Domača naloga 1, Tanja Gošnjak

Pregled podatkov

Iz repozitorija UCI smo prenesli datoteko podatkov Spambase. Ta vsebuje podatke neželene e-pošte (spam) in klasične e-pošte (ham). Podatki »neželene pošte« so raznoliki: oglasi za izdelke/splet, spletna mesta, sheme za hitri zaslužek, verižna pisma, pornografija ... Zbirka klasične e-poštne prihajajo iz službenih in osebnih e-poštnih sporočil.

Podatke smo analizirali v programskem jeziku Python. Podatki vsebujejo 4601 primerov e-poštnih sporočil in 57 atributov. Zadnji stolpec v podatkih je razred, ki določa ali je e-poštno sporočilo spam (1) ali ne (0). Večina atributov določa, kako pogosto se določen znak ali beseda pojavi v e-poštnem sporočilu. Drugi atributi merijo dolžino/povprečje besed, ki so napisane z velikimi črkami. Vse spremenljivke so zvezne, večina je delimalnih, dve pa predstavljajo cela števila. Razred je definiran nominalno, z 0 in 1.

V podatkovnem okvirju, kot navaja tudi dokumentacija, nismo našli manjkajočih vrednosti.

V podatkovnem okvirju je 2788 (60,60 %) primerov ham in 1813 (39,40 %) primerov spam.

Vsi atributi so podobno porazdeljeni, kjer je večina vrednosti skoncentriranih pri manjših vrednostih glede na rang, ki ga atribut zaseda. Atributi niso porazdeljeni normalno.

Ustvarjanje modelov

Podatke smo razdelili v testno in učno množico, tako da smo v učno množico razporedili 80% podatkov. Ker imamo več podatkov ham kot spam, smo določili parameter delitve, da bo tudi v testni oziroma učni množici enak delež obeh razredov.

Naredili smo več primerov modelov. Naredili smo pet primerov odločitvenih dreves. V prvem primeru smo uporabili privzete nastavitve, v drugem primeru smo določili maksimalno globino drevesa na 5, v tretjem primeru smo za mero uporabili entropijo (v privzetih nastavitvah je indeks Gini), v četrtem primeru smo uporabili kombinacijo mere entropije in maksimalne globine 5 ter v zadnjem primeru smo poskušali med različnimi parametri s cross-validacijo najti najprimernejše parametre za odločitveno drevo. Naredili smo še model Random forest s privzetimi nastavitvami in model z Naivnim Bayesom s privzetimi nastavitvami, kjer smo podatke najprej standardizirali.

Evalvacija modelov

Za vse modele smo izpisali točnost, ocenjevalno poročilo in matriko lažno in resnično pozitivnih ter lažno in resnično negativnih vrednosti, ki smo jih napovedali. Ugotovimo, da ima izmed testiranih modelov, največjo točnost Random Forest (0.948), najslabšo pa naivni Bayes (0.826). Izmed odločitvenih dreves smo najboljši rezultat dobili z izbiro med naborom parametrov (0.918), sledilo je odločitveno drevo z mero entropije (0.912), odločitveno drevo z mero entropije in maksimalno globino=5 (0.906), odločitveno drevo z maksimalno globino=5 (0.902) in odločitveno drevo s privzetimi nastavitvami (0.900). Vidimo, da so razlike majhne. Predvsem je majhna razlika med odločitvenim drevesom z optimalnimi parametri (izbrali smo mera entropija, brez makismalne globine in splitter': 'best') in tistim z izbrano mero entropije. To kaže, da je (v našem primeru) entropija boljša mera nečistoč. Z najboljšim modelom odločitvenih dreves smo določili (testni set, n = 921) 46 lažnih spam in 29 lažnih ham sporočil. Z Random Forest modelom smo določili 20 lažnih spam sporočil in 27 lažnih ham sporočil. Z naivnim Bayesom smo določili 150 lažnih spam sporočil in 10 lažnih ham sporočil.