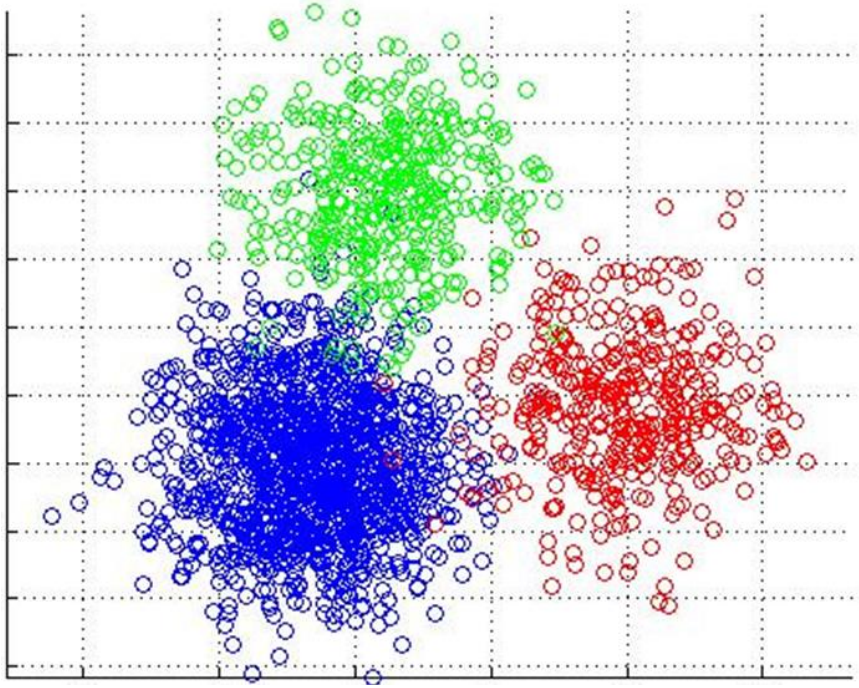
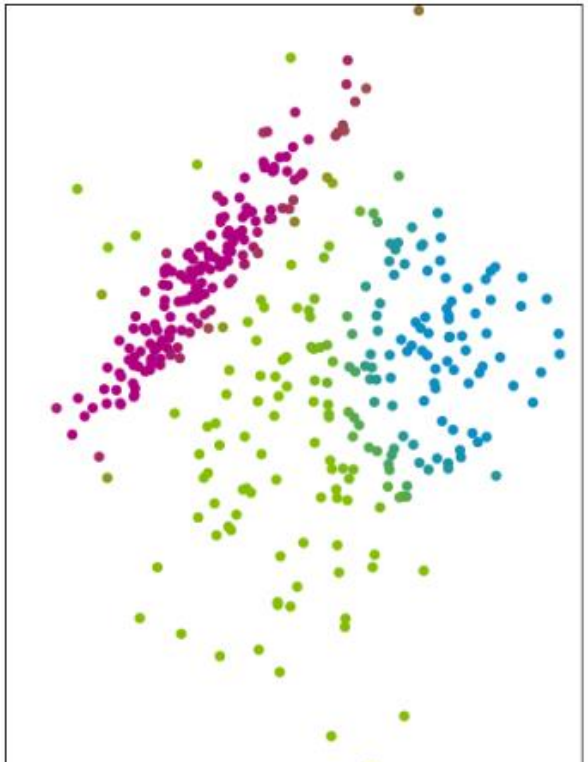


K-means



Gaussian Mixture Model

Klasterovanje

Gaussian Mixture Model

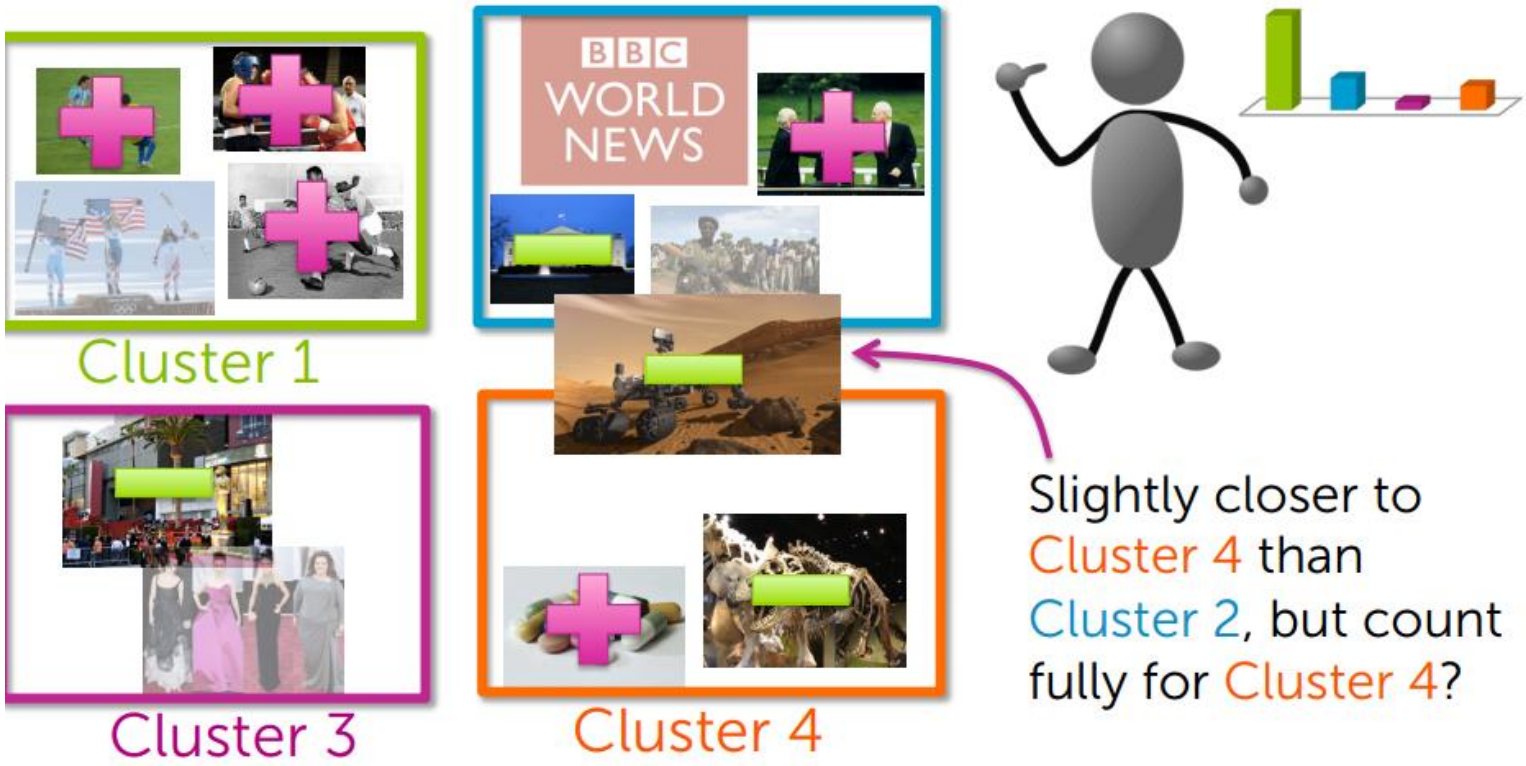
Modeli mešavina

- Upoznali smo se sa algoritamskim pristupom klasterovanju (*K-means*)

- Modeli mešavina: probabilistički pristup klasterovanju (klasterovanje zasnovano na modelu)

Motivacioni primer

- Učenje preferenci korisnika nad skupom tema
- Klasterovaćemo dokumente u neke skupove tema i onda koristiti korisnički *feedback* za članke koje je pročitao



Ograničenja *K-means*

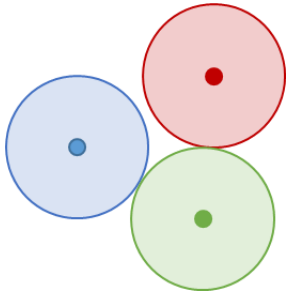
1. „Čvrsta“ (isključiva) dodela objekta klasteru, bez kvantifikacije nesigurnosti

Ograničenja *K-means*

2. Pretpostavlja sferično simetrične klastere
   - *K-means* dodeljuje objekat klasteru sa najbližim centroidom:

$$z_i \leftarrow \arg\min_j \left\| \mu_j - x_i \right\|_2^2$$

Samo centroid figuriše, ništa o obliku klastera nije uzeto u obzir

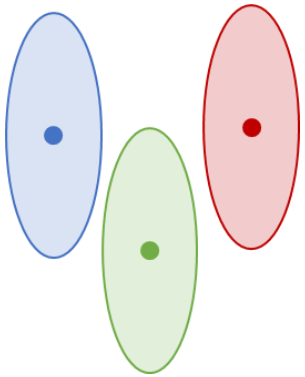


Možemo ga modifikovati da koristi otežinjenu Euklidsku udaljenost, ali…

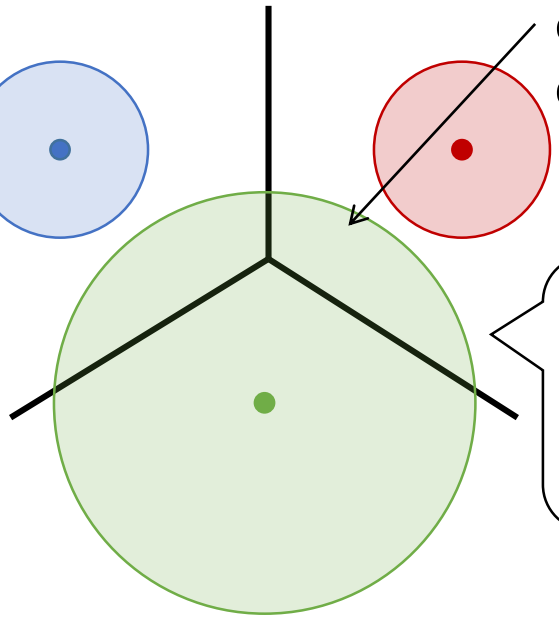
Klasteri su isključivo elipsoidni

Elipse su isključivo paralelne sa osama

Kako da odredimo težine?



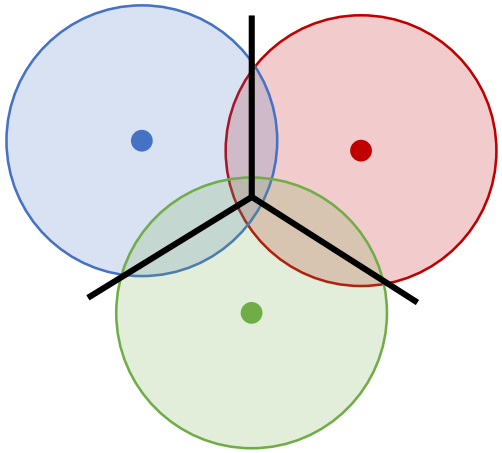
Slučajevi u kojima *k*-means ne uspeva

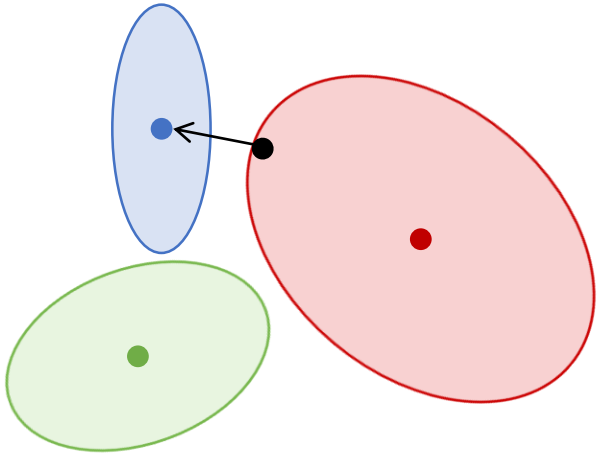


Voronoi: sve tačke unutar ove regije su najbliže crvenom centroidu

Klasteri sa različitom „rasutošću“

Klasteri koji se preklapaju





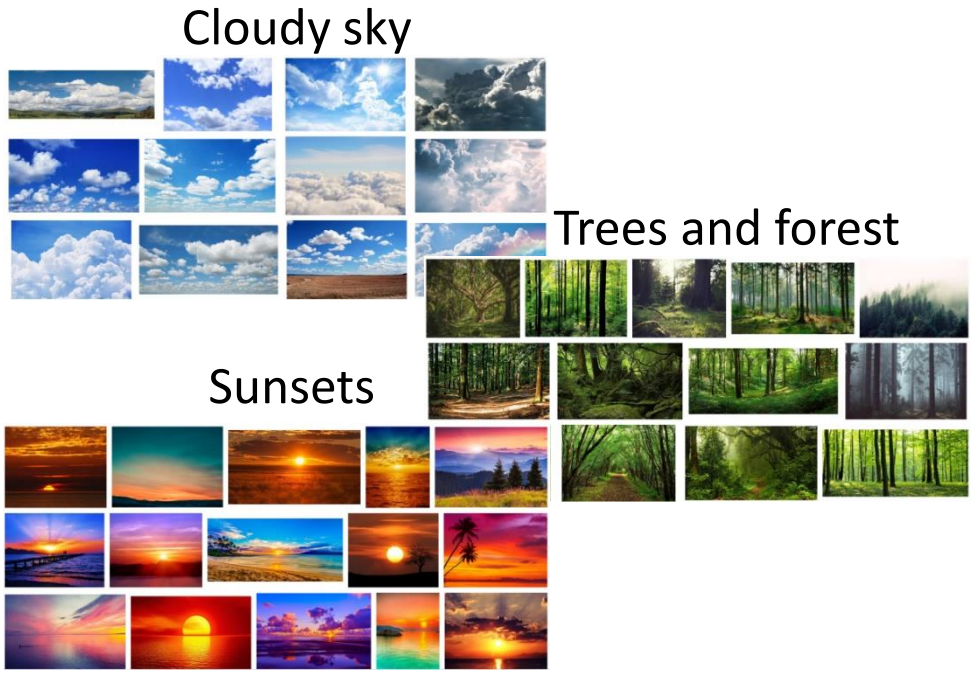
Klasteri različitih oblika i orijentacija

Model mešavina

- Ova ograničenja motivišu probabilistički model: model mešavina (*mixture model*)

1. Model mešavina nam omogućava „meku“ dodelu objekata klasterima
   - Npr. „svetske vesti“ 54%, „nauka“ 45%, „sport“ 1% i „zabava“ 0%
2. Pored centra klastera uzima u obzir i njegov oblik
3. Omogućava da učimo „težine“ različitih dimenzija
   - Npr. prilikom klasterovanja dokumenata – kakve težine da dodelimo svakoj reči iz rečnika (prilikom računanja sličnosti dokumenata)

Ilustrativni primer: klasterovanje slika



Reprezentacija: prosečna vrednost intenziteta crvenih, zelenih i plavih piksela [R, G, B]

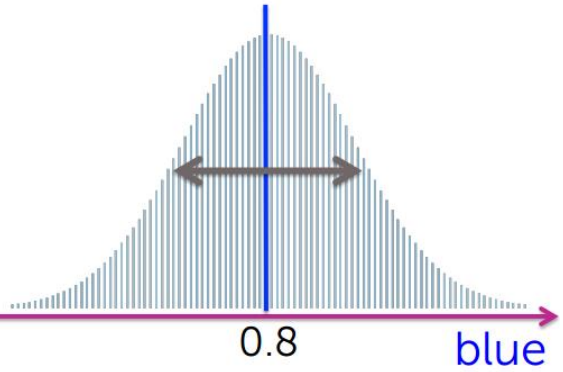

[R = 0.85, G = 0.05, B = 0.35]

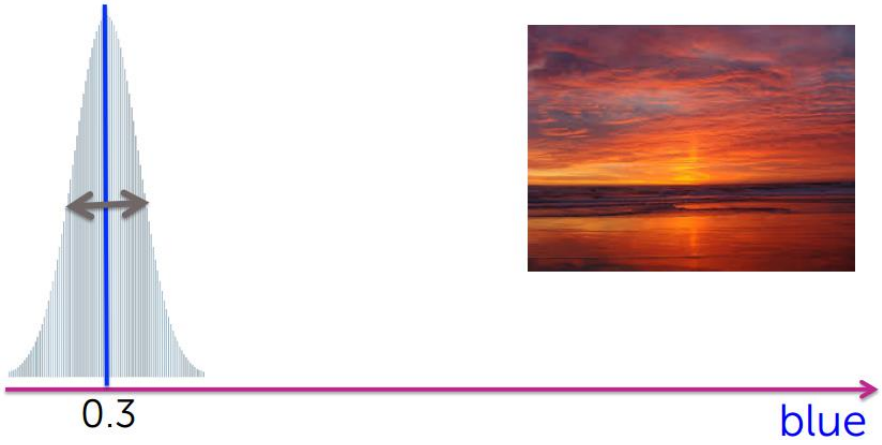

[R = 0.02, G = 0.95, B = 0.4]


[R = 0.05, G = 0.7, B = 0.9]

Histogram vrednosti plave boje

Ako bismo imali labele, mogli bismo se uveriti da je ova reprezentacija adekvatna

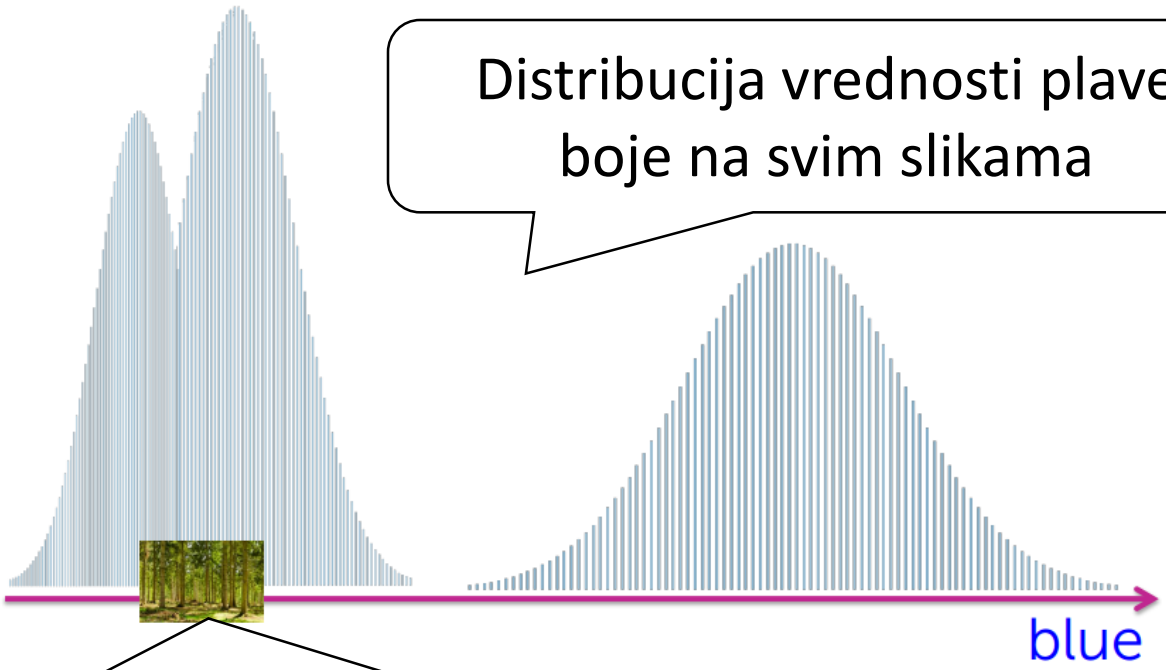




Histogram vrednosti plave boje

Ali mi nemamo labele...

Distribucija vrednosti plave boje na svim slikama

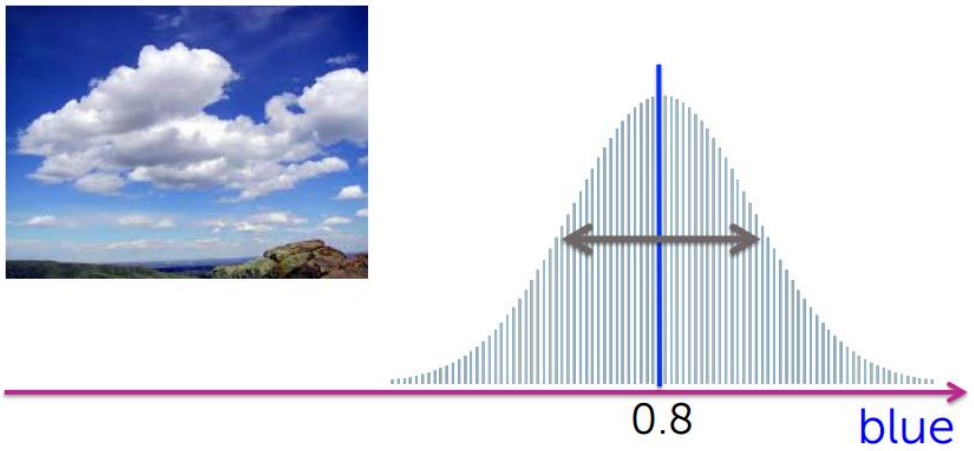


Nije jasno kojoj grupi pripada. Možda ćemo to lakše uočiti posmatrajući drugu dimenziju (npr. crvenu boju)

Histogram vrednosti crvene boje

- U „crvenoj“ dimenziji značajno lakše možemo razdvojiti kategorije „šume“ i „zalaska sunca“
- Prilikom klasterovanja, zakonitosti koje nam mogu pomoći da razdvojimo klastere mogu biti uočljivije ako posmatramo više dimenzija, a ne samo jednu

Fitovanje normalne (Gausove) raspodele



- Za svaku dimenziju [R, G, B]:
  - Intenzitet piksela ćemo tretirati kao slučajnu promenljivu
  - Naša pretpostavka je da su uočene opservacije $x^{(i)}$ izvučene nezavisno jedna od druge iz iste Gausove distribucije $\mathcal{N}\left(x^{(i)}\middle|\mu,\sigma^2\right)$
  - Želimo da na osnovu podataka odredimo parametre distribucije μ (srednju vrednost) i σ^2 (varijansu)