Nenadzorovano modeliranje

Amon Stopinšek (63150273)

10. april 2017

1 Uvod

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

2 Podatki

Pri nalogi sem uporabili podatkovno zbirko Movie Lens.

3 Metode

3.1 Iskanje osamelcev

Za iskanju osamelcev na podlagi variance ocen sem najprej preoblikoval originalno obliko podatkov v datoteki ratings.csv v matriko1 kjer stolpci predstavljajo filme, vrstice uporabnike, vrednosti pa predstavljajo oceno filma določenega uporabnika.

```
import pandas as pd
dfRatings = pd.read_csv('../data/ratings.csv')
df = dfRatings.pivot(index='userId', columns='movieId', values='rating')
```

Tabela 1: Izsek iz matrike ocen

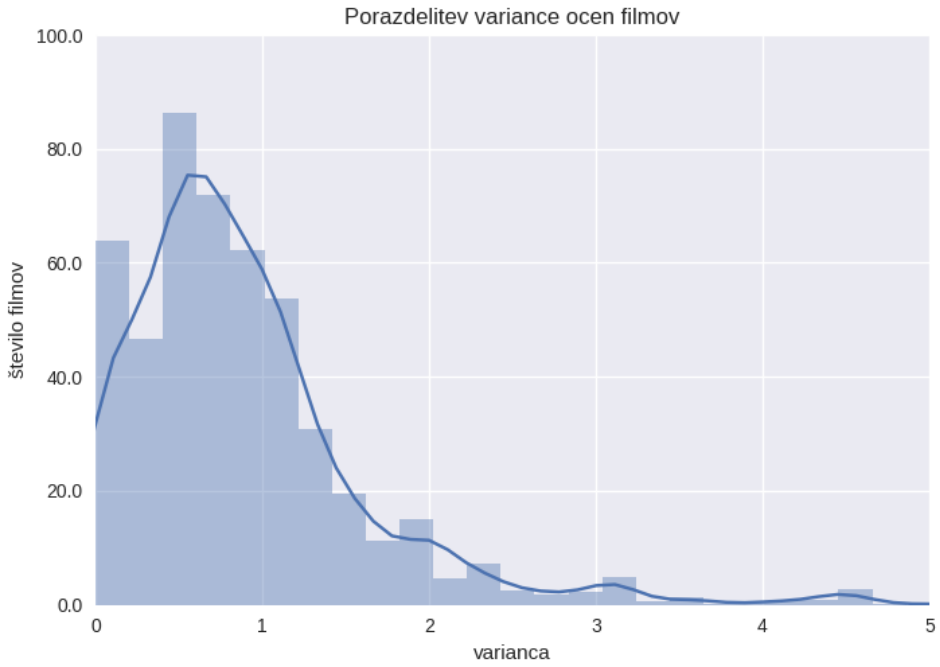
userId / movieId	1	2	3	5	6
15	2.0	2.0	NaN	4.5	4.0
16	NaN	NaN	NaN	NaN	NaN
17	NaN	NaN	NaN	NaN	4.5
18	NaN	NaN	NaN	3.0	4.0
19	3.0	3.0	3.0	NaN	3.0

Za izračun filmov o katerih so si gledalci najmanj enotni sem uporabil varianco. Izračun variance ocen vsakega filma je bil zaradi oblike podatkov in uporabe knjižice pandas precej preprost.

```
df.var()
```

Pri prikazu porazdelitve izračunanih varianc so lepo porazdelitev "pokvarili"1 filmi z malo ocenami. Problem sem rešil tako, da sem izločili filme z manj kot 10 ocenami. Za iskanje osamelcev je s tem ostalo 2245 filmov.

```
df = df.dropna(axis=1, how='any', thresh=10, subset=None, inplace=False)
```



Slika 1: Porazdelitev variance filmov brez izločanja filmov z malo ocenami

3.2 Gručenje filmov

Iskanja skupin med filmi sem se lotil na podoben način kot iskanja osamelcev. Iz podatkov sem najprej sestavil matriko film / uporabnik.

Gručenja sem se najprej lotil z metodo k-Means. Preizkusil sem različne možnosti za izračun razdalje (evklidska, manhatenska, razdalja je 1 - presek uporabnikov, ki so ocenili oba filma deljeno z unijo uporabnikov, ki so ocenili vsaj enega izmed filmov, ...). Najbolje se je izkazala evklidska razdalja, tudi z uporabo le te pa so bili rezultati gručenja še vedno neuporabni.

Zaradi slabih rezultatov z metodo k-means sem za gručenje preizkusil še metodo DBSCAN. Tudi uporaba te metode se ni izkazala za uspešno, vse filme je označila kot šum, ali pa vsak film razvrstila v svojo skupino.

Preizkusil sem še hirearhično gručenje z wardovo metodo. Slika 8 prikazuje dobljeni dendogram. Tudi ta metoda se ni izkazala z uporabnimi rezultati.

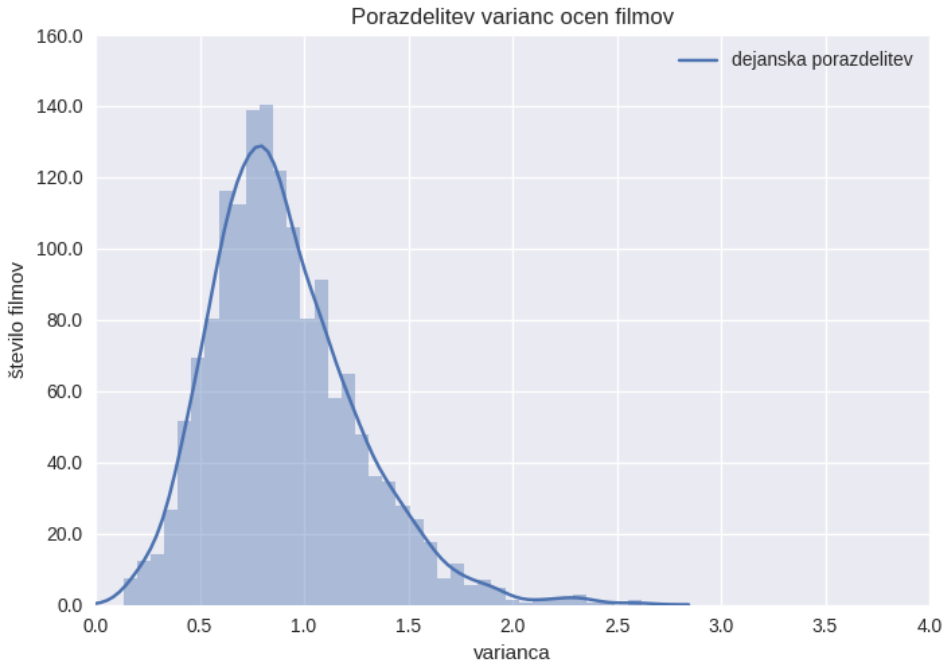
Podatke sem izvozil še v format .xslx in jih uporabil v programu Orange. Tudi tam mi ni uspelo iz danih podatkov poiskati dobrih skupin z uporabo metod k-means in hirearhičnega razvrščanja.

4 Rezultati

4.1 Iskanje osamelcev

Na vprašanje kateri filmi imajo najbolj razpršene ocene odgovori varianca ocen posameznega filma.

Graf2 prikazuje kako so porazdeljene variance ocen vseh filmov z več kot 10 ocenami.



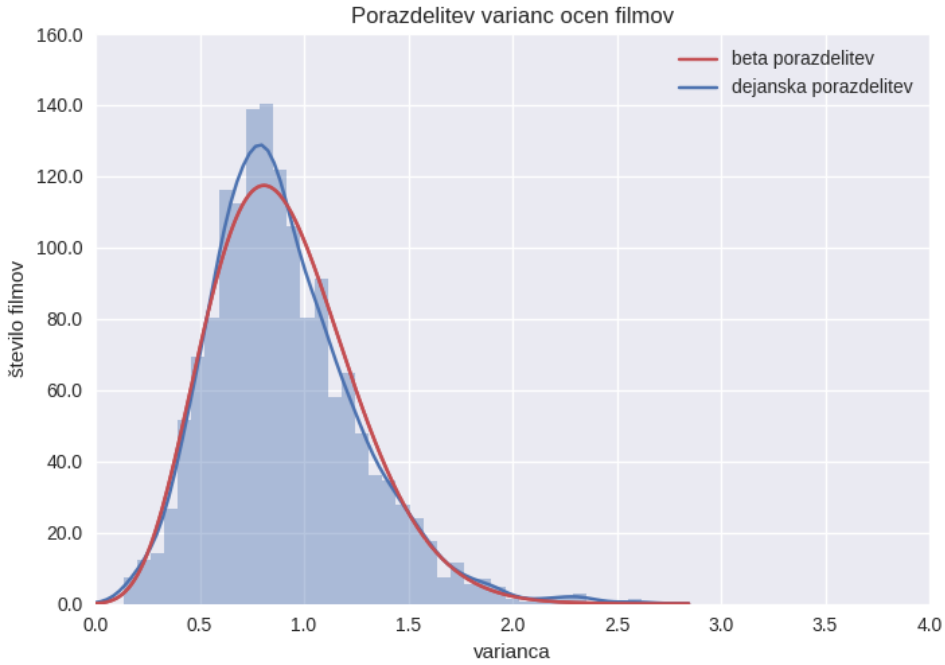
Slika 2: Porazdelitev varianc ocen filmov

Dobljena porazdelitev precej odstopa od normalne2. Od porazdelitev, ki smo jih spoznali porazdelitvi še najbolje ustreza porazdelitev beta. Izračunani oceni parametrov za beta porazdelitev so:

```
params = stats.beta.fit(x)
print(params)
> (6.90759787872952, 27.9450733031308, -0.11751603359648, 5.1490259125502)
```

Na grafu4, ki ima dorisano še krivuljo beta porazdelitve lahko primerjamo ujemanje beta porazdelitve z dejansko.

Preizkusil sem še nekaj ostalih porazdelitev. Za najboljšo se je izkazala Noncentral t-distribution7.



Slika 3: Primerjava krivulje porazdelitve z Beta

V zgornjih 5% statistično najbolj značilnih filmov spada 133 filmov od 2245. Tabela2 prikazuje 10 filmov, ki so med vsemi najbolj izstopali.

4.2 Gručenje filmov

Za gručenje filmov sem izbral metode k-means, dbscan in hirearhično razvrščanje. Za mero podobnosti sem uporabil evklidsko in manhatensko razdaljo ter wardovo metodo. Poleg naštetih sem podobnost meril tudi z razmerjem med številom uporabnikov ki so ocenili oba filma in unijo uporabnikov, ki so ocenili vsaj enega izmed filmov.

Izbran nabor algoritmov za gručenje in mer podobnosti težko utemeljim, saj so se vsi izkazali za neuporabne za dani problem.

Med izbranimi filmi je 53 skupin, če privzamemo, da je vsaka kombinacija žanrov svoja skupina. Če upoštevamo samo prvo napisano kategorijo, pa izbrani filmi spadajo v 11 različnih skupin3. Štirje filmi izmed izbranih filmov nimajo določenega žanra.

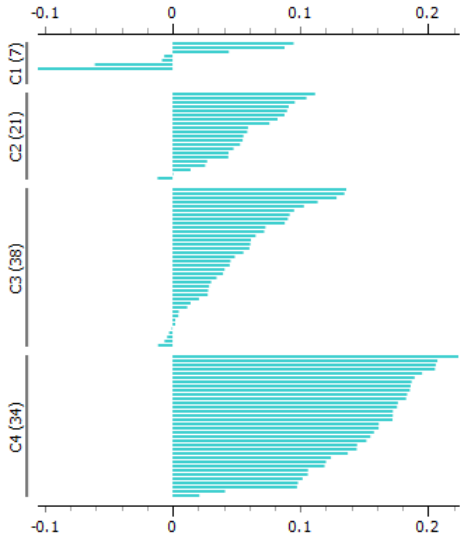
Obstaja veliko metod in ocen za ocenitev kvalitete razvrščanja v gruče kot so naprimer silhouette, dunn's index, Krzanowski-Lai index, R-square,..

Za validacijo gručenja sem uporabil silhouette score, ki omogoča še vizualizacijo ocene. Ocenil sem dobljene gruče, ki sem jih dobil po uporabi metode k-means4 in hirearhičnega razvrščanja5.

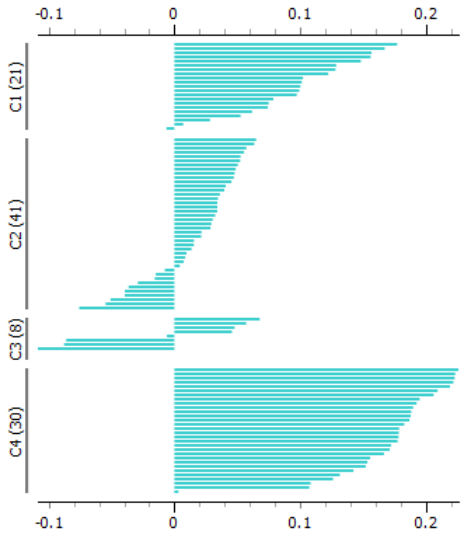
Dobljeni rezultati so zelo slabi. Gruče delujejo zelo naključno v primerjavi s pravimi žanri. Tudi dobljena ocena silhouette je bila v obeh primerih zelo nizka, pri hirearhičnem razvrščanju

Tabela 2: 10 najbolj statistično značilnih filmov

naslov filma	varianca
Killing Zoe (1994)	2.620
Dead Man (1995)	2.558
Christmas Carol, A (1938)	2.428
Cook the Thief His Wife & Her Lover, The (1989)	2.372
Stalker (1979)	2.323
Event Horizon (1997)	2.322
Great Expectations (1998)	2.317
Dungeons & Dragons (2000)	2.316
The Hunger Games: Mockingjay - Part 1 (2014)	2.281
Deadpool (2016)	2.273



Slika 4: K-means silhuette



Slika 5: silhuette score za hirearhično razvrščanje

ima veliko filmov tudi negativno silhuette oceno, kar pomeni, da niso podobni ostalim filmom v svojem razredu.

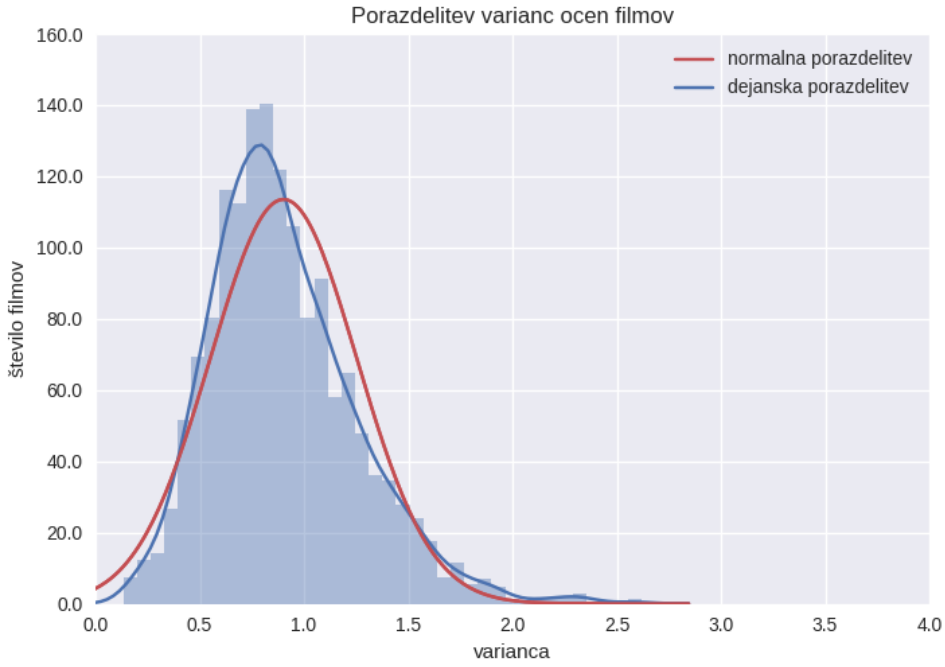
5 Izjava o izdelavi domače naloge

Domačo nalogo in pripadajoče programe sem izdelal sam.

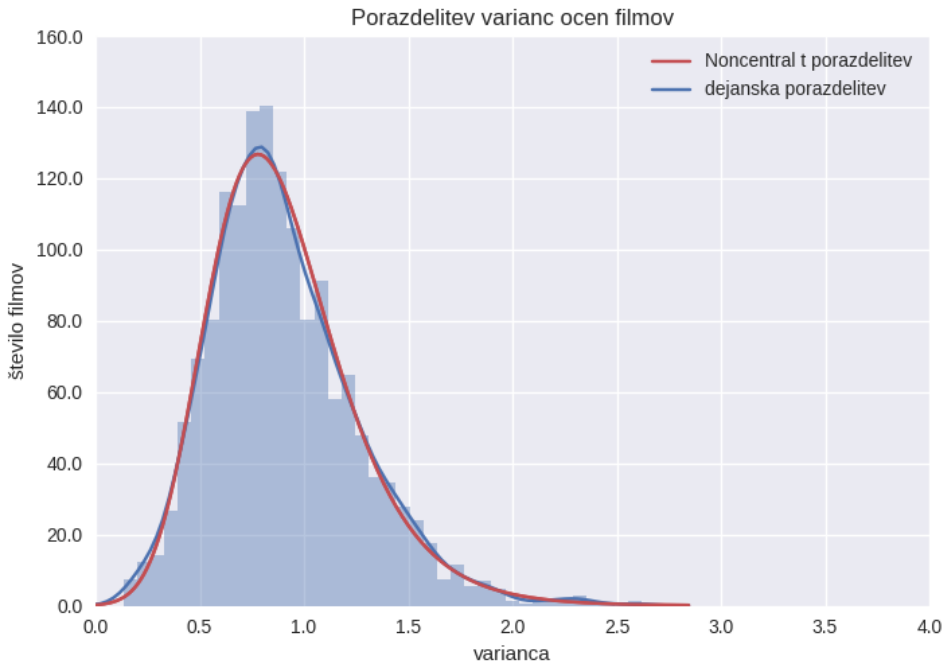
Priloge

A Podrobni rezultati poskusov

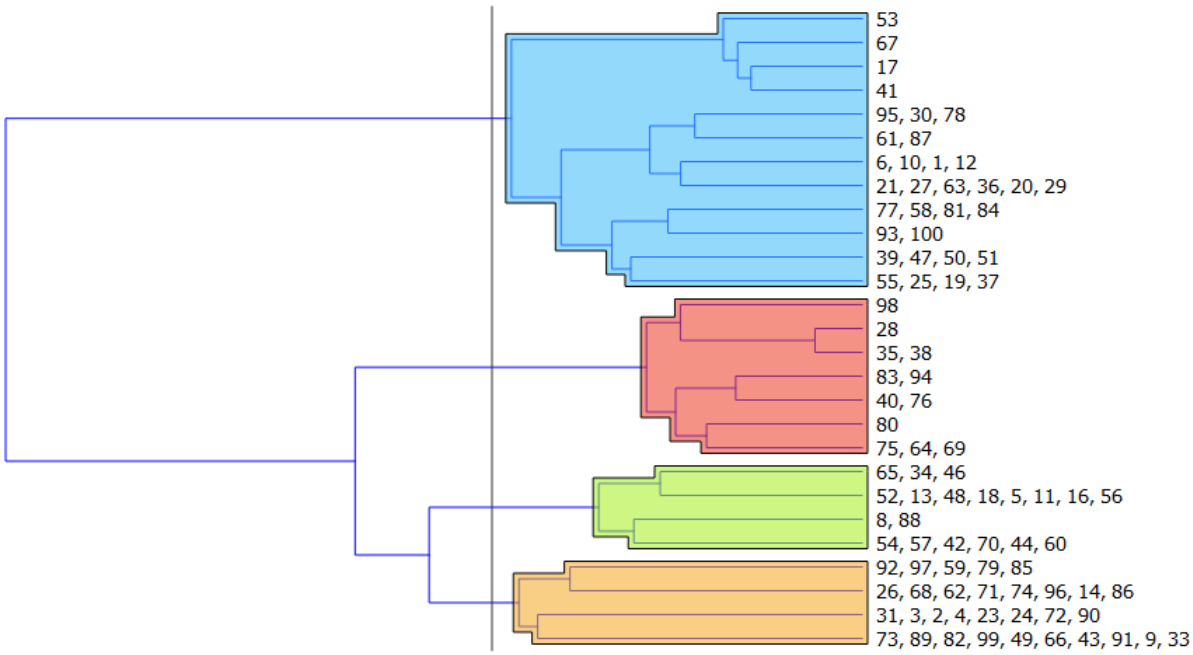
B Programska koda



Slika 6: Primerjava dejanske krivulje porazdelitve in krivulje normalne porazdelitve



Slika 7: Primerjava krivulje porazdelitve z Noncentral t-distribution



Slika 8: hirearhično razvrščanje

Tabela 3: Število filmov v posamezni kategoriji

žanr	število filmov
Comedy	28
Drama	22
Action	18
Crime	12
Horror	6
Adventure	5
Animation	1
Documentary	1
Mystery	1
Romance	1
Thriller	1