Nenadzorovano modeliranje

Amon Stopinšek (63150273)

2. april 2017

1 Uvod

V nalogi smo se lotili iskanja osamelcev in gruč. V prvem delu naloge smo poiskali filme o katerih so si gledalci najmanj enotni. Problema smo se lotili z izračunom variance za vsak film in p testa. V drugem delu naloge smo poiskali filme, ki so si med seboj najbolj podobni.

2 Podatki

Pri nalogi smo uporabili podatkovno zbirko Movie Lens.

3 Metode

3.1 Iskanje osamelcev

Za iskanju osamelcev na podlagi variance ocen smo najprej preoblikovali originalno obliko podatkov v datoteki ratings.csv v matriko kjer stolpci predstavljajo filme, vrstice pa uporabnike, vrednosti pa predstavljajo oceno filma določenega uporabnika.

Tabela 1: Atributi in njihove zaloge vrednosti.

0	1	2
3	5	6
15	2.0	2.0
NaN	4.5	4.0
16	NaN	NaN
NaN	NaN	NaN
17	NaN	NaN
NaN	NaN	4.5
18	NaN	NaN
NaN	3.0	4.0
19	3.0	3.0
3.0	NaN	3.0

Pri izračunu

4 Rezultati

V tem poglavju podaš rezultate s kratkim (enoodstavčnim) komentarjem. Rezultate lahko prikažeš tudi v tabeli (primer je tabela 2).

Odstavke pri pisanju poročila v LaTeX-u ločiš tako, da pred novim odstavkom pustiš prazno vrstico. Tudi, če pišeš poročilo v kakšnem drugem urejevalniku, morajo odstavki biti vidno ločeni. To narediš z zamikanjem ali pa z dodatnim presledkom.

Tabela 2: Atributi in njihove zaloge vrednosti.

ime spremenljivke	definicijsko območje	opis
cena	[0, 500]	cena izdelka v EUR
teža	[1, 1000]	teža izdelka v dag
kakovost	[slaba—srednja—dobra]	kakovost izdelka

Podajanje rezultati naj bo primerno strukturirano. Če ima naloga več podnalog, uporabi podpoglavja. Če bi želel poročati o rezultatih izčrpno in pri tem uporabiti vrsto tabel ali grafov, razmisli o varianti, kjer v tem poglavju prikažeš in komentiraš samo glavne rezultate, kakšne manj zanimive detajle pa vključite v prilogo (glej prilogi A in B).

5 Izjava o izdelavi domače naloge

Domačo nalogo in pripadajoče programe sem izdelal sam.

Priloge

A Podrobni rezultati poskusov

Če je rezultatov v smislu tabel ali pa grafov v nalogi mnogo, predstavi v osnovnem besedilu samo glavne, podroben prikaz rezultatov pa lahko predstaviš v prilogi. V glavnem besedilu ne pozabi navesti, da so podrobni rezultati podani v prilogi.

B Programska koda

Za domače naloge bo tipično potrebno kaj sprogramirati. Celotno kodo oddaj zapakirano skupaj s poročilom v datoteki zip. V kolikor je določen izsek kode nujen za boljše razumevanje poročila, ga vključi v prilogo poročila.

Čisto za okus sem tu postavil nekaj kode, ki uporablja Orange (`http://www.biolab.si/orange`) in razvrščanje v skupine.

```
import random
```

```
import Orange

data_names = ["iris", "housing", "vehicle"]
data_sets = [Orange.data.Table(name) for name in data_names]

print "%10s %3s %3s %3s" % ("", "Rnd", "Div", "HC")
for data, name in zip(data_sets, data_names):
    random.seed(42)
    km_random = Orange.clustering.kmeans.Clustering(data, centroids = 3)
    km_diversity = Orange.clustering.kmeans.Clustering(data, centroids = 3,
        initialization=Orange.clustering.kmeans.init_diversity)
    km_hc = Orange.clustering.kmeans.Clustering(data, centroids = 3,
        initialization=Orange.clustering.kmeans.init_hclustering(n=100))
    print "%10s %3d %3d %3d" % (name, km_random.iteration, \
    km_diversity.iteration, km_hc.iteration)
```