learning rate*. Parametri koji su razmatrani u *grid search*-u su:

```
{'learning_rate': [0.01, 0.1, 0.5],
 'max_depth': [1, 2, 5, 10],
 'max_features': [2, 4, 6],
 'min_samples_leaf': [1, 2, 3],
 'min_samples_split': [2, 4, 6],
 'n_estimators': [5, 10, 50, 100, 200],
 'subsample': [0.6, 0.8, 1]}
```

Tacnost na treningu i testu za *base* klasifikatora bez zadavanja hiperparametara i najboljeg klasifikatora pronadjenog *gird search*-om prikazana je na sledecim slikama:





Na prvoj slici moze se videti da s povecavanjem broja estimatora raste i tacnost klasifikacije na trening setu. S obzirom da tacnost klasifikacije na test setu ostaje ista, moze se zakljuciti da dolazi do *overfitting*-

a. Pojava *overfitting*-a je znacajno manja kod klasifikatora pronadjenog pomocu *grid search*-a u odnosu na *base* klasifikator.