Klasterovanje

Tim:

- **Matija Aleksić sw55-2016**

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. Zadatak se rešava upotrebom Modela Gausovih mešavina (eng. Gaussian Mixutre Model, GMM), tj. algoritmom Očekivanje - maksimizacija (eng. Expectation - maximization, EM). Da bi zadatak bio uspesno rešen treba da se ispuni sledeci uslov: **v measure > 0.4**

Analiza podataka:

Opis skupa:

- **Region** – geografksi 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 podataka se sastoji od 84 elementa.
Postoje dva kategoricna obeležja(Region, Oil) koja smo enkodirali ručno sa vrednostima:

- 'Africa' : 1
- 'Asia' : 2
- 'Americas' : 3
- 'Europe' : 4

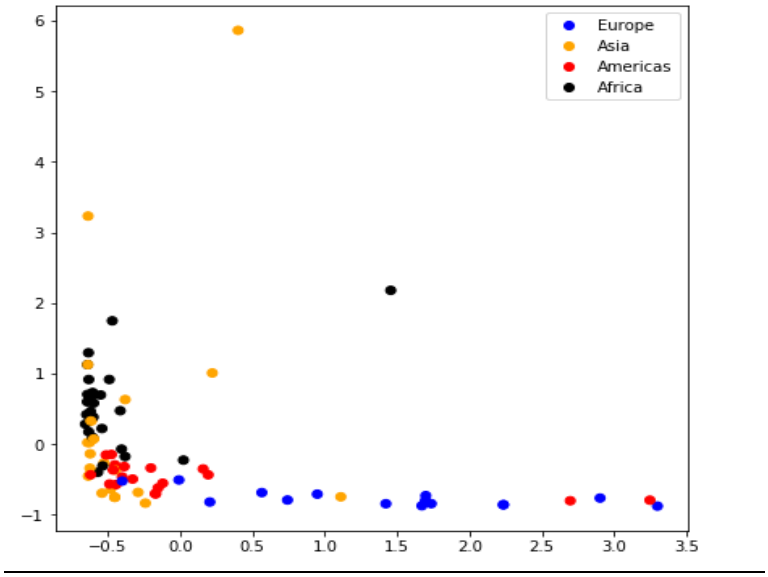
a za ulje:

- ’yes’ : 1
- ’no’ : 2

Takođe iz ova 84 elementa u koloni infant imaju neke vrednosti koje su NaN. Da bi resšio ovaj problem napravili smo pomoćnu metodu za LinearnuRegresiju svih ostalih članova koje nemaju NaN, i obučavanjem nad njima predvideli smo vrednosti nedostajućih elemenata koje su pre ovoga bili NaN.

Dve kolone infant i income imaju veliki raspon brojeva koje je potrebno skalirati da bi klasifikacija bolje radila. To sam uradio ručno sa posebnom metodom koja radi po formuli:

element = (element – element_mean) / element_std



Postojali su neki outlieri ali posto ukupno podataka za treniranje je bilo 84, smatrao sam da ne bi trebao da ih izbacujem jel model ne bi imao dovoljno podataka za treniranje.

Odabrani model:

Koriscen je GausianMixture model sa četiri komponente. Broj komponenti je izabran tako da mora da prepozna četiri kategorije iz kojeg kontinenta dolazi predikovana zemlja(Americas, Africa, Asia, Europe). Igranjem sa hiperparametrima ovoga modela dosli smo do zakljucka da najbolje radi sa tipom kovarijanse(covariance_type='diag'), odnosno da svaka komponenta ima svoju diagonalnu matricu kovariansi. Koriscenje ovakvog modela ostvareni rezutat je **v_measure: 0.484375429313693.**