KLASTERIZACIJA

JELENA JOVANOVIĆ

Email: jeljov@gmail.com

Web: http://jelenajovanovic.net

Pregled predavanja

- Šta je klasterizacija?
- Koje su oblasti/primeri primene?
- Klasterizacija primenom K-Means algoritma
  - Upoznavanje sa algoritmom kroz primer
  - K-Means algoritam
  - Potencijalni problemi pri primeni algoritma
  - Primer primene u WEKA-i

Šta je klasterizacija?

Klasterizacija je jedan od oblika nenadgledanog m. učenja

- ono što je raspoloživo od podataka su podaci o instancama koje je potrebno na neki način grupisati;
- ne posedujemo podatke o poželjnoj / ispravnoj grupi za ulazne instance

Šta je klasterizacija?

Klasterizacija je zadatak grupisanja instanci, tako da za svaku instancu važi da je *sličnija* instancama iz svoje grupe (klastera), nego instancama iz drugih grupa (klastera)

Sličnost instanci se određuje primenom neke od mera za računanje

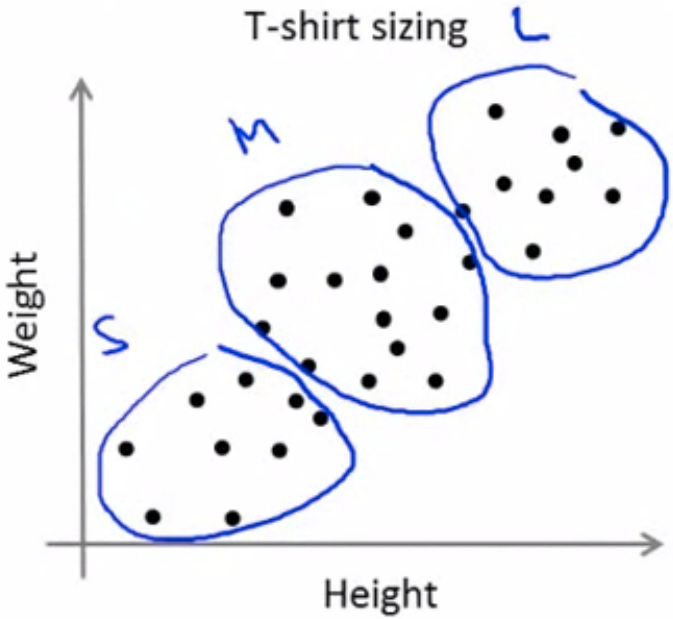
- sličnosti (npr. Kosinusna sličnost) ili
- udaljenosti dva objekta (npr. Euklidska udaljenost)

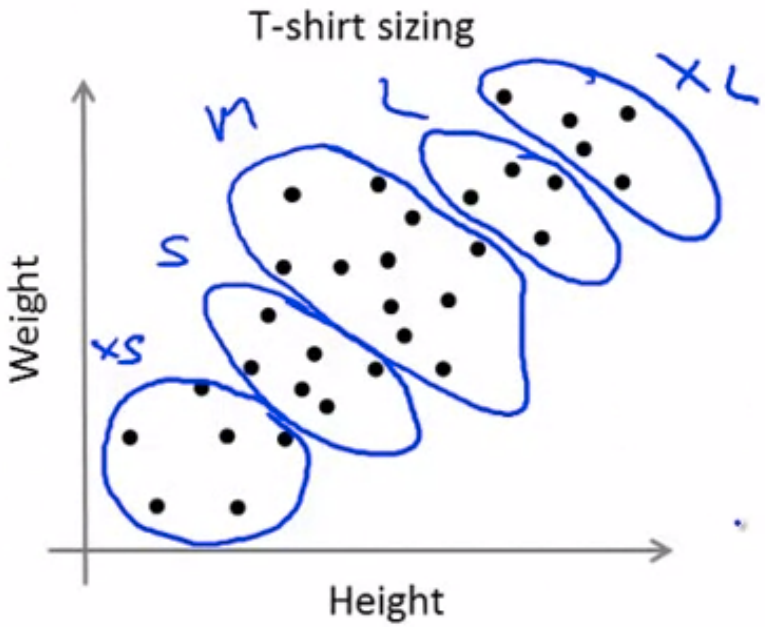
ŠTA JE KLASTERIZACIJA?

Za razliku od klasifikacije, ovde nemamo “tačno” rešenje

- ocena uspešnosti algoritma je dosta teža nego kod klasifikacije
- pogodnost 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

Primer različitih dobrih rešenja za isti polazni skup podataka





Oblasti primene

- Segmentacija tržišta
- Uočavanje grupa u društvenim mrežama
- Identifikacija paterna u ponašanju korisnika nekog Web sajta
- Grupisanje objekata (npr., slika/dokumenata) prema zajedničkim karakteristikama
- …

K-MEANS ALGORITAM

K-MEANS

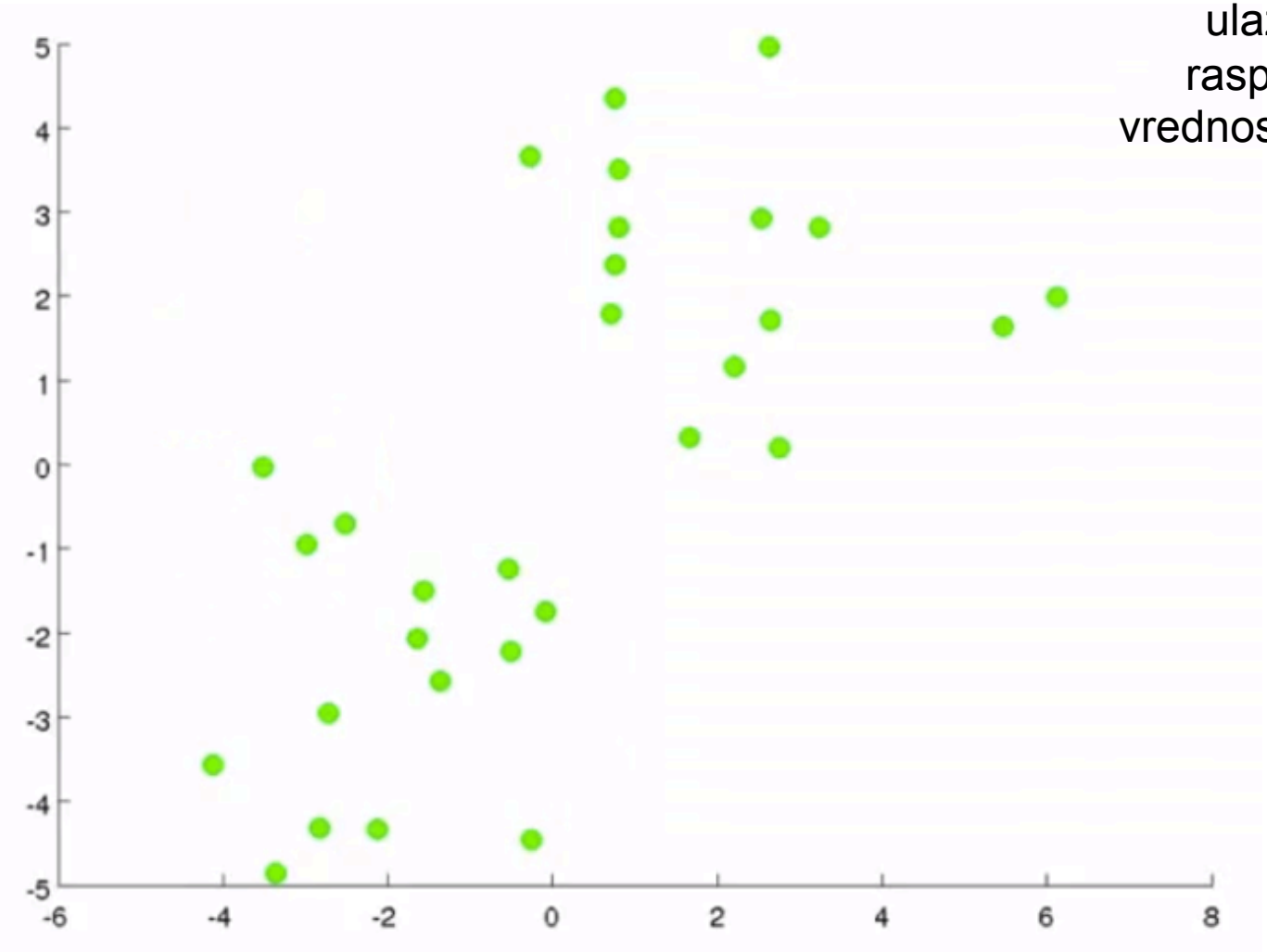
Jedan od najpoznatijih i najjednostavnijih algoritama klasterizacije

Najlakše ga je razumeti na primeru, pa ćemo prvo razmotriti jedan primer

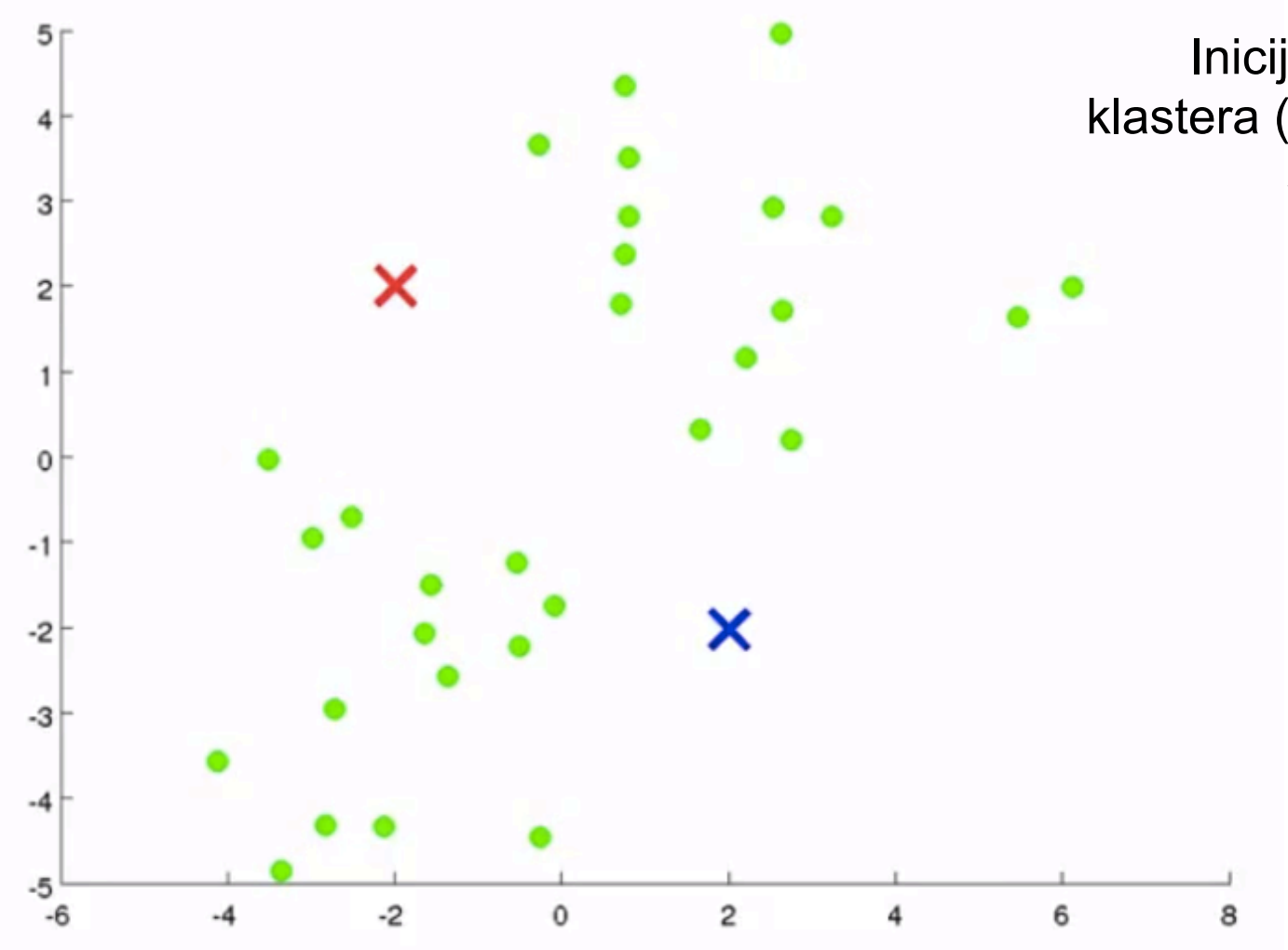
Primer je preuzet iz kursa: https://www.coursera.org/course/ml

K-MEANS ALGORITAM – PRIMER

Pretpostavimo da su ovo ulazni podaci kojima raspolažemo, opisani vrednostima dva atributa

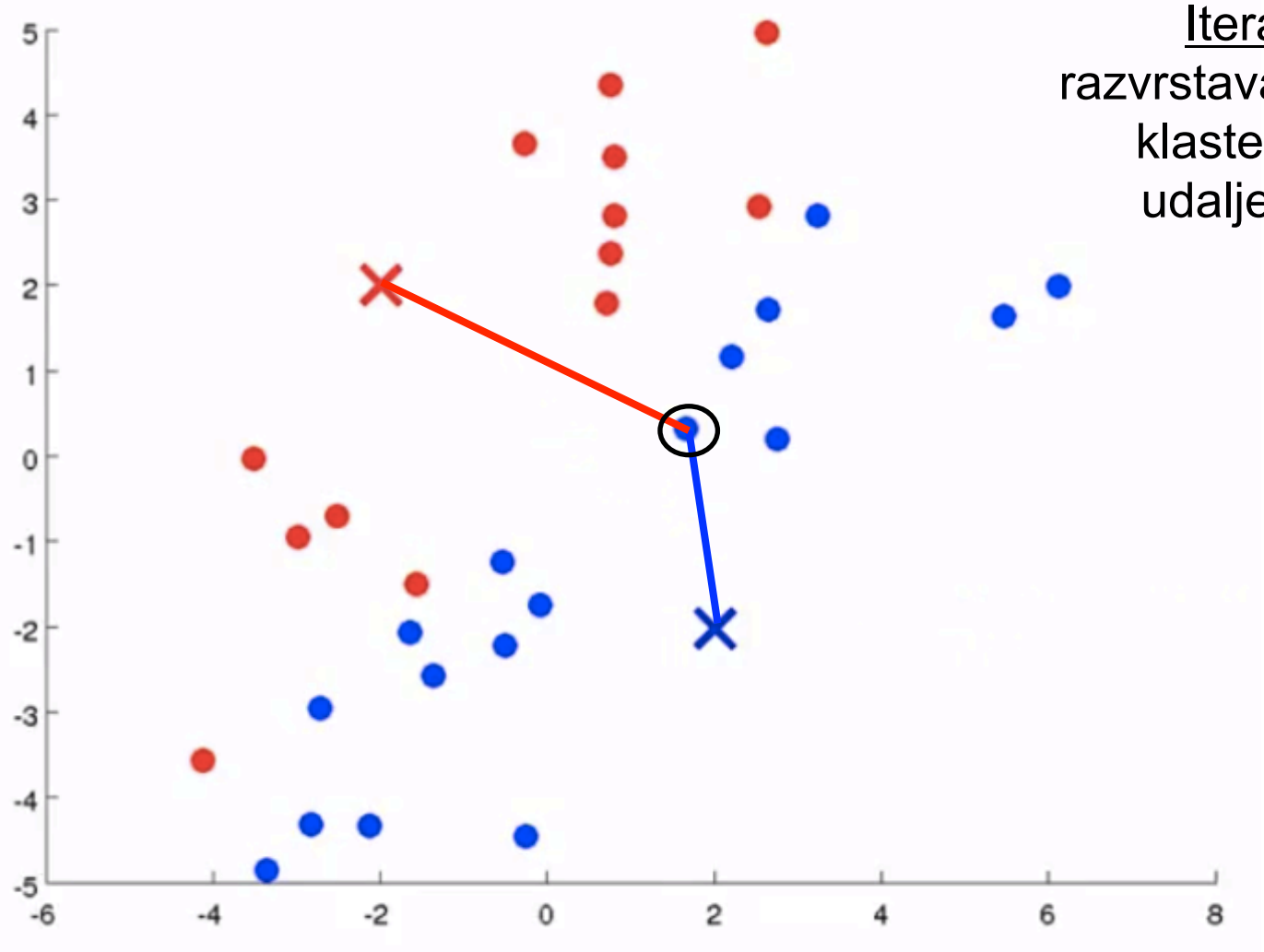


K-MEANS: PRIMER



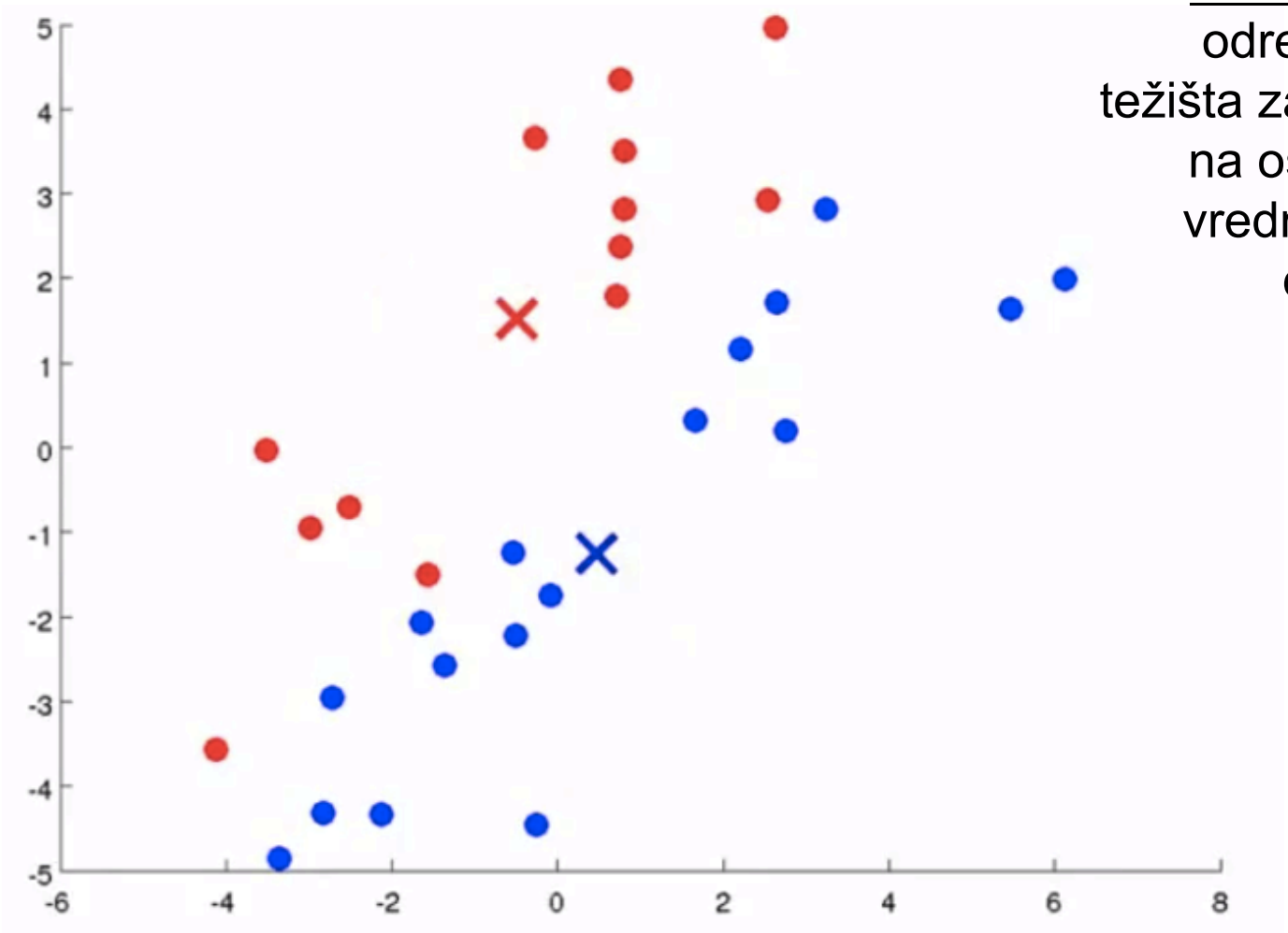
<u>Inicijalizacija</u>:
Inicijalni izbor težišta klastera (K = 2) metodom slučajnog izbora

K-MEANS: PRIMER



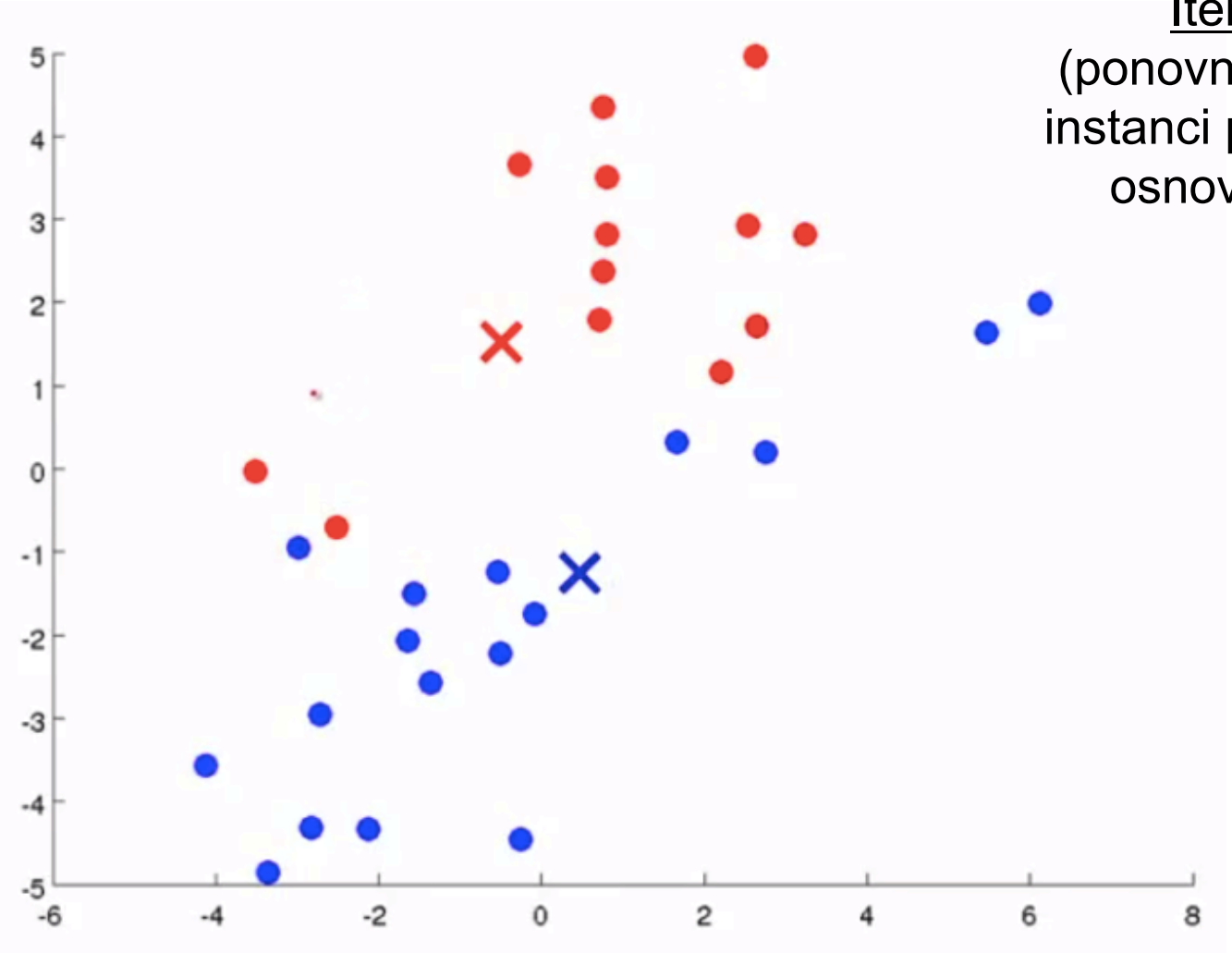
Iteracija 1, korak 1:
razvrstavanje instanci po klasterima na osnovu udaljenosti od težišta klastera

K-MEANS: PRIMER



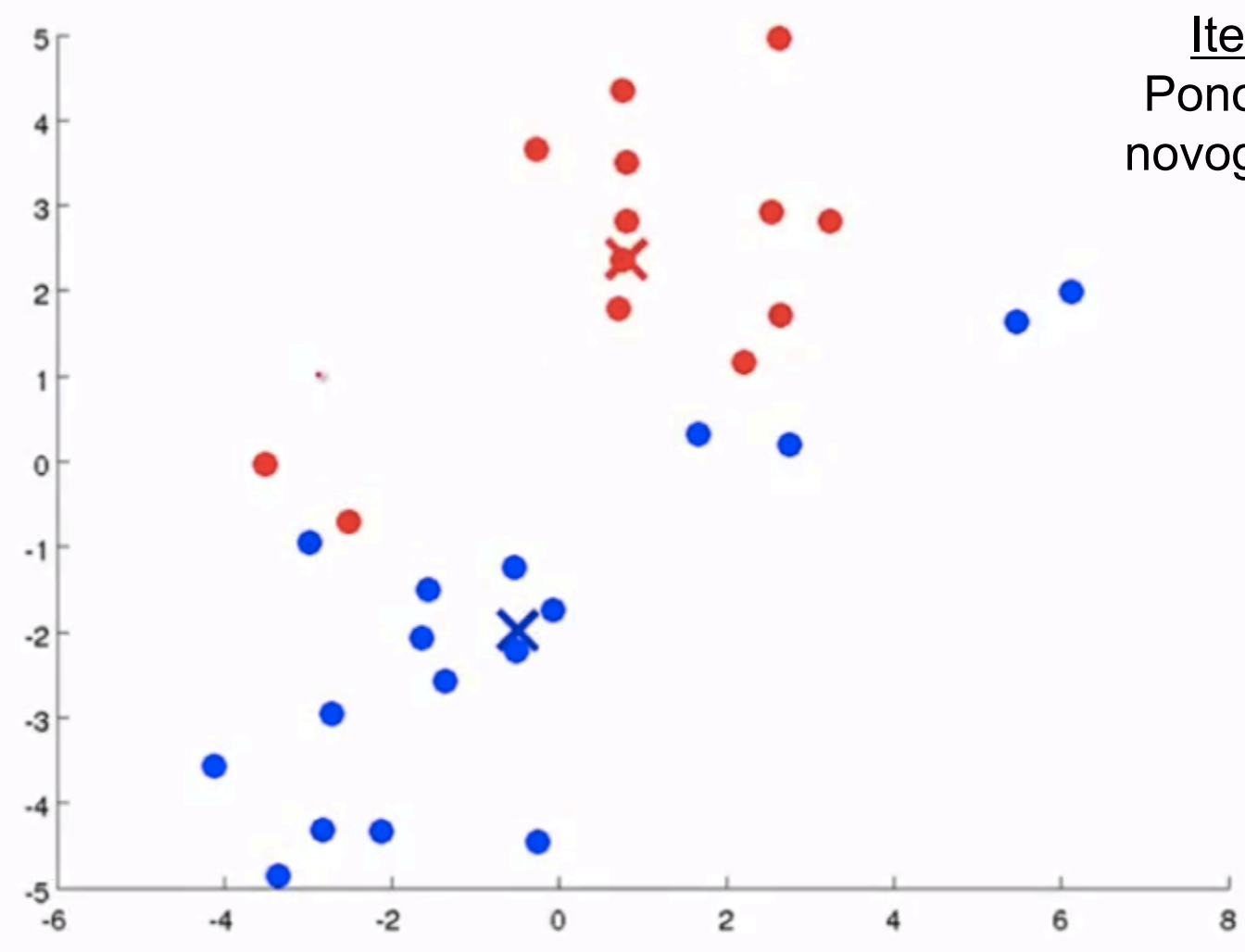
Iteracija 1, korak 2: određivanje novog težišta za svaki klaster, na osnovu proseka vrednosti instanci u datom klasteru

K-MEANS: PRIMER



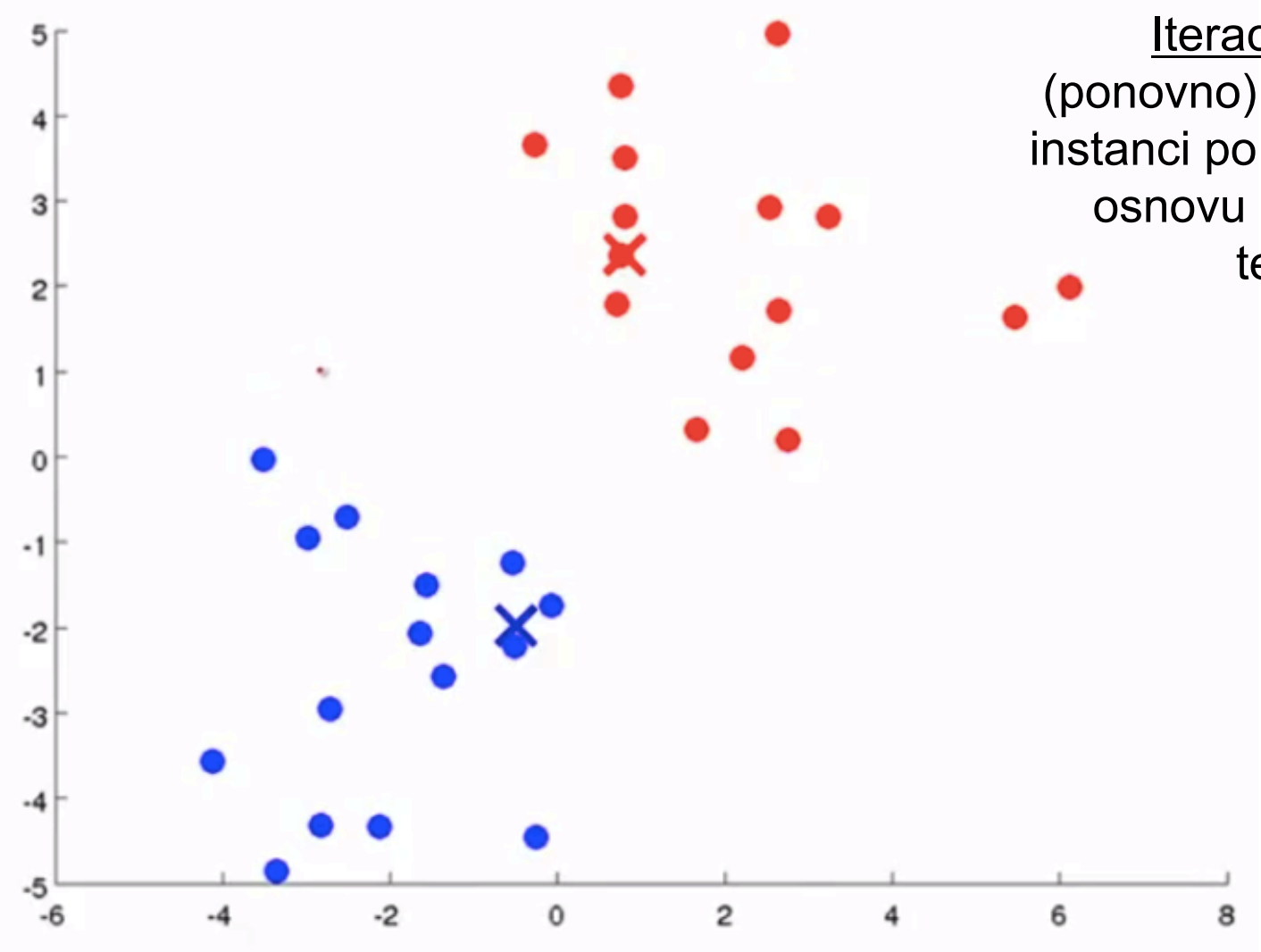
Iteracija 2, korak 1:
(ponovno) razvrstavanje instanci po klasterima na osnovu udaljenosti od težišta klastera

K-MEANS: PRIMER



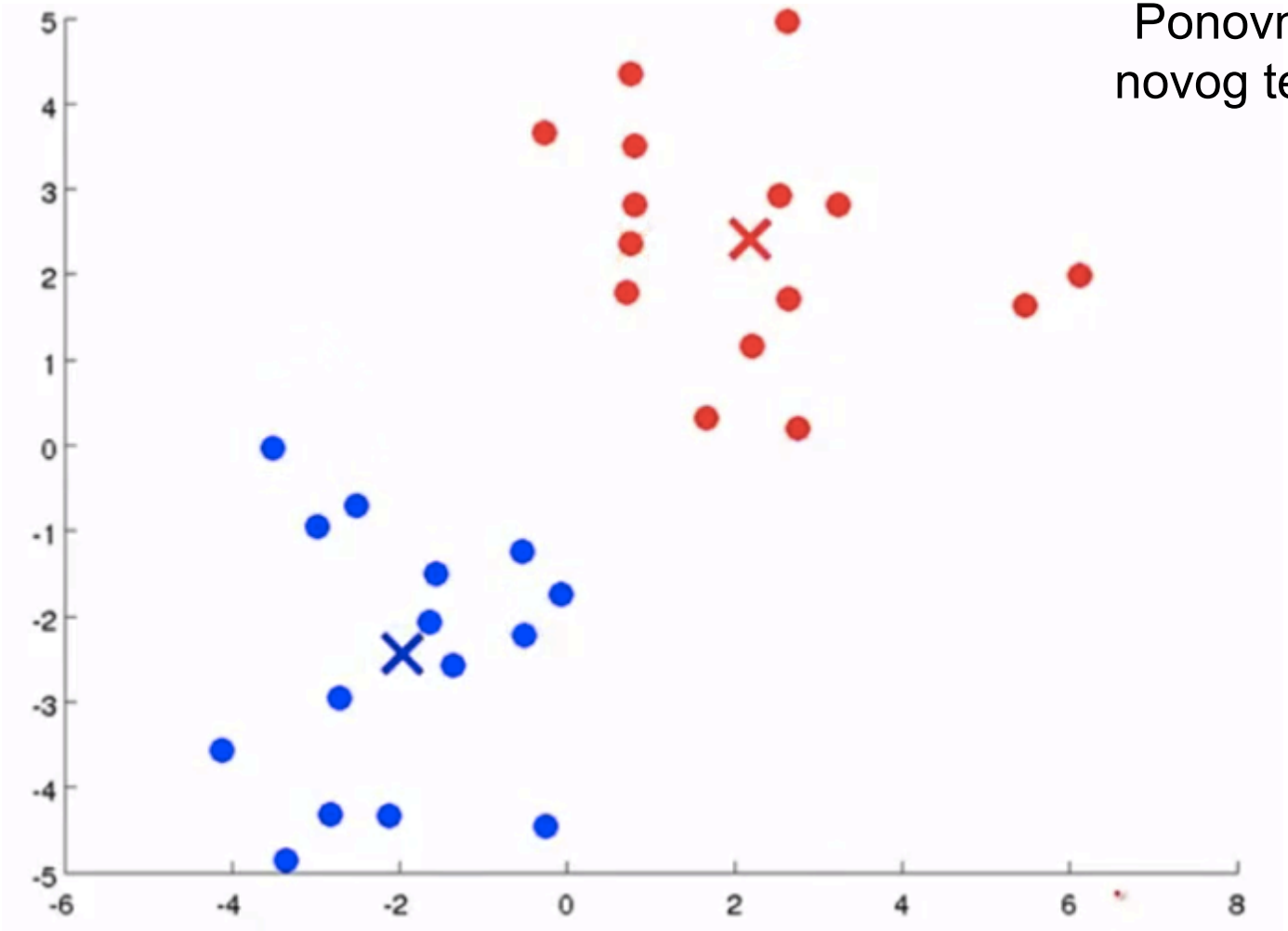
Iteracija 2, korak 2:
Ponovno određivanje novog težišta za svaki klaster

K-MEANS: PRIMER



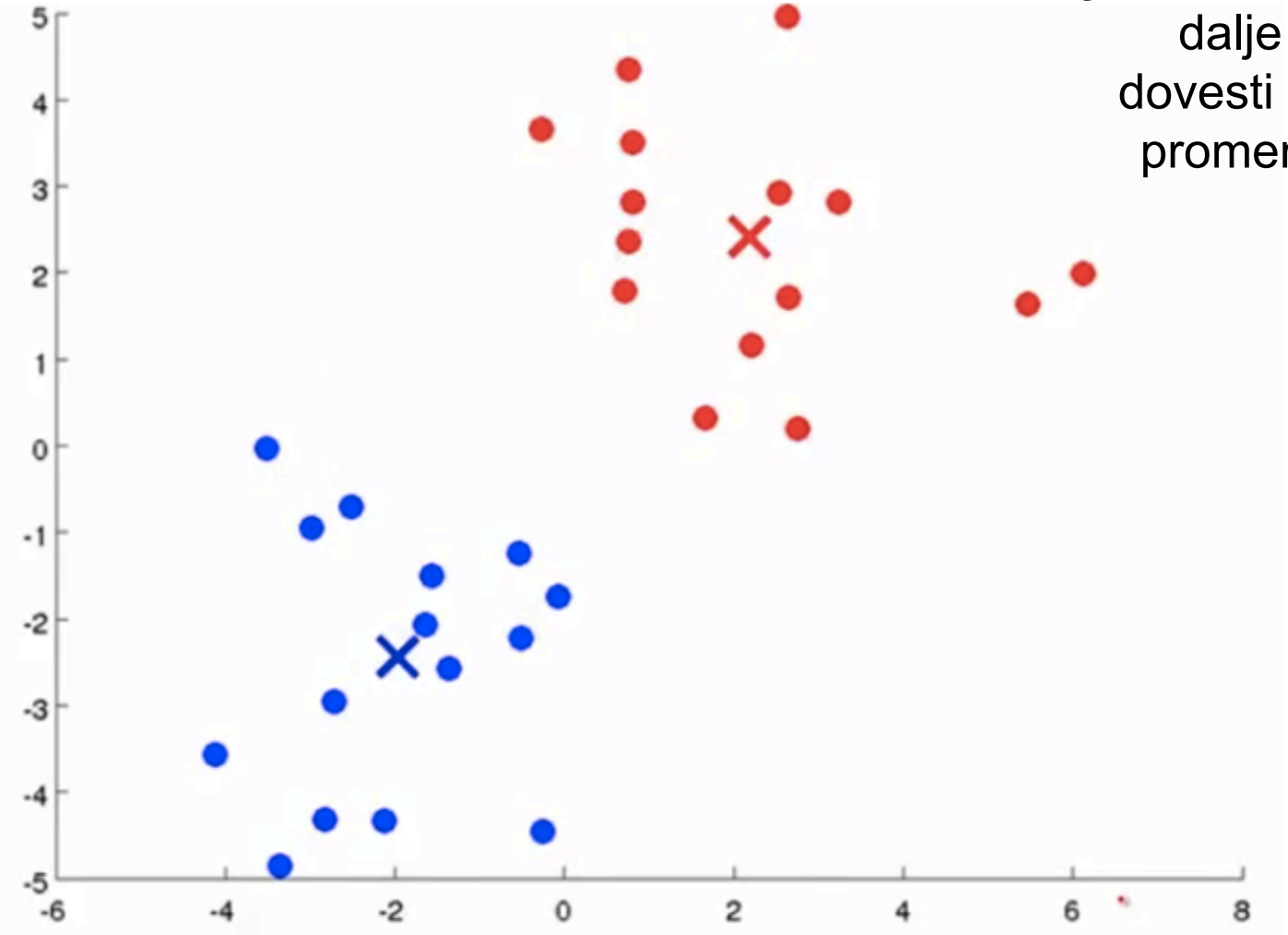
Iteracija 3, korak 1:
(ponovno) razvrstavanje instanci po klasterima na osnovu udaljenosti od težišta klastera

K-MEANS: PRIMER



Iteracija 3, korak 2:
Ponovno određivanje novog težišta za svaki klaster

K-MEANS: PRIMER



Algoritam već konvergira: dalje iteracije neće dovesti do značajnijih promena i proces se zaustavlja

K-MEANS: ALGORITAM

Ulaz:

- *K* - broj klastera
- (neobeležen) skup za trening sa *m* instanci; svaka instanca u skupu je vektor opisan sa *n* atributa (x_1, x_2, …, x_n)
- *max* - max broj iteracija (opcioni parametar)

K-MEANS: ALGORITAM

Koraci:

1) Inicijalni izbor težišta klastera, slučajnim izborom
   - težišta se biraju iz skupa instanci za trening, tj. *K* instanci za trening se nasumično izabere i proglasi za težišta

2) Ponoviti dok algoritam ne konvergira ili broj iteracija <= *max*:
   1) *Grupisanje po klasterima*: za svaku instancu iz skupa za trening, i = 1,m, identifikovati najbliže težište i dodeliti instancu klasteru kome to težište pripada
   2) *Pomeranje težišta*: za svaki klaster izračunati novo težište uzimajući prosek tačaka (instanci) koje su dodeljene tom klasteru

K-MEANS ALGORITAM: FUNKCIJA KOŠTANJA

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

$$J(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K) = \frac{1}{m} \sum_{i=1}^{m} ||x^{(i)} - \mu_{c^{(i)}}||^2$$

$x^{(i)}$ – *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

$\mu_{c(i)}$ – težište klastera u koji je instanca $x^{(i)}$ trenutno raspoređena

Ova funkcija se zove i funkcija distorzije (distortion function)

K-MEANS ALGORITAM: FUNKCIJA KOŠTANJA

$$\min_{\substack{c^{(1)},\ldots,c^{(m)},\\ \mu_1,\ldots,\mu_K}} J(c^{(1)},\ldots,c^{(m)},\mu_1,\ldots,\mu_K)$$

Minimizacija funkcije koštanja ***J*** kroz K-means algoritam:

- faza *Grupisanja po klasterima* minimizuje ***J*** po parametrima $c^{(1)}$, …,$c^{(m)}$, držeći μ_1,…, μ_K fiksnim
- faza *Pomeranja težišta* minimizuje ***J*** po parametrima μ_1,…, μ_K, držeći $c^{(1)}$,…,$c^{(m)}$ fiksnim

K-MEANS: EVALUACIJA

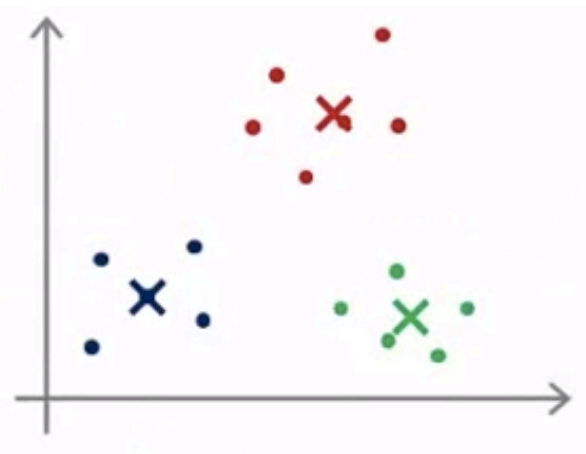
Kriterijumi za procenu “kvaliteta” kreiranih klastera:

- Međusobna udaljenost težišta
  - što su težišta dalje jedno od drugog, to je stepen preklapanja klastera manji, i njihov kvalitet viši
- St. devijacija pojedinačnih instanci u odnosu na težište
  - što je st. devijacija manja, to su instance tešnje grupisane oko težišta i klasteri se smatraju boljim
- Suma kvadrata greške unutar klastera (within cluster sum of squared errors)
  - daje kvantitativnu meru za procenu kvaliteta kreiranih klastera
  - razmotrićemo detaljnije na primeru (slajd 23)

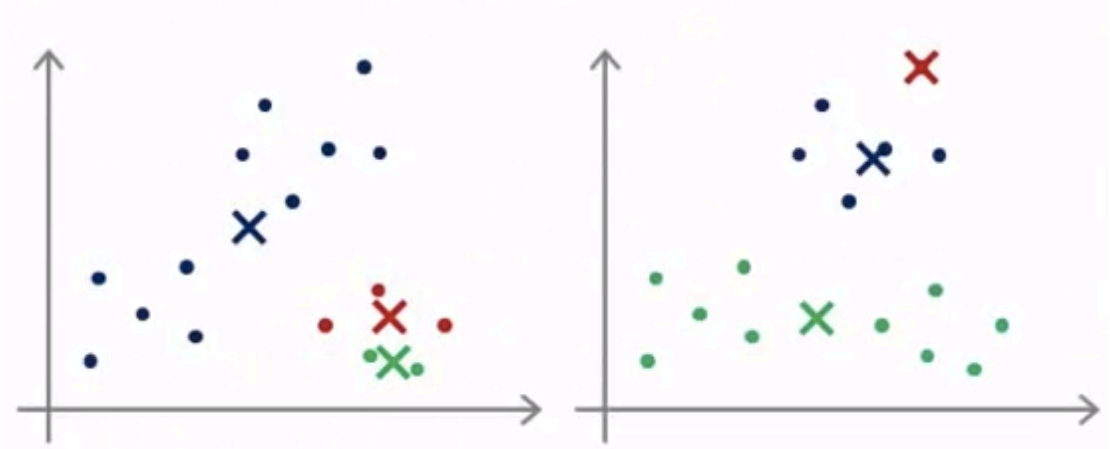
K-MEANS:
PROBLEM INICIJALNOG IZBORA TEŽIŠTA

Zavisno od inicijalnog izbora težišta:

- K-means algoritam može konvergirati brže ili sporije;
- Takodje, može “upasti” u tzv. lokalni minimum i dati loše rešenje
  - reč je o lokalnom min. funkcije koštanja



Dobra inicijalizacija



Inicijalizacija koja vodi u lokalne minimume

K-MEANS:
VIŠESTRUKA NASUMIČNA INICIJALIZACIJA

Omogućuje da se izbegnu situacije koje K-means dovode u lokalni minimum

Sastoji se u sledećem:

```
for i = 1 to n { //n obično uzima vrednosti 50 - 1000
      Nasumično odabrati inicijalni skup težišta;
      Izvršiti K-Means algoritam;
      Izračunati funkciju koštanja (cost function)
}
Izabrati instancu algoritma koja daje najmanju vrednost
za f. koštanja
```

Ovaj pristup daje dobre rezultate ukoliko je broj klastera relativno mali (2 - 10); za veći broj klastera ne bi ga trebalo koristiti

K-MEANS: KAKO ODREDITI K?

Kako odrediti broj klastera K?

- U slučaju da posedujemo znanje o fenomenu/pojavi koju podaci opisuju
  - Pretpostaviti broj klastera (K) na osnovu domenskog znanja
  - Testirati model sa K-1, K, K+1 klastera i uporediti grešku*

- Ukoliko ne posedujemo znanje o fenomenu/pojavi
  - Krenuti od malog broja klastera i u više iteracija testirati model uvek sa jednim klasterom više
  - U svakoj od iteracija, uporediti grešku* tekućeg i prethodnog modela i kad smanjenje greške postaje zanemarljivo, prekinuti postupak

*Na primer, within cluster sum of squared errors

K-MEANS: PRIMER PRIMENE U WEKA-I

Primer je preuzet iz članka sa IBM Developer Works sajta:

http://www.ibm.com/developerworks/library/os-weka2/

ZAHVALNICA I PREPORUKA

Stanford

Machine Learning

Andrew Ng

Learn about the most effective machine learning techniques, and gain practice implementing them and getting them to work for yourself.

Workload: 5-7 hours/week

Taught In: English

Subtitles Available In: English

Preview

Sessions:

Oct 14th 2013 (10 weeks long) Sign Up

Apr 22nd 2013 (10 weeks long) Sign Up

3,484 Tweet 12k +1 13k Like

Coursera:
https://www.coursera.org/course/ml

Stanford YouTube channel:
http://www.youtube.com/view_play_list?p=A89DCFA6ADACE599

(Anonimni) upitnik za vaše kritike, komentare, predloge:

http://goo.gl/cqdp3l