Nadzorovano modeliranje

Primož Pečar (63150213)

7. maj 2017

1 Uvod

Cilj naloge je bilo izvajati nadzorovano modeliranje oz. napovedovanje ocen uporabnikov. Potrebno je bilo narediti zvezno napoved, torej smo naredili regresijski model in z njim smo napovedovali ocene uporabnikov. Narediti je bilo treba še klasifikacijski model, kjer smo ugotavljali, ali je uporabnik ocenil film z manj ali več kot 3.

2 Podatki

Delal sem na podatkovni množici MovieLens in sicer z datotekami:

1. Movies
2. Cast

Movies in Cast sem oboje spremenil v xlsx format, da sem ga lahko obdelal s python modulom OpenPyXl.

Kreiral sem še dve nove datoteke, katere sem uporabil pri obeh nalogah, te bodo bolj podrobno opisanje kasneje.

3 Regresija

Potrebno je bilo narediti regresijski model, za vsakega od uporabnikov, to sem dosegel na sledeč način, izbral sem 100 najbolj gledanih filmov, torej filmi kateri so imeli rating več kot 100. Vzel sem vse filme, jih sortiral po padajočem vrstnem redu (glede na oceno) in omejil seznam na 100.

Nato sem še podobno naredil za uporabnike, imel sem slovar uporabnikov, katerega sem sortiral po vrednosti (št. ocenah), nato pa omejil na 100 elementov.

Kar je bilo potrebno še narediti, je preveriti ali je uporabnik podan film ocenil, če ga ni sem pa tam zapisal 0 (po zahtevi profesorja).

Tako dobimo tabelo 100x100, kjer imamo v stolpcih uporabnika in njegove ocene za podan film, v vrstici. Nad tem za vsakega uporabnika posebaj kreiramo regresijski model, kjer razdelimo učno množico na 75%, testno 25%. Za mero ocene sem uporabljal MSE in MAE.

3.1 Rezultati

Naredil sem model za vsakega uporanika posebaj, in si zapomnil njegov MSE oceno in MAE, to sem nakoncu delil s št. modelom, da sem dobil povprečno napoved. Primer napovedi so bile sledeče

Tabela 1: Rezultati regresije Python

Uporabnik	MSE	MAE
User 42	3.43	1.61
User 43	3.30	1.55
User 44	3.44	1.61
User 45	3.44	1.61
User 46	3.45	1.59

Povprečni Mean Squared Error je bil **3.3989**, Mean Absolute Error pa **1.5975**.

Potem sem še meritve testira v Orange3, te so bile sledeče (izvedene z naključnim samplingom, 75% testna, 25% učna, 3x ponovno učenje/testiranje). Uporabil sem Ridgevo metodo z alfo 0.0001. Načeloma, kar se tiče MSE, čim bližje 0 smo, boljši je model.

Tabela 2: Rezultati regresije Orange3

Uporabnik	MSE	MAE
User 42	13.239	2.789
User 43	10.186	2.623
User 44	7.041	2.223
User 45	15.629	3.156
User 46	6.774	1.958

Sedaj vidim veliko razliko med mojimi rezultati in med temi, ki jih vrne Orange3. Sicer so metode ocenjevanja različna, vendar so to še kar velike razlike. Če bi namesto MSE pri Orange3 gledal RMSE, pa bi imel bolj smiselne rezultate, kjer so za vsakega uporabnika okol 2.7-3.6.

4 Klasifikacija

Podobno kot pri regresiji, sem naredil novo xlsx datoteko, dimenzij 100x100, kjer imamo uporabnike in njihove ocene, namesto samih ocen pa sem po formuli, če je y¿3 = 1, else 0, nastavil vrednosti za vsako oceno. Izvajal sem klasifikacijo z Naivnim Bayesovim klasifikatorjem, kjer je bila učna množica 0.4, testna 0.6.

Zgradil sem klasifikacijski model za vsakega od uporabnikov, in dobil sledeče rezultate.

Povprečna kasifikacijska točnost je bila **0.8342**. To se zdi kot super model, ampak kar jaz domnevam je, da narobe uporabljam funkcijo v pythonu (predvideval sem da funkcija sama odstrani testne iz učnih, saj drugače nisem moral podati podatkov). Tukaj najvrjetneje prihaja

Tabela 3: Rezultati klasifikacije Python

Uporabnik	CA
User 84	0.725
User 85	0.825
User 86	0.775
User 87	0.9
User 88	0.85

do overfittinga učnih podatkov, ali pa tudi, da primer na katerem se učimo, je vsebovan v učni množici. Torej ta model je bolj ali manj neuporaben.

Enako kot pri regresiji sem testiral še v Orange3, tam pa z kNN in NB.

Tabela 4: Rezultati klasifikacije NB Orange3

Uporabnik	CA	Precision
User 84	0.5	0.917
User 85	0.508	0.783
User 86	0.625	0.686
User 87	0.642	0.673
User 88	0.583	0.617

Tukaj pa so rezultati dokaj slabši, vendar bolj smiselni kot zgornji. Problem tukaj je, da smo še od prej ohranjali vrednosti 0, tam kjer uporabnik ni ocenil filma, tukaj se to tudi prevede v to da je ocenil film z manj kot 3. Torej imamo veliko več ničel kot pa enic, zato model tudi slabše ocenuje celotno zadevo. Poskusil sem še kNN, dobil pa sem sledeče rezultate.

Tabela 5: Rezultati klasifikacije KNN Orange3

Uporabnik	CA	Precision
User 84	0.9	0.810
User 85	0.825	0.681
User 86	0.7	0.676
User 87	0.742	0.753
User 88	0.65	0.637

Ta pa dosega bistveno boljše rezultate, kot pri NB, to pa tudi zaradi tega ker izvedemo preverjanje 10x čez iste podate, tukaj imamo tudi možnost overfittinga.

5 Ocena samega sebe

Izpolnil sem excel datoteko z novim uporabnikom, in mu dodal vrednosti filmov, katere sem si ogledal. Vse je bilo subjektivno. Ker veliko filmov nisem še gledal, predvidevam, da se bodo

ocene nagibali proti 1. Za napoved 25 filmov iz množice, sem dobil naslednje napovedi zame (2.25031577 2.4623364 2.4623364 2.50717537 2.25684872 2.51240173 2.25031577 2.25684872 2.4083513 2.45711004 2.30822064 2.41096448 2.25684872 2.25031577 2.4623364 2.25423554 2.41096448 2.25554213 2.51370832 2.51370832 2.25684872 2.25684872 2.25162236 2.35959256 2.256848723).

Rezultati se mi zdijo smiselni, napoved se giblje proti 0, vendar ker sem nekaj filmov ocenil z 5, veliko tudi z 2,3. Bi rekel, da so sledeče ocene vredu, mogoče malce preveč nizke.

6 Izjava o izdelavi domače naloge

Domačo nalogo in pripadajoče programe sem izdelal sam.