2. domača naloga pri predmetu Podatkovno rudarjenje

Nenadzorovano modeliranje

27. marec 2017

1 Uvod

Skupni imenovalec vseh metod modeliranja je iskanje strukture v podatkih. Metode nenadzorovanga modeliranja uporabljamo, da razložimo zgodbo v ozadju. Povedano drugače, zanima nas *proces*, ki generira podatke.

V tej nalogi bomo uporabili modeliranje verjetnostnih porazdelitev ter metode za iskanje skupin (gručenje). Praktična cilja, ki ju bomo zasledovali sta

1. iskanje osamelcev,
2. iskanje sorodnih primerov v podatkih.

2 Podatki

Opis podatkovne zbirke MovieLens 1995-2016 ostaja enak prvi nalogi.

3 Vprašanja

Z uporabo principov, ki ste jih spoznali na vajah in predavanjih, odgovorite na spodnja vprašanja. Pri vsakem vprašanju dobro premislite, na kakšen način boste najbolje podali, prikazali oz. utemeljili odgovor. Bistven del so odgovori na vprašanja in ne implementacija vaše rešitve.

3.1 Iskanje osamelcev

1. (50 %) O ocenah katerih filmov so si uporabniki najmanj enotni? Povedano drugače, za katere filme so pripadajoče ocene najbolj razpršene?

   Formuliraj problem kot modeliranje verjetnostne porazdelitve. Premisli o naslednjih vprašanjih, naredi ustrezne poizkuse in odgovori.

   (a) Katera je ustrezna naključna spremenljivka (količina) v podatkih, ki odgovarja na vprašanje?

(b) Nariši njeno porazdelitev (npr. s pomočjo histograma).

(c) Ali porazdelitev spominja na kakšno znano porazdelitev? Ali je porazdelitev morda normalna ali katera druga?

(d) Oceni parametre te porazdelitve s pomočjo postopkov, ki smo jih spoznali na vajah.

(e) Izmed porazdelitev, ki smo jih spoznali na vajah, izberi tisto, ki se podatkom najbolj prilega.

(f) Izpiši filme z vrednostjo naključne spremenljivke, ki spada v zgornjih 5 % statistično značilnih primerov.

3.2 Gručenje filmov

2. (50 %) Priporočilni sistemi pogosto odkrivajo predmete (v našem primeru filme), za katere velja visoka podobnost.

   Poiščite 100 najbolj gledanih filmov. Ali med njimi obstajajo skupine? Uporabite ustrezen algoritem za gručenje. Na film lahko gledamo kot vektor, s številom komponent enakim številom uporabnikov. Vektorji vsebujejo tudi *neznane vrednosti*. Primer vektorjev za deset filmov prikazuje Tabela 1.

   Algoritme gručenja lahko izvajamo v izvornem prostoru (koordinatni sistem filmi-uporabniki), ali pa filme primerjamo z merami podobnosti, ki smo jih spoznali na vajah. Premisli, kateri način je primernejši glede na obliko podatkov.

	Movie	u_0	u_1	u_2	$\cdots$
$\vec{x}_0$	Fight Club (1999)	?	?	?	$\cdots$
$\vec{x}_1$	Twelve Monkeys (a.k.a. 12 Monkeys) (1995)	?	?	2.5	$\cdots$
$\vec{x}_2$	Independence Day (a.k.a. ID4) (1996)	?	?	?	$\cdots$
$\vec{x}_3$	Dances with Wolves (1990)	4.0	?	?	$\cdots$
$\vec{x}_4$	Fargo (1996)	?	?	?	$\cdots$
$\vec{x}_5$	Speed (1994)	?	?	?	$\cdots$
$\vec{x}_6$	Apollo 13 (1995)	?	2.0	?	$\cdots$
$\vec{x}_7$	Seven (a.k.a. Se7en) (1995)	?	?	?	$\cdots$
$\vec{x}_8$	Sixth Sense, The (1999)	3.0	?	4.0	$\cdots$
$\vec{x}_9$	Aladdin (1992)	?	?	?	$\cdots$
$\cdots$	$\cdots$	$\cdots$			

Tabela 1: Primer vektorjev ocen za filme

Pri tem odgovori na naslednja vprašanja.

(a) Utemelji izbiro algoritma in mere podobnosti.

(b) Koliko skupin filmov je med izbranimi? Ali poznamo kvantitativne ocene za različne možnosti razvrščanja v skupine?

(c) Prikaži rezultate z uporabo ustrezne vizualizacije.

(d) Komentiraj smiselnost dobljenih rezultatov.

4 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.