NSU 2022/23 – prva domača naloga

17.3.2023

Na voljo sta dve nalogi, osnovna in napredna. Vsaka je vredna 10 točk. Rešuješ lahko samo osnovno nalogo, ali pa obe.

Podatkovje podatki.csv je bilo vzeto s strani https://www.openml.org/, nato pa malo spremenjeno: 1. Izbrisane so bile nekatere značilke, vrstni red preostalih pa je bil premešan. 2. Izbrisani so bili nekateri primeri, vrstni red preostalih pa je bil premešan. 3. Vrednosti stolpcev so bile spremenjene, tako da ne morejo bistveno vplivati na lastnosti podatkov.

Vedi, da za število izbrisov velja naslednje:

- Izbrisali smo največ 100 značilk.
- Izbrisali smo največ 1000 primerov.

Na koncu smo vse stolpce še preimenovali. Ciljna spremenljivka v podatkih nosi ime y in je dvojiška. Ker je precej neuravnotežena, uporabi za ocenjevanje zmogljivosti modelov ploščino pod ROC krivuljo namesto točnosti.

1 Izbira metode in optimizacija hiperparametrov

1.1 Ročno

Ročno poišči metodo strojenga učenja, ki na danih podatkih deluje dobro. Nastavi tudi njene hiperparametre, tako da bo delovala čim boljše.

1.2 Avtomatizirano

Uporabi orodje za avtomatizirano strojno učenje (kot je hyperopt) da poiščeš najboljšo metodo strojnega učenja in najboljše vrednosti hiperparametrov za dane podatke. Izberi vsaj tri klasifikacijske algoritme. Vsak od njih naj ima vsaj en hiperparameter, katerega optimalno vrednost je treba najti. Vsaj en od algoritmov naj ima več kot le en hiperparameter. Optimizacijo hiperparametrov izvedi s prečnim preverjanjem, zmogljivost izbranega algoritma z optimalno konfiguracijo parametrov pa preizkusi na testni množici. Primerjaj zmogljivost modela, ki si ga nastavil sam, ter modela, izbranega in optimiziranega avtomatsko. Ne pozabi, za ocenjevanje zmogljivosti uporabljaj ploščino pod ROC krivuljo.

Poročilo za prvi del domače naloge naj bo dolgo največ 300 besed in naj vsebuje:

- opisk postopka ročnega iskanja in nastavljanja metode ter izbrano konfiguracijo (1.1),
- natančen opis preiskovanega prostora konfiguracij v (1.2),
- najboljšo konfiguracijo (1.2),
- graf porazdelitev zmogljivosti (ki jih porodijo različne konfiguracije) za vsakega od algoritmov, ter tvoj komentar grafov (1.2),
- zmogljivosti najboljših konfiguracij, pridobljenih ročno in avtomatsko (1.1 in 1.2).

2 Meta učenje

Izberi si primerne metaznačilke, s katerimi opišeš podatkovja. V prostoru, ki ga te značilke razpenjajo, s pomočjo openml.org najdi tri podatkovja, ki so najbolj podobna temu iz podatki.csv. Za iskanje sosedov lahko uporabiš algoritem k najbližjih sosedov (k = 3). Ko najdeš sosede, ugotovi, kateri algoritem za klasifikacijo se je najbolje odrezal na njih, sodeč po obstoječih rezultatih na OpenML. Tako dobiš največ tri različne kandidate – za vsakega soseda enega.

Izberi enega od treh kandidatnih algoritmov in ga poženi na podatki.csv. Njegovo zmogljivost oceni na isti testni množici kot v prvem delu domače naloge.

Za konec še primerjaj zmogljivosti algoritma, izbranega s pomočjo meta učenja s tistim, ki si ga izbral_a ročno, ter s tistim, ki si ga izbral_a avtomatizirano. Je bilo meta učenje koristno? Bi zdaj v prvem delu naloge naredil_a kaj drugače?

Poročilo za drugi del domače naloge naj bo dolgo največ 300 besed in naj vsebuje:

- imena uporabljenih meta značilk,
- imena treh najbližjih sosedov,
- najboljši algoritem za klasifikacijo za vsakega od treh sosedov ter njegovo zmogljivost na podatki.csv,
- utemeljitev izbire končnega algoritma za klasifikacijo,
- diskusijo zmogljivosti vseh treh izbranih algoritmov (1.1, 1.2, 2.).

Primer iskanja kandidatov. Uporabimo zavihek Tasks na openml.org. Če je eden od odkritih sosedov podatkovje iris z ID-jem 61, izvedemo poizvedbo
https://www.openml.org/search?type=data&status=active&id=61 in izberemo zavihek Tasks. Potem izberemo enega od zadetkov z veliko poganjanji, navigiramo na zavihek Analysis ter najdemo najboljši algoritem. Za iris je to metoda podpornih vektorjev (SVC).