Klasterovanje

Tim - Mašinski_učenjaci

SW-03/2016 Nikola Zubić,
SW-11/2016 Filip Mladenović,
SW-53/2016 Mihajlo Kušljić

Zadatak

Klasterovati države na osnovu njihovih karakteristika u klastere koji predstavljaju geografske regione (*region*): ***Africa***, ***Americas***, ***Asia*** i ***Europe***. Zadatak je uspešno urađen ukoliko se na kompletnom testnom skupu podataka dobije *v* mera (eng. *v measure score*) veća od **0.40**.

Analiza podataka

Opis skupa:

- ***region*** – geografski regioni (kolona koju je potrebno prediktovati):
  - *Africa*
  - *Americas*
  - *Asia*
  - *Europe*
- ***income* –** prihod po glavi stanovnika u dolarima
- ***infant* –** smrtnost odojčadi na 1000 živorođenih
- ***oil*** – da li je država izvoznik nafte:
  - *yes* – da
  - *no* – ne

Trening skup sadrži 84 zapisa.

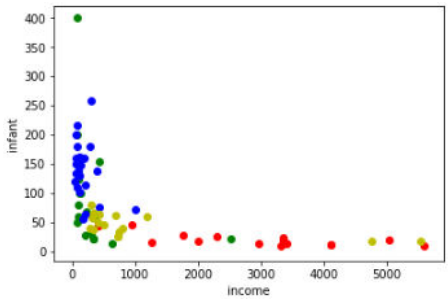
Trening skup sadrži nedostajuće vrednosti u koloni *infant* - zbog malog skupa podataka odlučili smo se da nedostajuće vrednosti popunimo srednjom vrednošću (engl. *mean*), jer je dala najbolje rezultate prilikom validacije.

Problem je male dimenzionalnosti pa su sva obeležija uzeta u razmatranje pri treniranju i prediktovanju klastera.

Za kodovanje kategoričkih obeležja je korišćen *LabelEncoder*.

Vrednosti su normalizovane *StandardScaler*-om.

Budući da je problem klasterovanje, pokušali smo da saznamo koji bi to oblik klastera bio najkompatibilniji našem problemu. Takođe, uspešnosti klasterovanja doprinosi i što bolje određivanje njihovih inicijalnih centara i težina.



Napomena: Na slici se ne može uvideti da li je konkretna zemlja izvoznik nafte, ali pošto je to binarna vrednost to bi klasteru dalo samo, da kažemo, loptasti oblik po z osi.

Odabrani model

Korišćen je *GausianMixture* model sa četiri komponente (zato što postoje četiri regiona koja je potrebno prediktovati). Tačnost modela je određivana korišćenjem *Repeated Stratified K Fold* unakrsne validacije sa parametrima: n_splits = 4, n_repeats = 2 i random_state = 12345 (videli smo da se često koristi ovakav *random_state* na *sklearn*-ovom zvaničnom sajtu sa primerima). Tokom eksperimentisanja povećavan je i broj komponenti – zaključak je da je značajno bolji rezultat dobijen kada se koristi klasifikator sa devet komponenti. Ispitujući težine takvog klasifikatora zaključeno je da se formira suptilnija granica između klastera (**sa više sigurnosti možemo reći da li nekom regionu pripada neki podatak**) prilikom odlučivanja. Od devet težina, postoje dve koje dominiraju sa uticajem od oko 39.25% dok su ostale težine imale znatno manje vrednosti. To se može uvideti i na priloženoj slici jer bismo mogli jasnije da odredimo crveni i plavi klaster dok je za žuti i zeleni potrebno preciznije biranje. Odlučeno je da se takvi odnosi težina iskoriste za inicijalizaciju modela sa četiri komponente gde se jednom klasteru postavi težina koja je jednaka zbiru dve dominantne težine (od 78.5%), dok su ostale bliske nuli (čime je očuvan odnos između težina kao i sa devet klastera - jedna dominantna težina na četiri komponente, naspram dve dominante težine od oko 39.25% u slučaju sa devet komponenti). Kasnije, koristeći *GridSearchCV* i dodatnim ručnim podešavanjem su nad tim modelom varirani ostali parametri dok nije dobijen najbolji rezultat na validaciji.

Uz postavljanje inicijalnih težina, ključnu ulogu je odigrao i broj iteracija koji je postavljen na manji broj zato što u svim iteracijama nakon te model iskonvergira u neki drugi lokalni minimum u kome ostaje zaglavljen, a koji je lošiji od prethodnog.

Dobijen je *v-measure score* koji na kompletnom test skupu iznosi: **0.699675259419155**.