13M051MU, 4. domaći zadatak 2021/22
Izbor odlika. Stabla. Ansambli.

Podaci sa kojima radite su u datotekama `data_1.csv` i `data_2.csv`. Poslednja kolona označava klasu, a ostale kolone sadrže vrednosti prediktora.

1 Izbor prediktora

Za podatke iz `data_1.csv`, treba da sortirate prediktore od najboljeg do najgoreg, na 2 različita načina:

1) na osnovu koeficijenata korelacije sa ciljnom promenljivom,
2) pomoću "omotač-algoritma", koristeći logističku regresiju bez regularizacije kao klasifikator.

Za svaku od metoda, na istom grafiku nacrtajte tačnosti na trening i validacionom skupu u zavisnosti od broja korišćenih prediktora. Na x-osi naznačite poslednji dodati prediktor. Dakle, za n-tu tačku na ovim graficima, oznaka na x-osi treba da bude indeks n-tog najboljeg prediktora, a vrednost na y-osi treba da odgovara tačnosti na trening, odnosno validacionom skupu, za model treniran samo sa n najboljih prediktora.

2 Obučavanje stabla

Za najbolji par prediktora iz prethodnog zadatka obučite klasifikaciono stablo. Dozvoljeno je korišćenje ugrađenih f-ja i klasa: `sklearn.tree.DecisionTreeClassifier` u Pythonu, odnosno `ClassificationTree` u Matlabu.

Skicirajte zavisnost grešaka klasifikacije na trening i validacionom skupu u zavisnosti od maksimalne dubine stable. Ukoliko biblioteka koju koristite ne

omogućava zadavanje maksimalne dubine, zadajte odgovarajući maksimalan broj čvorova.

Skicirajte granicu odlučivanja u ravni prediktora za 3 različite dubine (ili tri različita maksimalna broja čvorova). Izaberite ove dubine tako da jedno stablo bude pod-obučeno, a jedno pre-obučeno. Na ovim graficima prikažite i primere iz obučavajućeg i validacionog skupa, tako da se sa slike vidi koji odbirak pripada kom skupu i kojoj klasi. Na istoj slici naznačite i granicu odluke.

3 Ansambli

Na podacima iz `data_2.csv` ispitajte kako hiper-parametri utiču na performanse sledeća dva algoritama:

1) Random Forest (RF),
2) AdaBoost (AB), ako radite u Matlabu, ili Gradient Boosting (GB), ako koristite Python.

U oba algoritma koristite stabla odlučivanja kao članove ansambla. Dozvoljeno je korišćenje ugrađenih f-ja za obučavanje klasifikatora: `fitcensemble` u Matlabu, odnosno `RandomForestClassifier` i `GradientBoostingClassifier` u Pyhon modulu `sklearn.ensemble`.

Za svaku kombinaciju algoritma i hiper-parametra treba da generišete dva grafika sa po dve krive. Na prvom grafiku prikažite tačnosti na obučavajućem skupu, a na drugom tačnosti na validacionom skupu, za (bar) dve različite vrednosti hiper-parametra, u zavisnosti od veličine ansambla. Za oba algoritma ispitajte uticaj veličine stabala. Dodatno, za RF analizirajte i maksimalan broj odlika koje se razmatraju pri generisanju čvorova stabla, a za AB i GB stopu učenja.

Kôd u Pythonu ili Matlab/Octave, i izveštaj sa traženim graficima u pdf formatu, predaje se putem MS Teamsa. *Ne zaboravite da kliknete na Turn In!*