3. domača naloga pri predmetu Podatkovno rudarjenje

Nadzorovano modeliranje

24. april 2017

1 Uvod

Spoznali bomo praktično uporabo enostavnih metod nadzorovanega modeliranja oz. napovedovanja. Skupna lastnost vseh omenjenih metod je, da s pomočjo naključnih spremenljivk (atributov) modelirajo vrednosti posebne spremenljivke, ki ji pravimo *razred* (v kontekstu uvrščanja v razrede, klasifikacije) ali *odziv* (v kontekstu regresije). Osnovne razlike med kontekstoma smo spoznali na predavanjih in vajah.

Praktična cilja, ki ju bomo zasledovali sta

1. modeliranje ocen posameznega uporabnika (odziva) s pomočjo vseh ostalih uporabnikov,
2. primerjava metod nadzorovanega modeliranja.

2 Podatki

Opis podatkovne zbirke MovieLens 1996-2016 ostaja enak prvi nalogi.

3 Predpriprava podatkov

Za potrebe te naloge bomo podatke pripravili na naslednji način:

1. Izberi m filmov z vsaj 100 ogledi.
2. Izberi n uporabnikov, ki si je ogledalo vsaj 100 filmov.
3. Pripravi matriko $\mathbf{X}$ velikosti $m \times n$, kjer vrstice predstavljajo filme, stolpci pa uporabnike. Neznane vrednosti zamenjaj z 0.

Za vsakega od izbranih n uporabnikov bo zgrajen klasifikacijski in regresijski model, katerega cilj bo napoved ocen za filme.

4 Vprašanja

1. (50 %) **Regresija**. Za vsakega od uporabnikov postavite regresijski model. Uporabite eno ali več metod za učenje regresijskih modelov (linearna regresija, Ridge, Lasso, itd.).

   Za vsakega od n uporabnikov izberite ustrezni stolpec v matriki podatkov. Za uporabnika i imamo torej

   - Vektor odziva $\mathbf{y}^{(i)}$,
   - Matriko podatkov $\mathbf{X}^{(i)}$, ki vsebuje vse stolpce *razen* i.

   Za lažjo predstavo si oglej Tabelo 1. Nekajkrat (npr. trikrat) ponovite postopek preverjanja s pomočjo učne in testne množice:

   (a) Množico filmov, ki si jih je uporabnik ogledal, *naključno* razdelite v razmerju 75% (učna množica) in 25 % (testna množica).

   (b) Naučite regresijski model na učni množici (izberite ustrezne vrstice v $\mathbf{X}$ in $\mathbf{y}$).

   (c) Ovrednotite model na testni množici (ponovno izberite ustrezne vrstice v $\mathbf{X}$ in $\mathbf{y}$).

   Oceno vrednotenja nato delite s številom poizkusov, da dobite končno oceno.

   Poročajte o uspešnosti vašega modela. Pri tem se osredotočite na naslednja vprašanja:

   - Utemeljite ustrezno mero vrednotenja. Ali model dobro napoveduje ocene?
   - Z izbrano mero ocenite modele za vseh n uporabnikov.

		$\mathbf{y}^{(0)}$	$\mathbf{X}^{(0)}$		
	Film/uporabnik	u_0	u_1	u_2	$\cdots$
f_1	Twelve Monkeys (a.k.a. 12 Monkeys) (1995)	0	0	2.5	$\cdots$
f_2	Dances with Wolves (1990)	4.0	0	0	$\cdots$
f_3	Apollo 13 (1995)	0	2.0	0	$\cdots$
f_4	Sixth Sense, The (1999)	3.0	0	4.0	$\cdots$
$\cdots$	$\cdots$	$\cdots$			

		$\mathbf{y}^{(1)}$	$\mathbf{X}^{(1)}$		
	Film/uporabnik	u_1	u_0	u_2	$\cdots$
f_1	Twelve Monkeys (a.k.a. 12 Monkeys) (1995)	0	0	2.5	$\cdots$
f_2	Dances with Wolves (1990)	0	4.0	0	$\cdots$
f_3	Apollo 13 (1995)	2.0	0	0	$\cdots$
f_4	Sixth Sense, The (1999)	0	3.0	4.0	$\cdots$
$\cdots$	$\cdots$	$\cdots$			

Tabela 1: Razdelitev podatkov za model uporabnika u_0 (zgoraj) in uporabnika u_1 (spodaj).

2. (50 %) Izberite poljubni klasifikacijski model. V enem odstavku opišite njegove lastnosti, morebitne predpostavke, parametre in algoritem za učenje. Vrednosti v vektorju $\mathbf{y}$ razdelite v dva *razreda*:

$$y' = \begin{cases} 0 & \text{če } y \leq 3; \\ 1 & \text{če } y > 3. \end{cases}$$

   Ponovite analizo iz prejšnjega vprašanja z uporabo klasifikacijskih modelov. Izberite ustrezno mero vrednotenja za klasifikacijo. Ali je klasifikacijski problem lažji od regresijskega? Utemelji odgovor.

3. (Bonus 15 %) Ustvarite novega uporabnika, ki predstavlja vaše ocene filmov. Ocenite nekaj filmov po lastnem okusu in preverite, kako modeli ocenijo neizbrane filme. Ali se vam zdijo napovedi primerne?

5 Zapiski

Implementacijo, opis in vrednotenje metod za nadzorovanjo učenje vsebujejo knjižnice `sklearn` ali `Orange`.

6 Oddaja poročila

Oddaja vključuje datoteko `vpisnast_priimek_ime.zip` z naslednjo vsebino:

- Poročilo z odgovori na vprašanja. Oddajte tako datoteko `.tex` kot `.pdf`. **Pomembno: oddaje, ki ne bodo vsebovale poročil, ne bodo ocenjene.** Vzorec poročila najdete na **spletni učilnici predmeta**.
- morebitne slike, ki jih vsebuje poročilo,
- vso izvorno kodo za pridobitev rezultatov.