k-Srednjih vrednosti (k-Means)

TEHNIKA KLASTEROVANJA UZORAKA

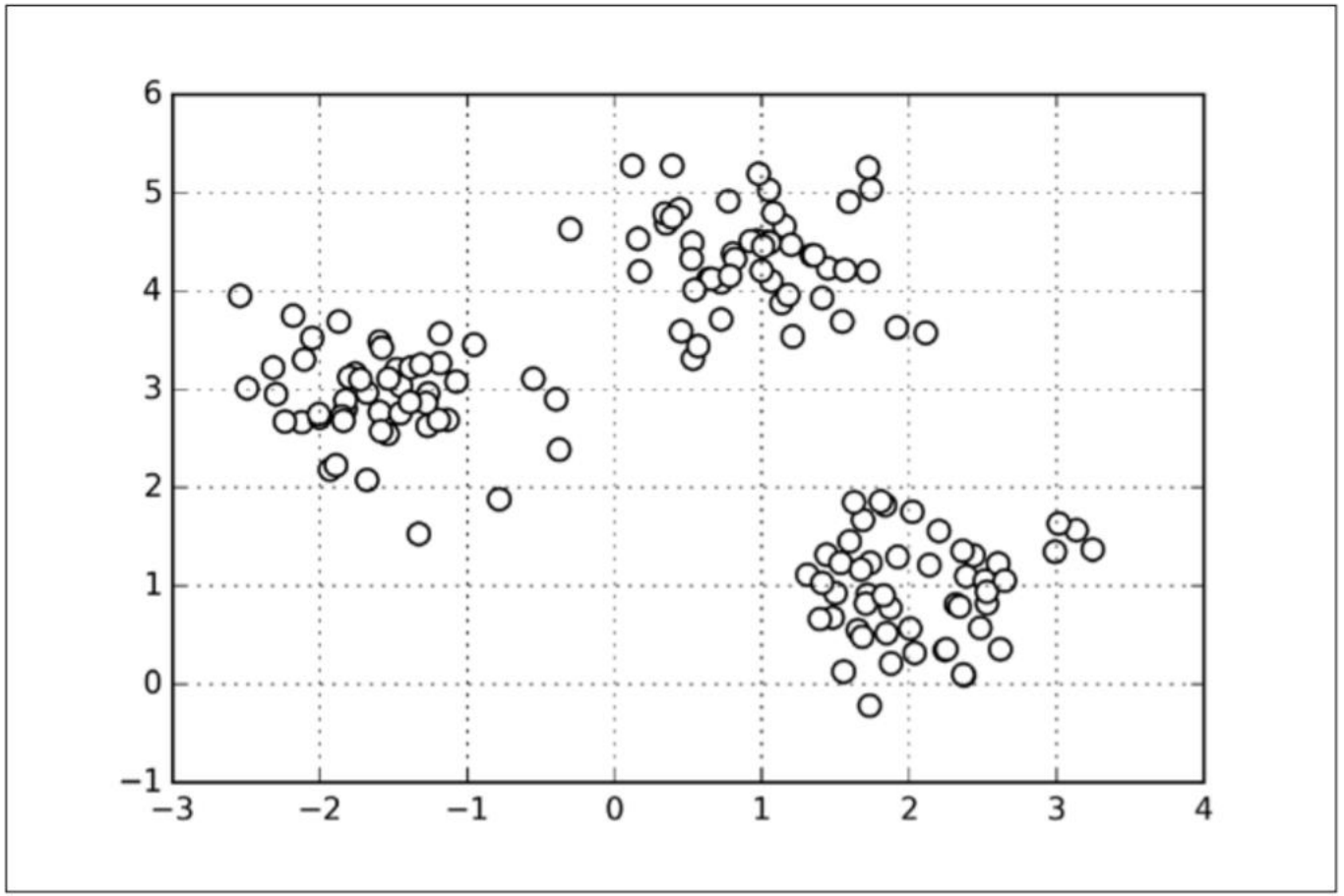
Klasterovanje

- Za razliku od klasifikacije, ovde ne postoji "tačno" rešenje
- Zato je ocena uspešnosti algoritma teže doneti u odnosu na klasifikaciju
- Rezultat rešenja zavisi od domena i slučaja primene – jedno isto rešenje može biti različito ocenjeno u različitim slučajevima primene
- Zahteva angažovanje domenskih eksperata koji će evaluirati rešenje

Tipovi klasterovanja

- Hijerarhijski algoritmi
  - Pronalaze sledeće klastere koristeći prethodno uspostavljene
  - bottom-up – kreću od svakog elementa kao pojedinačnog klastera i spajaju ih sukcesivno u veće
  - top-down – kreću od celog skupa kao jednog klastera i uzastopno ga smanjuju u manje klastere
- Parcionalno klasterovanje – prepoznaju se svi klasteri odjednom
  - K-means

Klasterovanje



Rastojanje

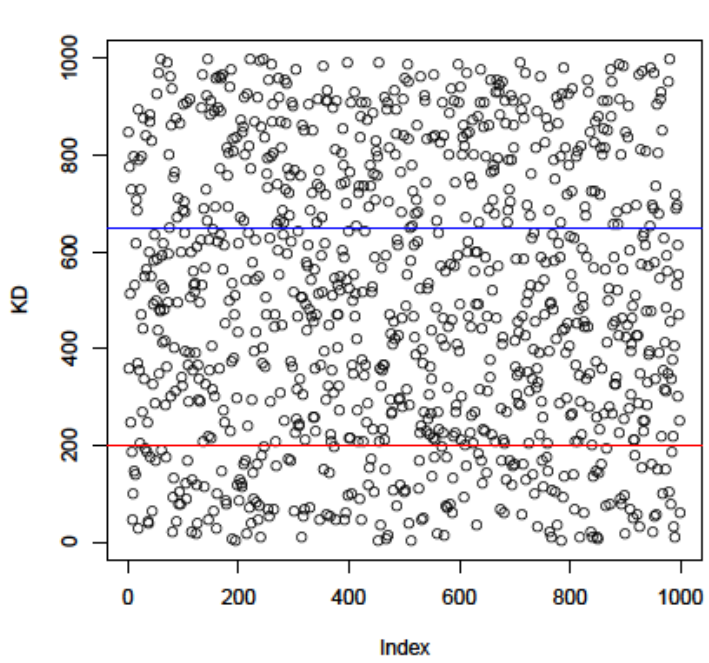
- Mera udaljenosti određuje kako se računa sličnost dva elementa i utiče na oblik klastera
  - Euklidska distanca $d(x,y) = \sum_{i=1}^{p} |x_i - y_i|$
  - Menhetn $d(x,y) = \sqrt[2]{\sum_{i=1}^{p} |x_i - y_i|^2}$
  - Maksimum $d(x,y) = \max_{1 \le i \le p} |x_i - y_i|$
- Ugao između dva vektora može se koristiti kao mera udaljenosti kada postoji više dimenzija podataka
- Minimalan broj zamena da bi se jedan element promenio u drugi se može koristiti kao mera udaljenosti

K-means

- Najpoznatiji i najjednostavniji algoritam klasterovanja (klasterizacije)
- Pripada tehnikama nenadgledanog učenja
- Sličnost instanci se može procenjivati primenom neke od mera za računanje:
  - Sličnosti (npr. kosinusna sličnost ili koeficijent korelacije)
  - Udaljenosti dve instance (npr. Euklidsko ili Menhetn rastojanje)

K-means – Osnovna verzija

- Slučajan izbor K težišta klastera (centroida)

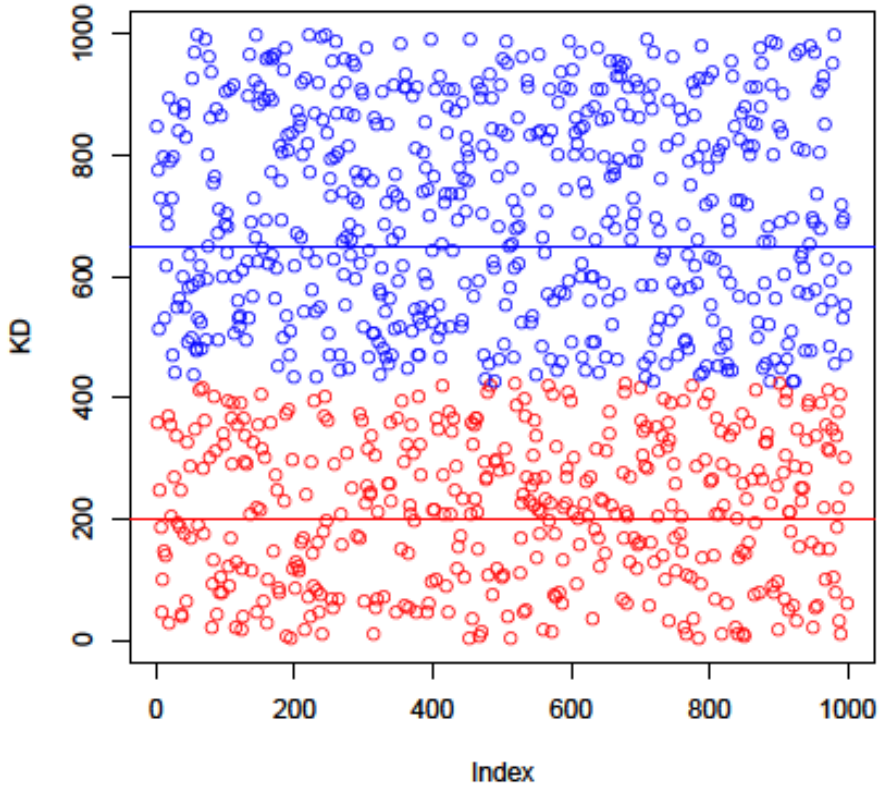


Seed 1
650

Seed 2
200

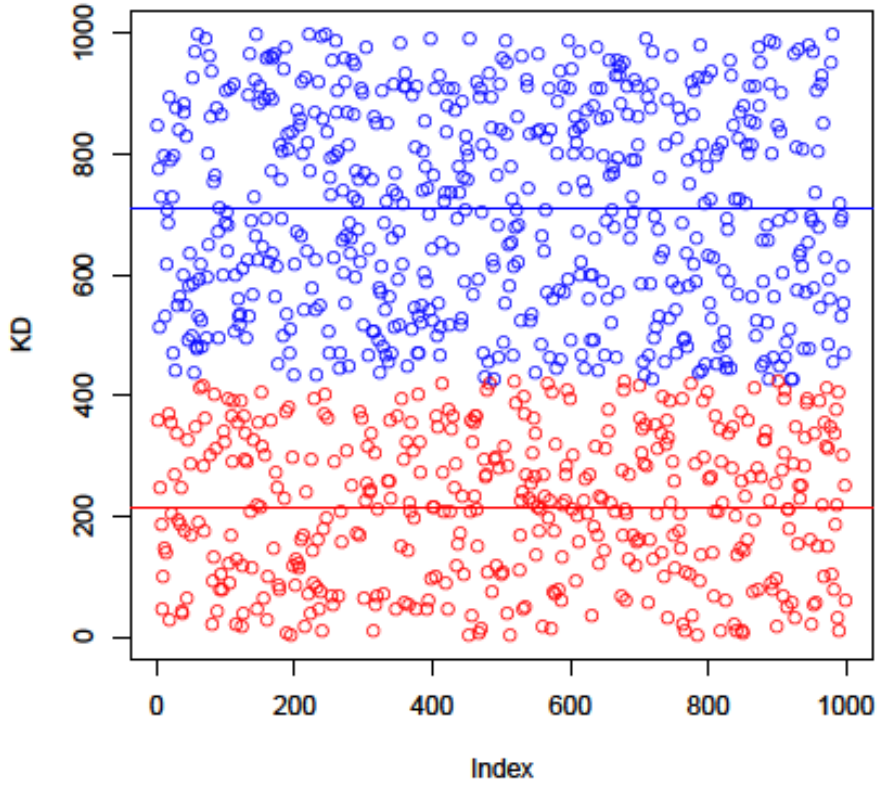
K-means – Osnovna verzija

- Izračunati rastojanje svakog objekta od svakog klastera (Euklidsko rastojanje)
- Pridružiti svaki objekat najbližem klasteru (650, 200)



K-means – Osnovna verzija

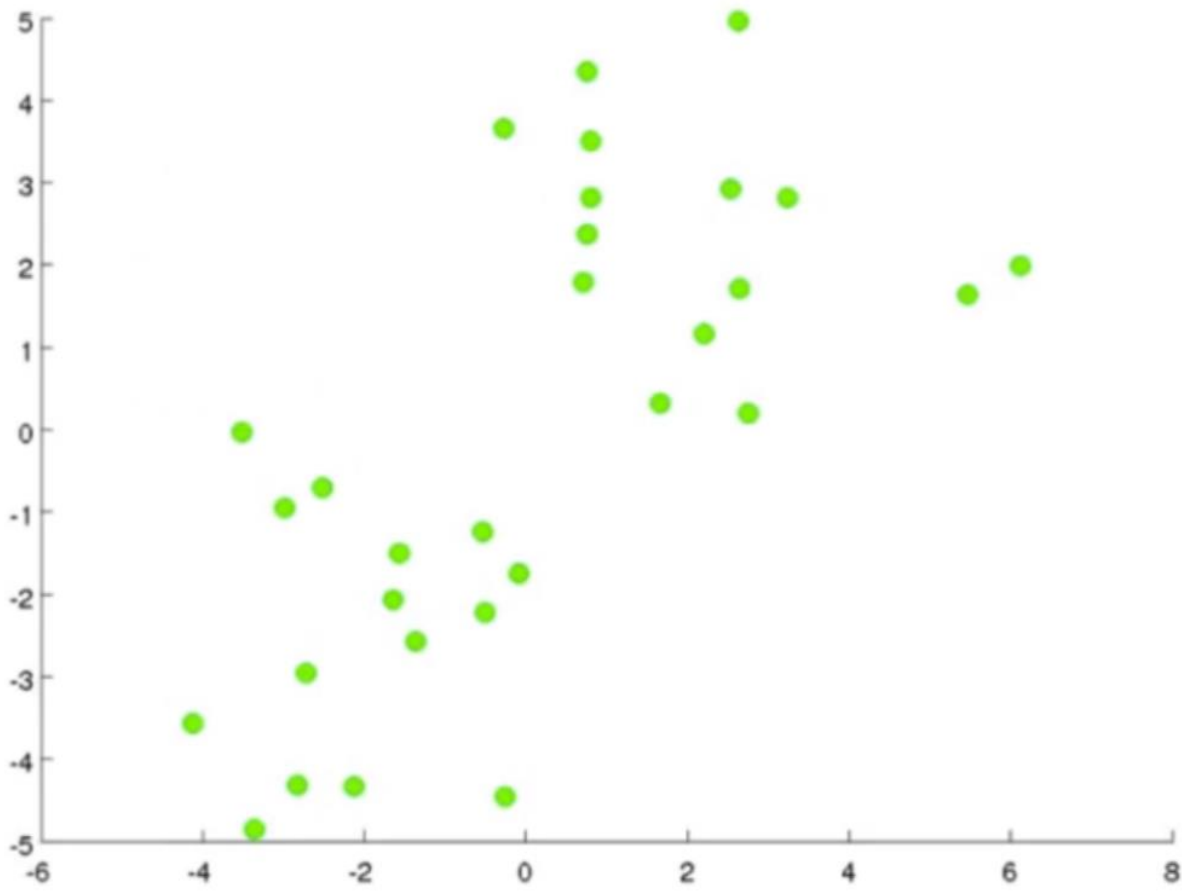
- Izračunati novi centroid za svaki klaster (708.9, 214.2)

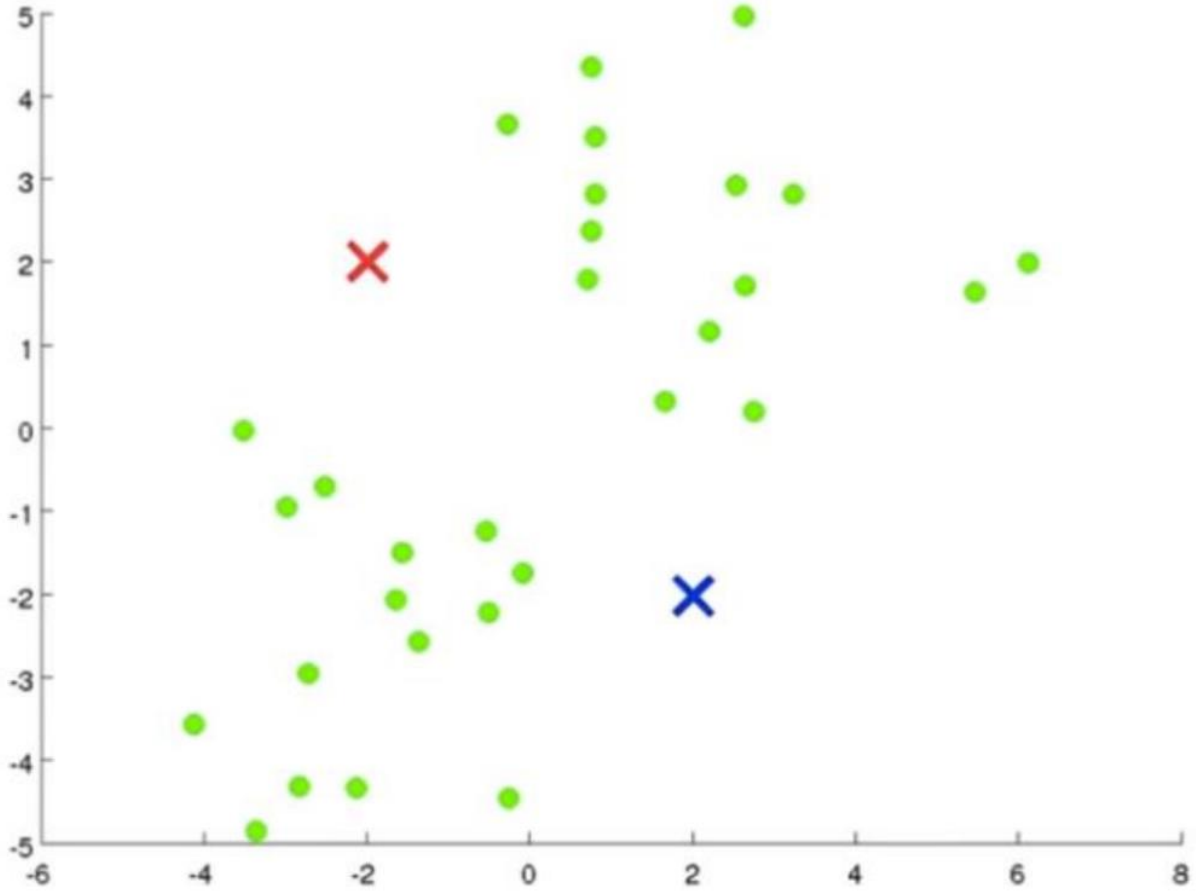


K-means – Osnovna verzija

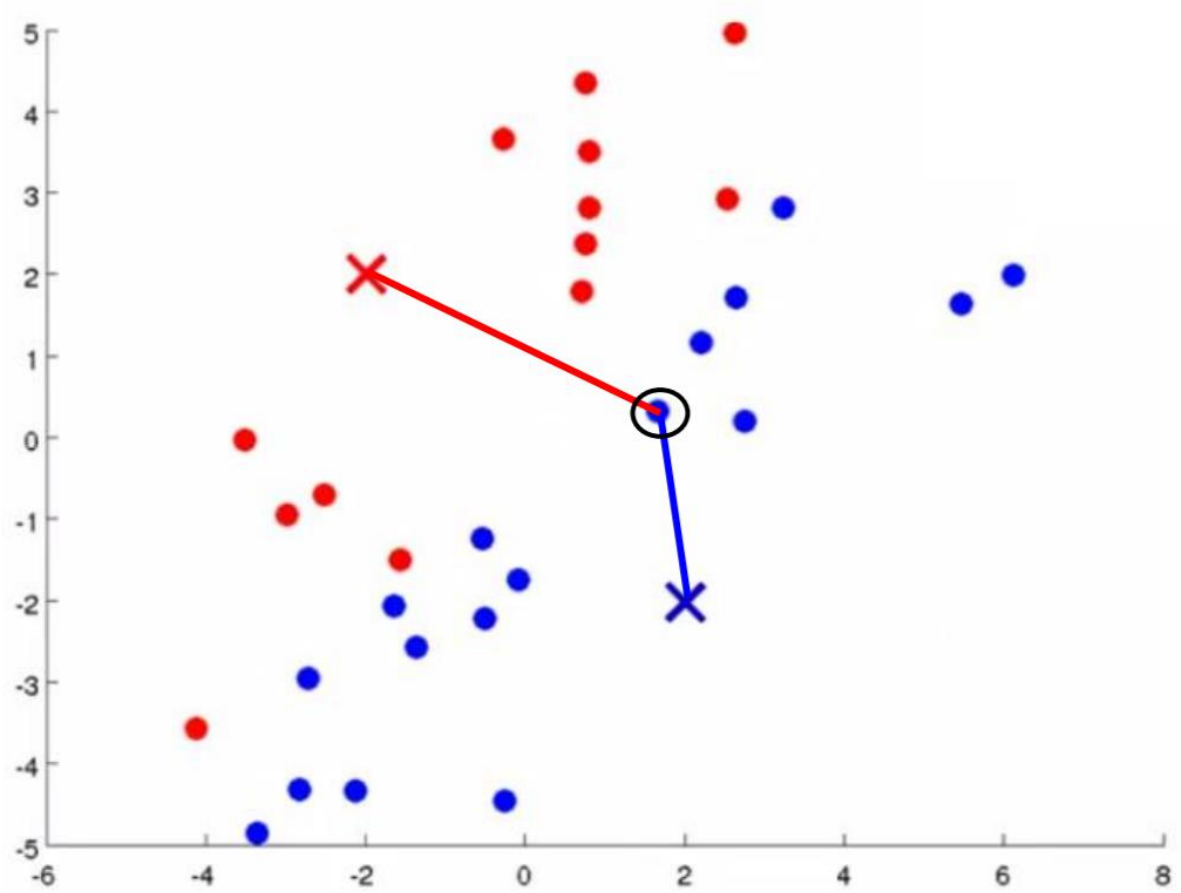
- Nastaviti iteracije:
  - Izračunati rastojanje od objekata do centroida klastera
  - Pridružiti objekte najbližim klasterima
  - Reizračunati nove centroide
- Završetak iteracija se zasniva na kriterijumu konvergencije
  - Nema promene u klasterima
  - Maksimalan broj iteracija

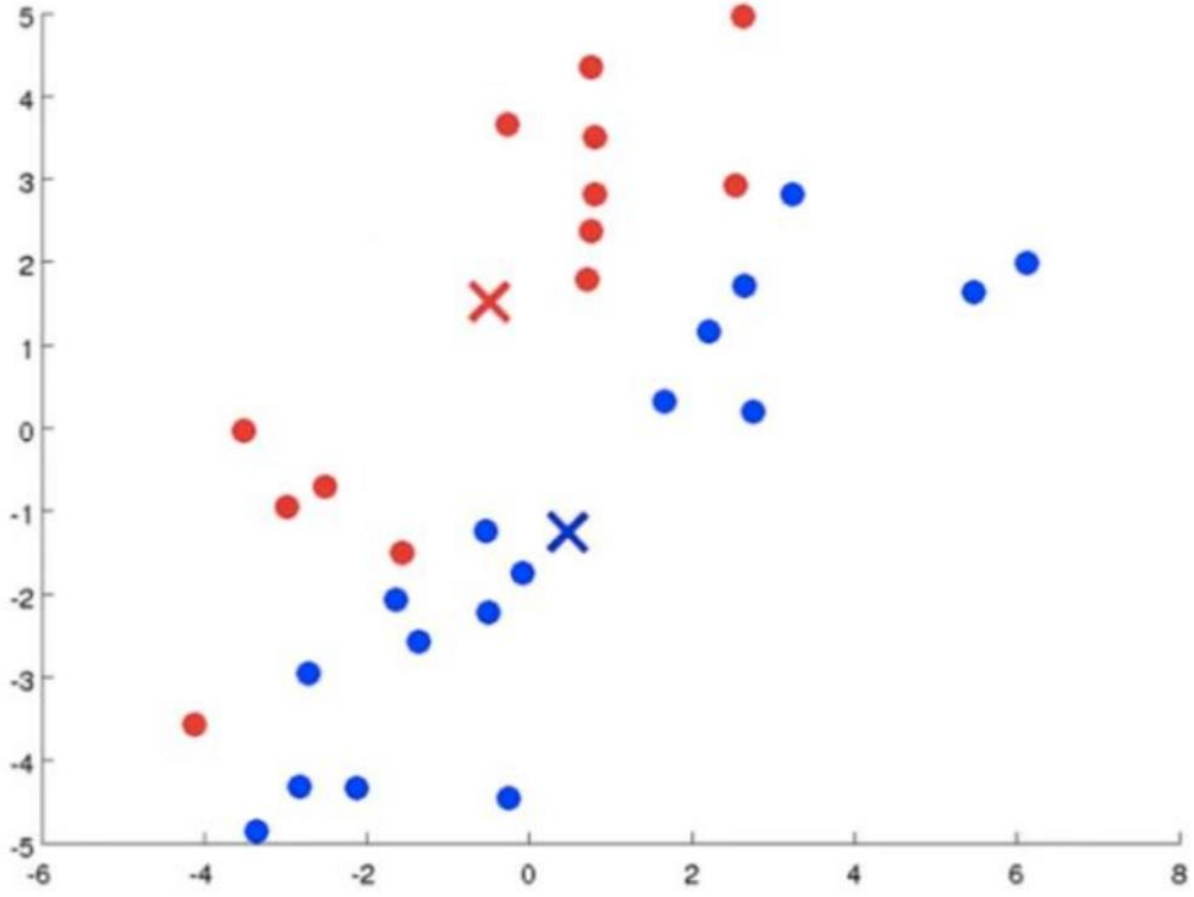
K-means (1) - Inicijalno



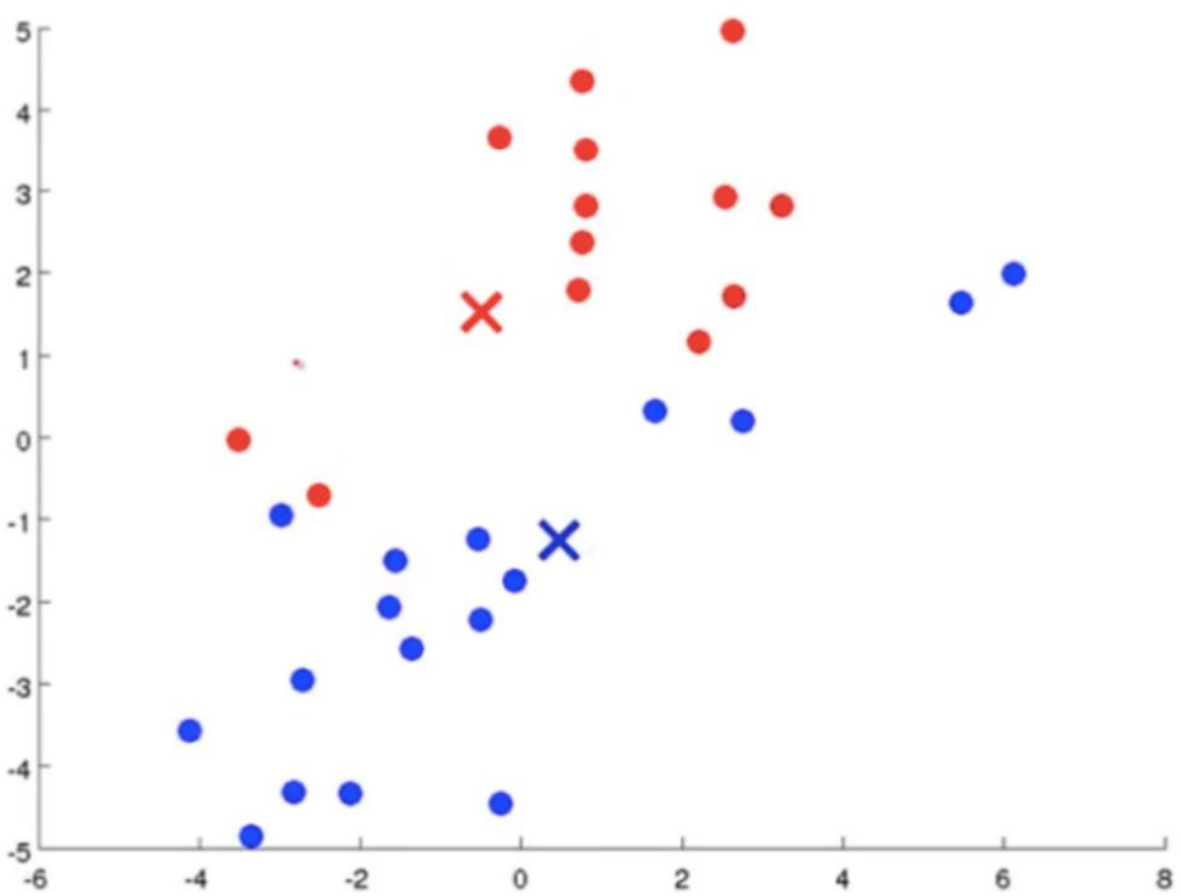


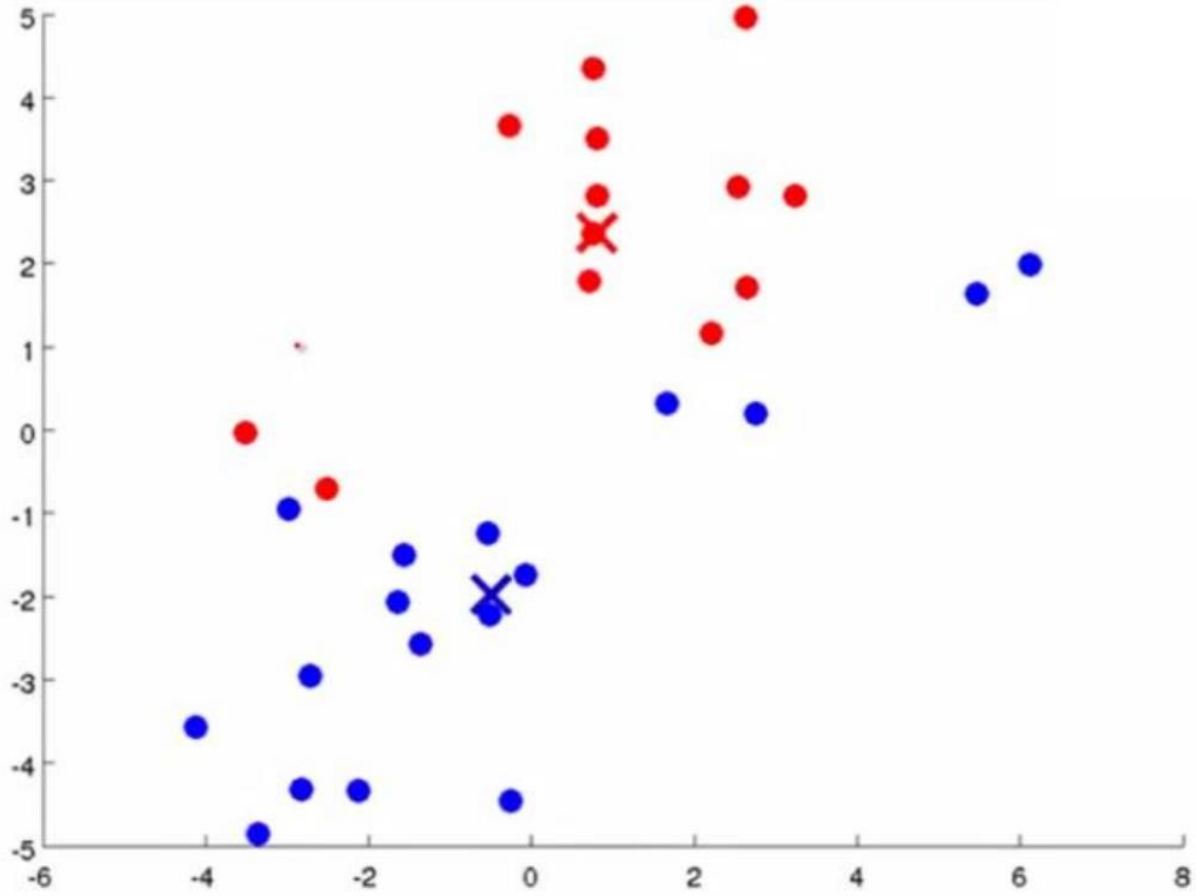
K-means (2) – Iteracija br. 1



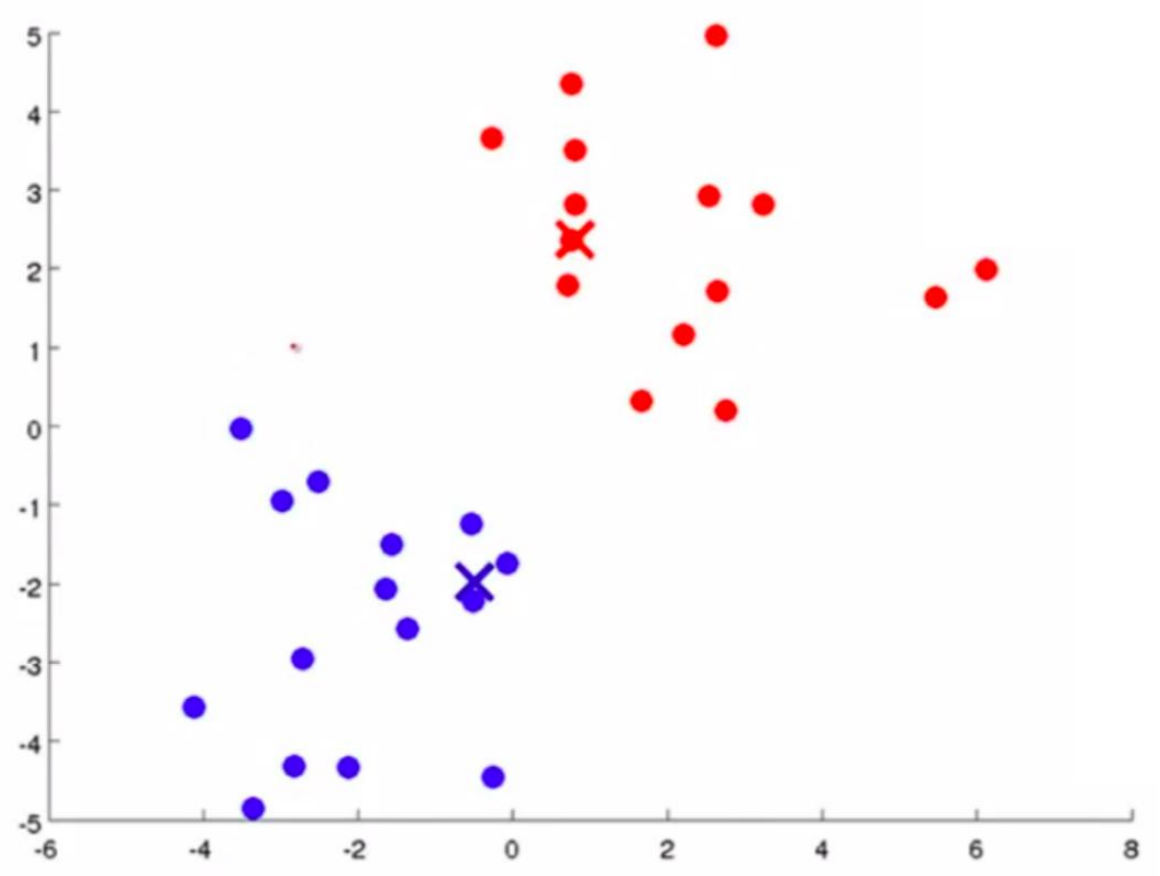


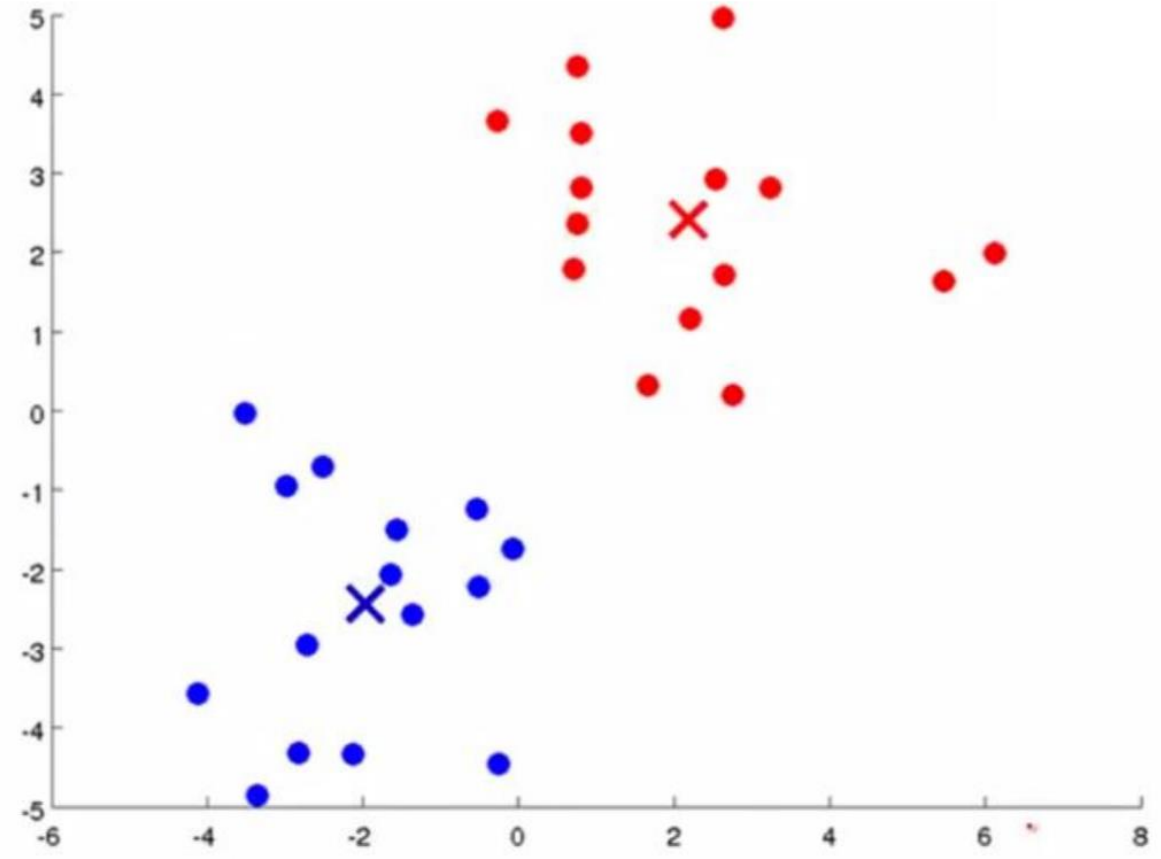
K-means (3) – Iteracija br. 2



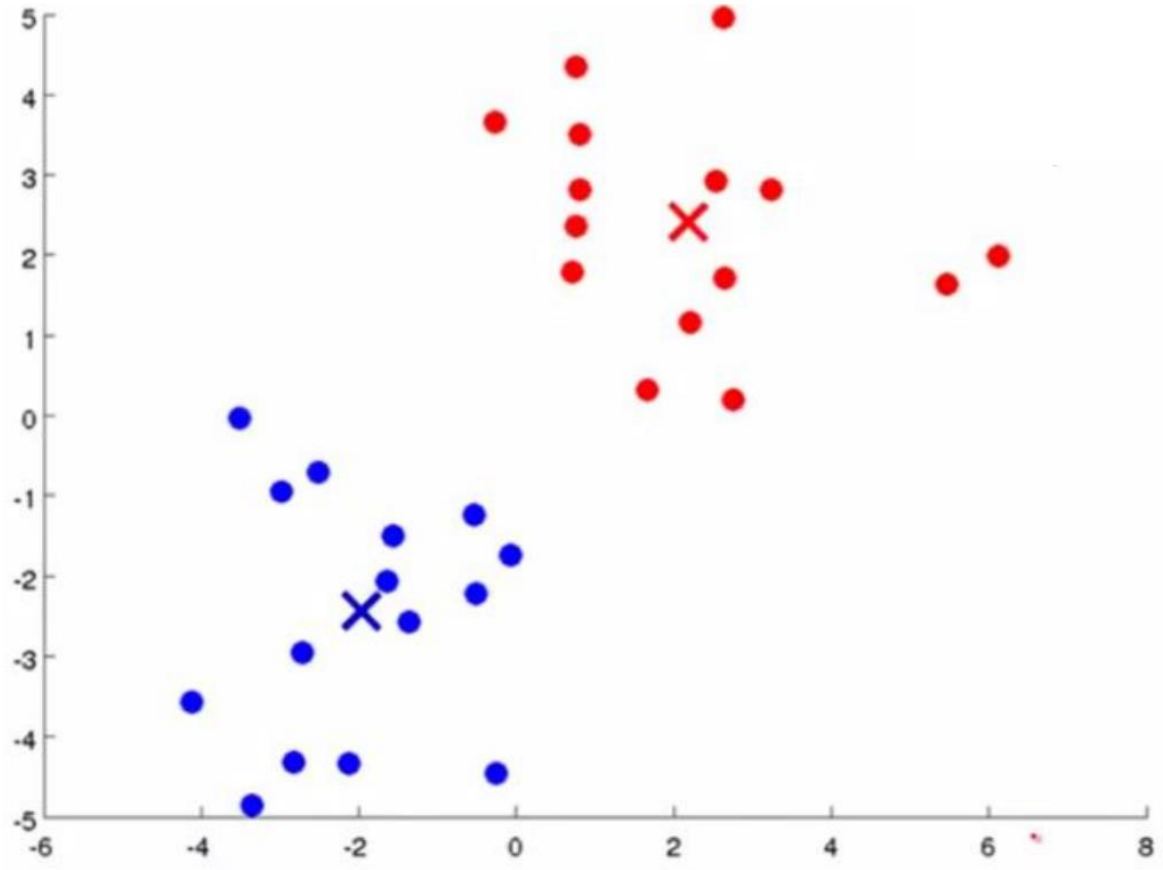


K-means (4) – Iteracija br. 3





K-means (5) - Završetak



K-means – Koraci algoritma

- 1) Slučajan odabir K težišta klastera (centroida)
- 2) Grupisanje po klasterima - svaki uzorak (instancu) dodeljujemo najbližem centroidu (najbliže težište na osnovu odabrane metrike)
- 3) Pomeranje težišta (centroida) ka centru svih uzoraka u klasteru – za svaki klaster se izračuna novo težište uzimajući prosek instanci koje su dodeljene tom klasteru
- Koraci 2) i 3) se ponavljaju sve dok algoritam ne konvergira ili broj iteracija ne dostigne MAX

K-means - Ulazni podaci

- Ulaz:
  - K – broj klastera
  - skup za trening sa m uzoraka (instanci), a svaki uzorak u skupu je vektor opisan sa n atributa (x1, x2, ... Xn)
  - Opciono: maksimalan broj iteracija koji se izvršava (max)

K-means - Funkcija koštanja/distorzije

- Smisao K-means algoritma je minimizacija funkcije koštanja J (eng. cost function):

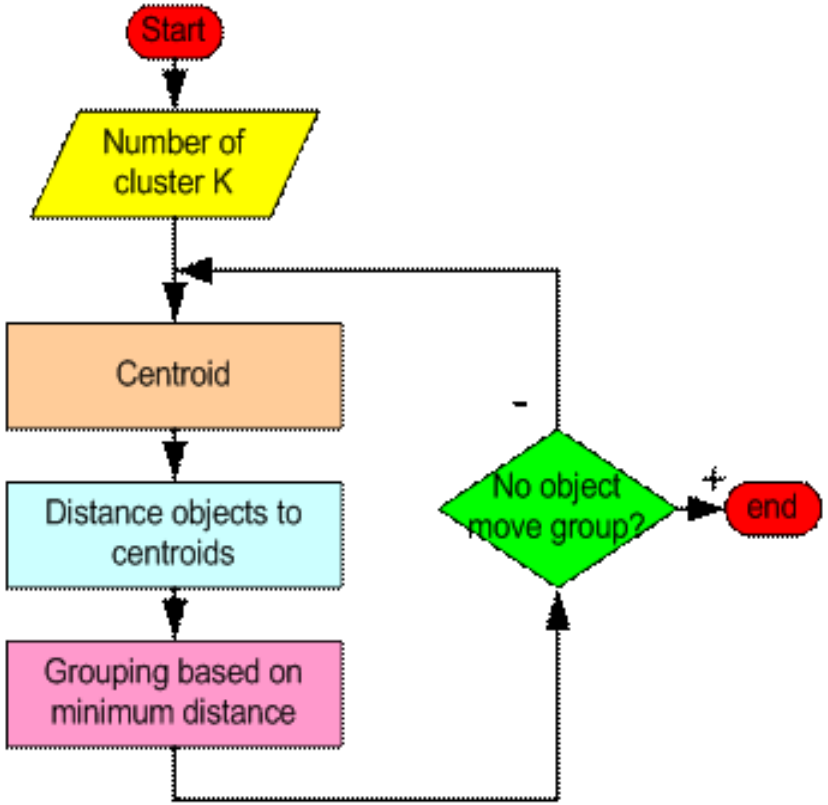
$$J\left(c^{(1)}, \dots, c^{(m)}, \mu_{(1)}, \dots, \mu_K\right) = \frac{1}{m}\sum_{i=1}^{m}\left\|x^{(i)} - \mu_{c^{(i)}}\right\|^2$$

$x^{(i)}$ – i-ta instanca u skupu podataka za trening, i = 1,...m
$c^{(i)}$ – indeks klastera u koji je instanca $x^{(i)}$ trenutno raspoređena
μ_j – težište klastera j, j = 1,... K, gde je K ukupan broj klastera
$\mu_{c(i)}$ – težište klastera u koji je instanca $x^{(i)}$ trenutno raspoređena

Minimizacija funkcije koštanja

- Minimizacija funkcije koštanja J kroz K-means algoritam:
  - Faza Grupisanja po klasterima (korak 2) minimizuje J po parametrima c^1, c^2,... c^m, držeći μ_1, μ_2,... μ_k fiksnim
  - Faza Pomeranja težišta minimizuje J po parametrima μ_1, μ_2,... μ_k, držeći c^1, c^2,... c^m, fiksnim

K-mean algoritam

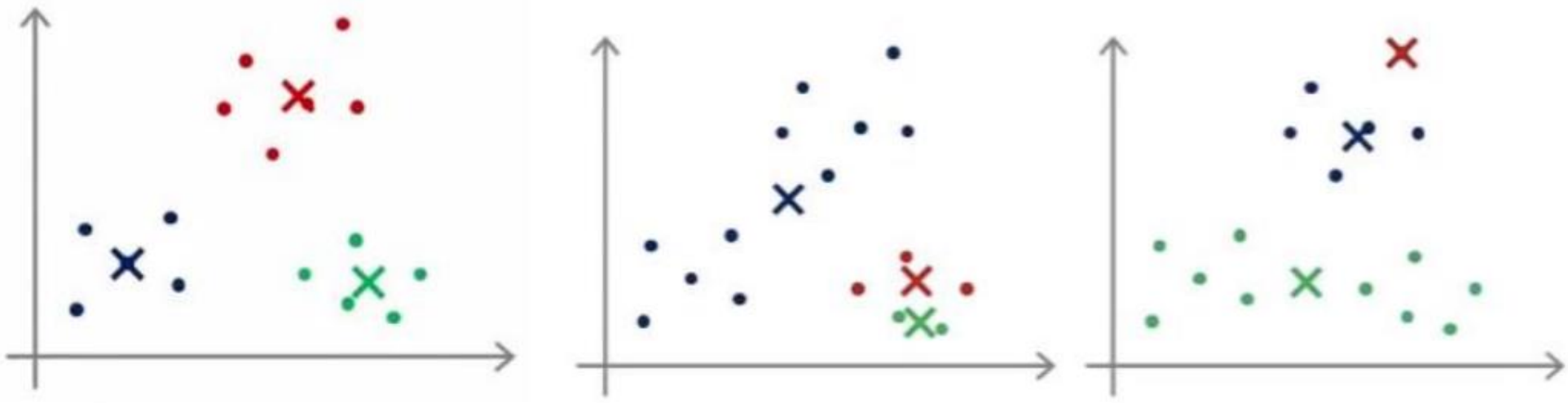


K-means evaluacija

- Da li imamo tačno rešenje?
- Kriterijumi za procenu kvaliteta kreiranih klastera
  - Međusobna udaljenost težišta
  - Standarda devijacija pojedinačnih instanci u odnosu na težište
  - Suma kvadrata unutar klastera

Problem inicijalnog izbora težišta

- U zavisnosti od inicijalnog izbora težišta:
  - K-means algoritam može konvergirati brže ili sporije
  - Može se doći do lokalnog minimuma i time imamo loše rešenje (lokalni minimum funkcije koštanja)



K-means - višestruka nasumična inic.

- Omogućava da se izbegne ulazak u lokalni minimum
- Sastoji se u sledećim koracima:
  for i = 1 to N { // N obicno ima vrednosti izmedju 50 i 1000
  Nasumicno odabrati inicijalni skup težišta
  Izvršiti K-means algoritam
  Izračunati funkciju koštanja
  }
  Izabrati instancu algoritma koja daje najmanju vrednost funkcije koštanja.
- Ovaj pristup daje dobre rezultate ukoliko je broj klastera relativno mali (između 2 i 10), za veći broj klastera ne treba da se koristi.

K-means: Kako odrediti K?

- Kako odrediti broj klastera K?
  - Ukoliko imamo znanje o oblasti/pojavi koju podaci opisuju
    - Pretpostaviti broj K na osnovu domenskog znanja
    - Testirati model sa K-1, K i K+1 klastera i uporediti grešku
  - Ukoliko ne posedujemo znanje o oblasti/pojavi koju podaci opisuju
    - Krenuti od malog broja iteracija i u više iteracija testirati model uvek sa jednim klasterom više
    - U svakoj iterarciji uporediti grešku (nekom metodom) tekućeg i prethodnog modela i kad smanjenje greške postane zanemarljivo, prekinuti postupak

K-means

- Prednosti
  - Jednostavan iterativni metod
  - Korisnik određuje „K“
- Mane
  - Suviše jednostavno – loši rezultati
  - Ponekada je teško odrediti korektnu vrednost za „K“

Hvala na pažnji

Kontakt: drazen.draskovic@etf.bg.ac.rs

bosko.nikolic@etf.bg.ac.rs