KLASTERIZACIJA

JELENA JOVANOVIĆ

jelena.jovanovic@fon.bg.ac.rs

Pregled predavanja

- Šta je klasterizacija?
- Tipični primeri primene
- Klasterizacija primenom K-Means algoritma

Šta je klasterizacija?

Klasterizacija je oblik 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 (klasi) za date instance

Šta je klasterizacija?

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

Procena sličnosti instanci

Sličnost (blizina) instanci se procenjuje primenom neke od mera za računanje:

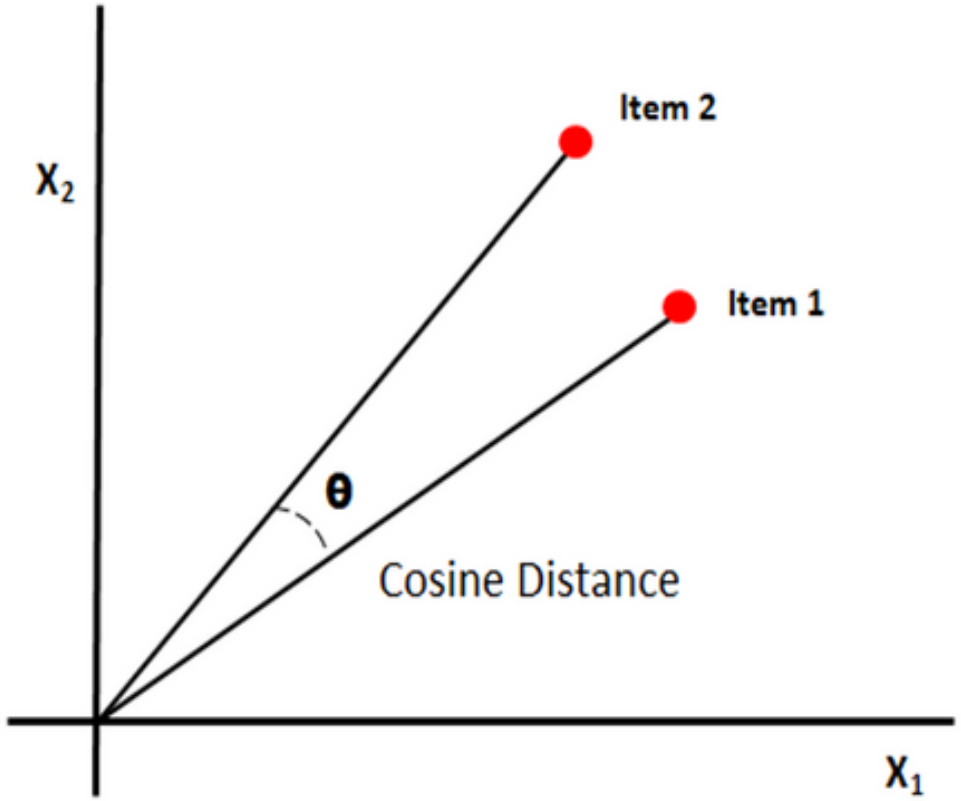
- sličnosti (npr. kosinusna sličnost ili koeficijent korelacije)

ili

- udaljenosti dve instance (npr. Euklidska ili Menhetn udaljenost)

PROCENA SLIČNOSTI INSTANCI

Kosinusna sličnost



$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

Izvor slike: https://www.tyrrell4innovation.ca/miword-of-the-day-iscosine-distance/

PROCENA SLIČNOSTI INSTANCI

Euklidska udaljenost:

$$d = \sqrt{\sum_{j=1}^{n} (x_{sj} - x_{tj})^2}$$

Manhattan udaljenost:

$$d = \sum_{j=1}^{n} |x_{sj} - x_{tj}|$$

n – broj atributa kojima su instance opisane

Izvor slike: http://goo.gl/9kgb4A

Procena rezultata klasterizacije

- Ocena uspešnosti modela je dosta teža nego kod nadgledanog m. učenja
- Ovde nemamo egzaktne metrike koje nedvosmisleno ukazuju na to koliko je model “dobar”

Procena rezultata klasterizacije

- Ocena uspešnosti modela je dosta teža nego kod nadgledanog m. učenja
- Ovde nemamo egzaktne metrike koje nedvosmisleno ukazuju na to koliko je model “dobar”

Procena rezultata klasterizacije

Pod “dobrim” rešenjem se podrazumeva model koji:

- Dobro deli instance u nepreklapajuće grupe (klastere) (objektivna procena)
- Koristan je za dati zadatak / problem zbog koga se klasterovanje i radi (subjektivna procena)

Procena rezultata klasterizacije

Neki od objektivnih kriterijuma za procenu kvaliteta klastera:

- Međusobna udaljenost težišta
  - što su težišta dalje jedno od drugog, to je stepen preklapanja klastera manji, i njihov kvalitet viši
- Max udaljenost instanci u okviru istog klastera
- Min udaljenost instanci iz različitih klastera
- Suma kvadrata unutar klastera
  - suma kvadrata odstupanja instanci u okviru klastera od težišta klastera

Procena rezultata klasterizacije

- Subjektivna procena korisnosti klastera za dati domen i zadatak je značajnija od opisanih objektivnih metrika

Procena rezultata klasterizacije

- Subjektivna procena korisnosti klastera za dati domen i zadatak je značajnija od opisanih objektivnih metrika
- Problem: ne postoje metrike koje ukazuju na to koliko je neko rešenje sveukupno dobro, odnosno korisno za dati zadatak

Procena rezultata klasterizacije

- Subjektivna procena korisnosti klastera za dati domen i zadatak je značajnija od opisanih objektivnih metrika
- Problem: ne postoje metrike koje ukazuju na to koliko je neko rešenje sveukupno dobro, odnosno korisno za dati zadatak
- Domensko znanje presudno za evaluaciju, tj izbor optimalnog skupa klastera

Oblasti primene

- Segmentacija tržišta
- Uočavanje grupa u društvenim mrežama
- Identifikacija korisnika koje karakterišu slični oblici interakcije sa sadržajima nekog Web sajta/aplikacije
- Grupisanje objekata (npr., slika/dokumenata) radi lakše i efektivnije pretrage
- …

K-MEANS ALGORITAM

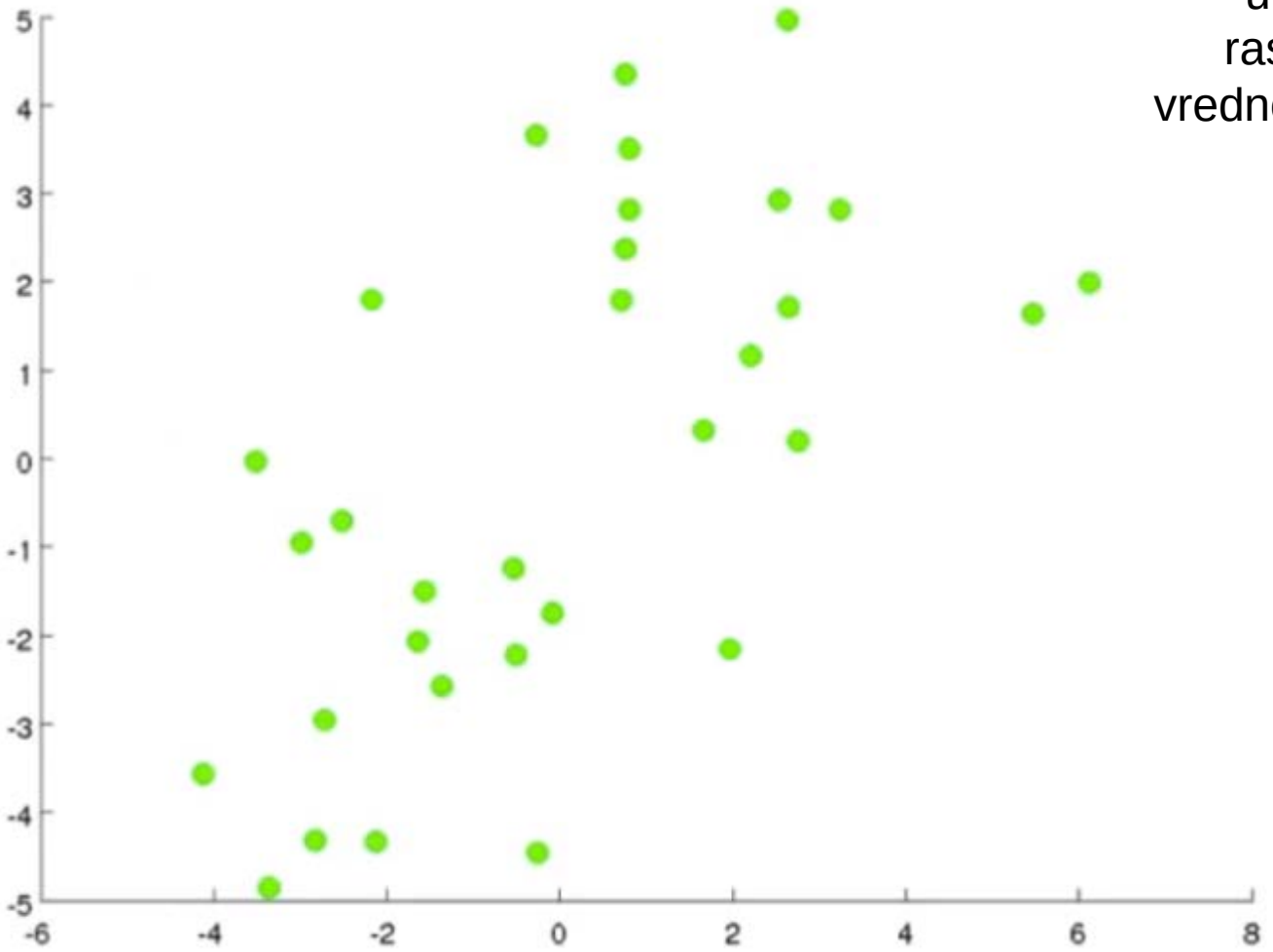
K-MEANS

Jedan od najpoznatijih i najjednostavnijih algoritama klasterizacije

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

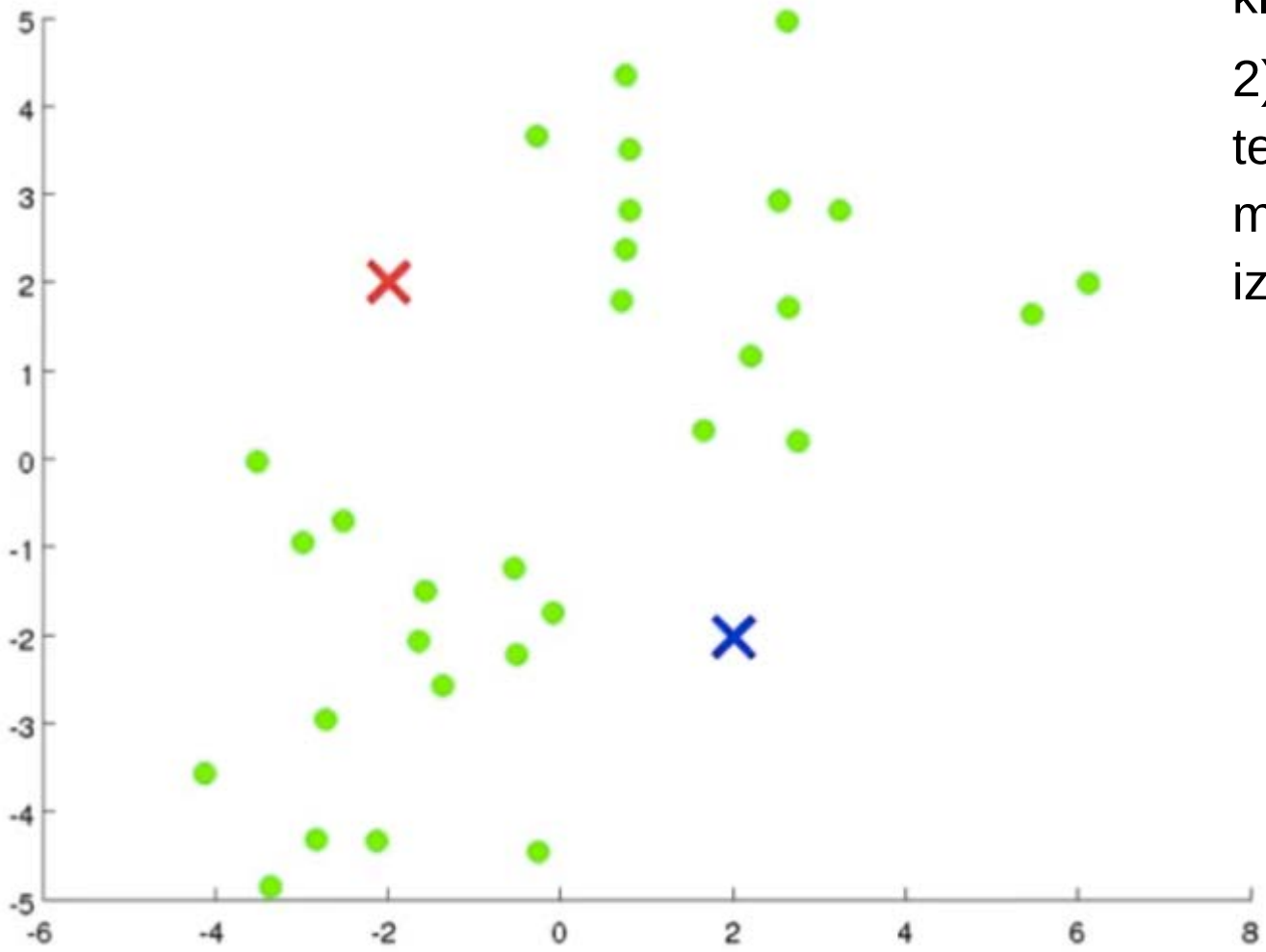
Primer je preuzet iz kursa: https://www.coursera.org/course/ml

K-MEANS ALGORITAM – PRIMER



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

K-MEANS: PRIMER

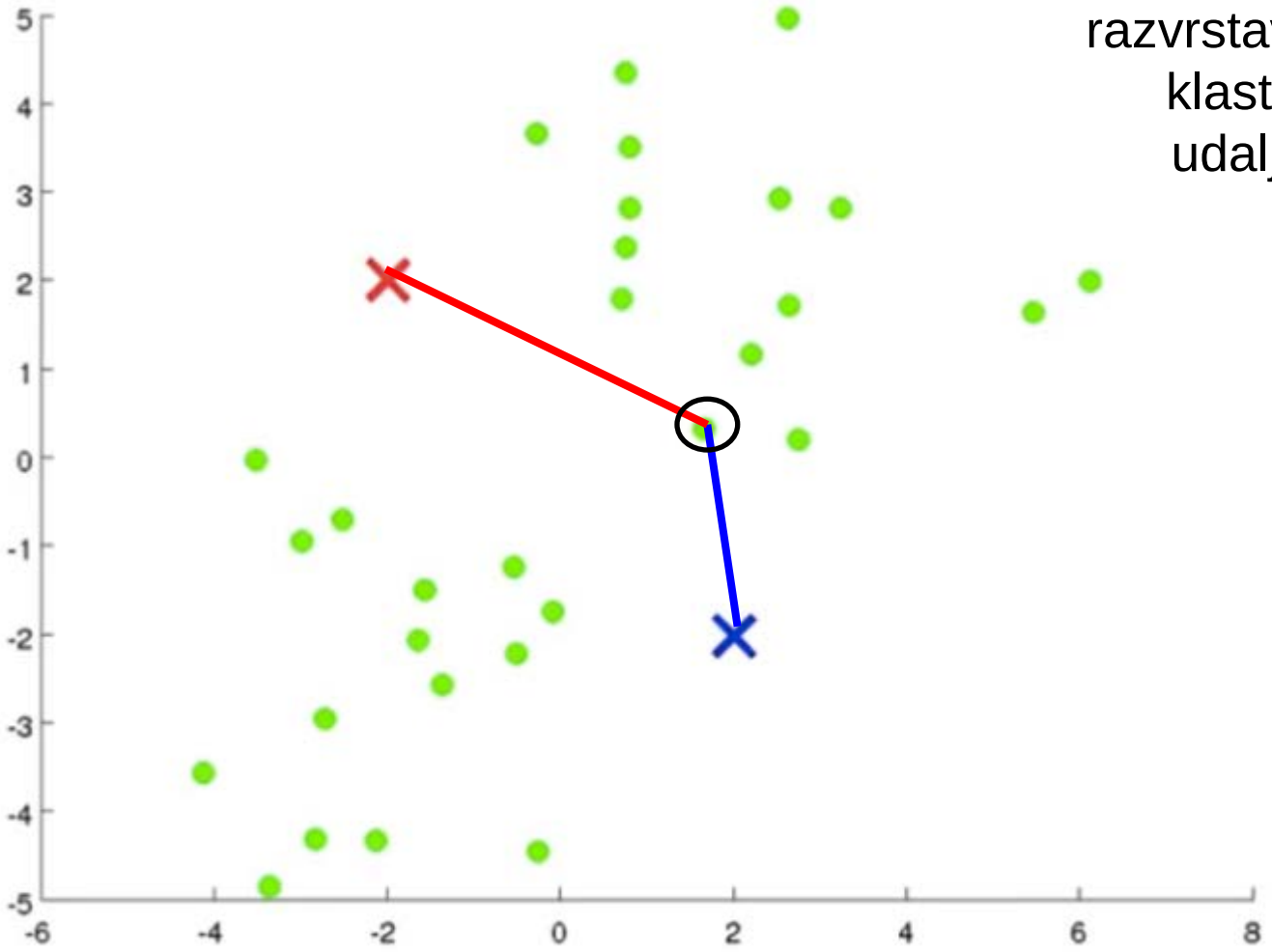


Inicijalizacija:

1) definisanje broja klastera, npr. K=2

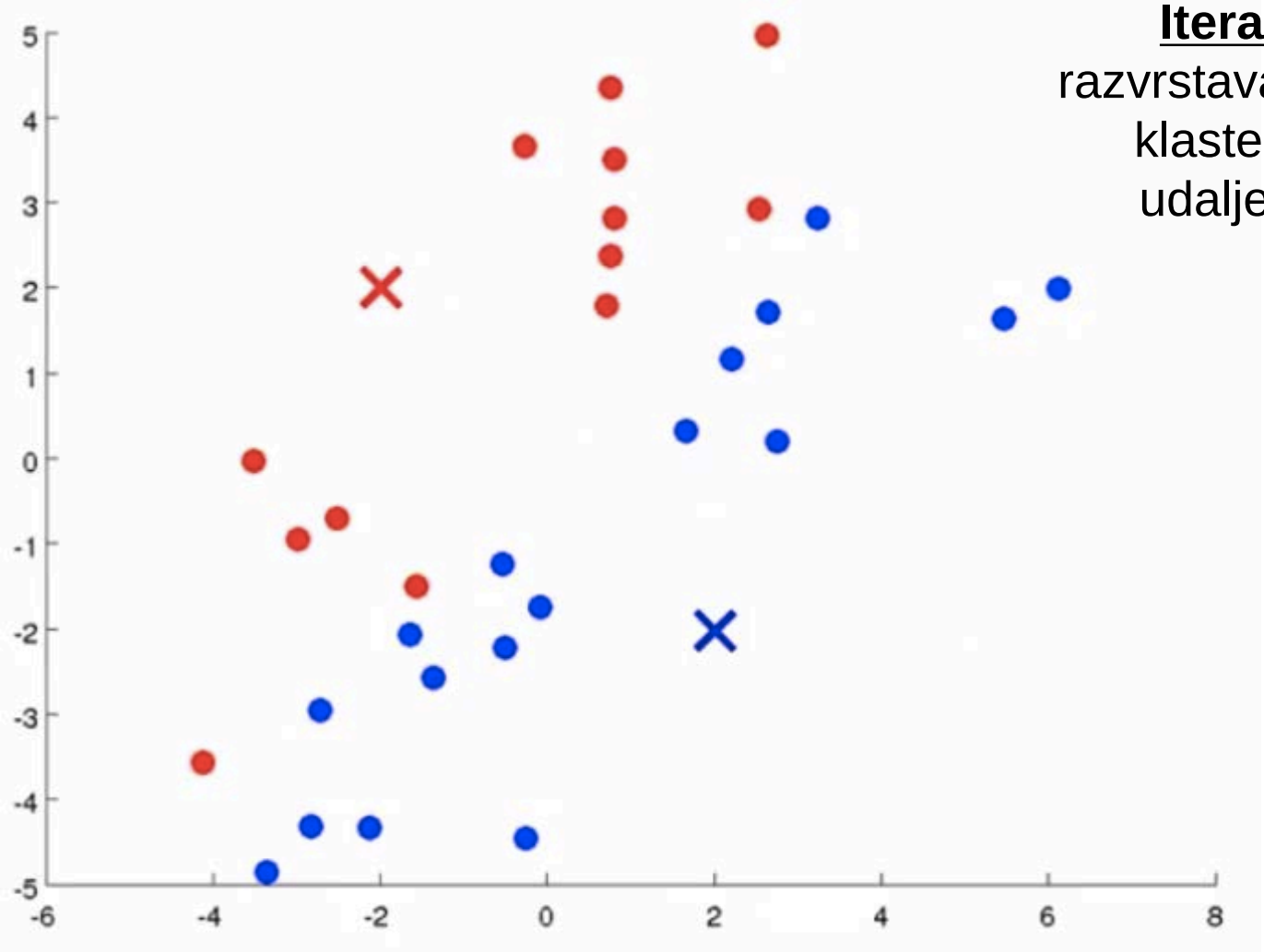
2) inicijalni izbor težišta klastera metodom slučajnog izbora

K-MEANS: PRIMER



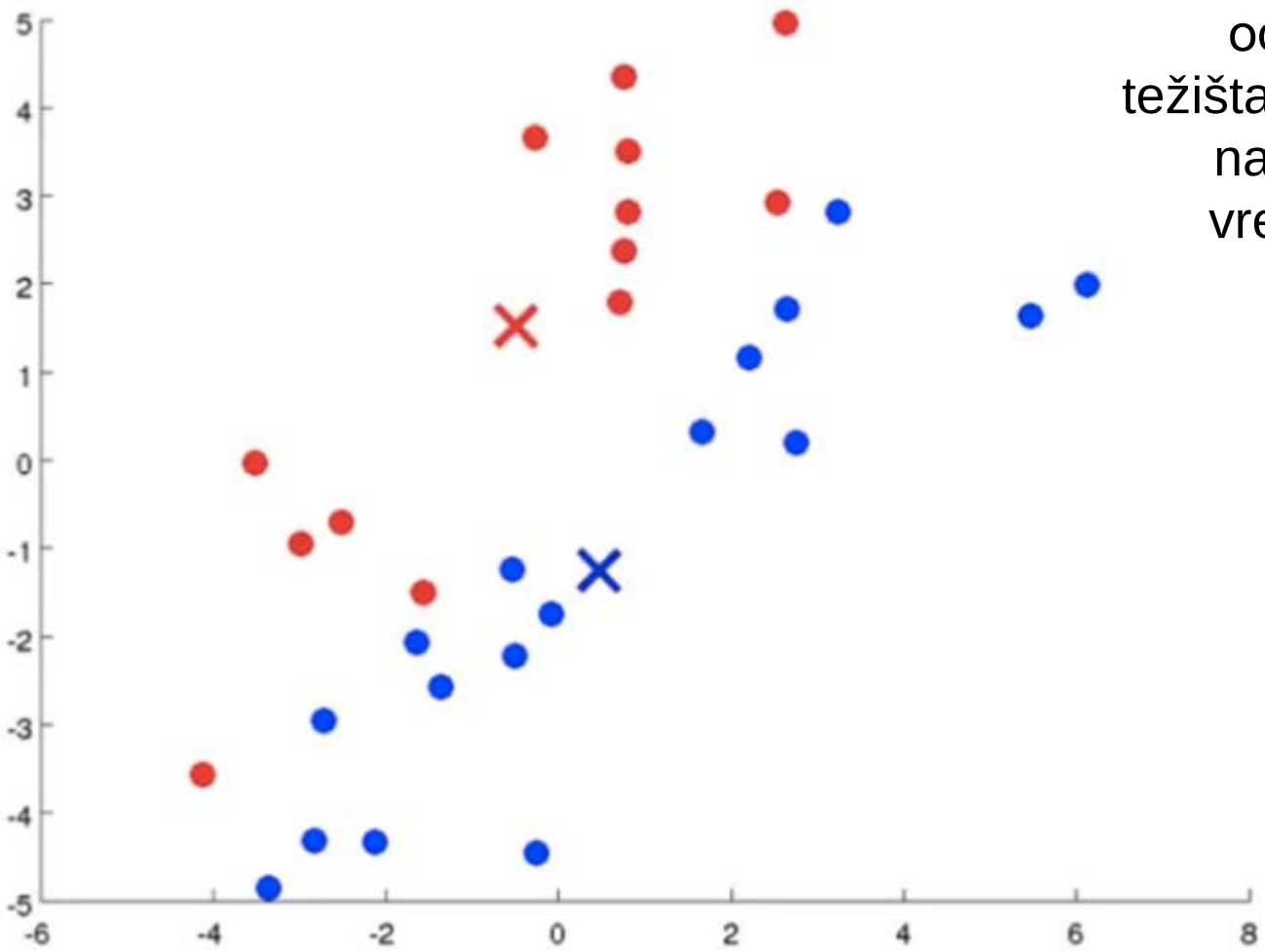
<u>Iteracija 1, korak 1</u>:
razvrstavanje instanci po klasterima na osnovu udaljenosti od težišta klastera

K-MEANS: PRIMER



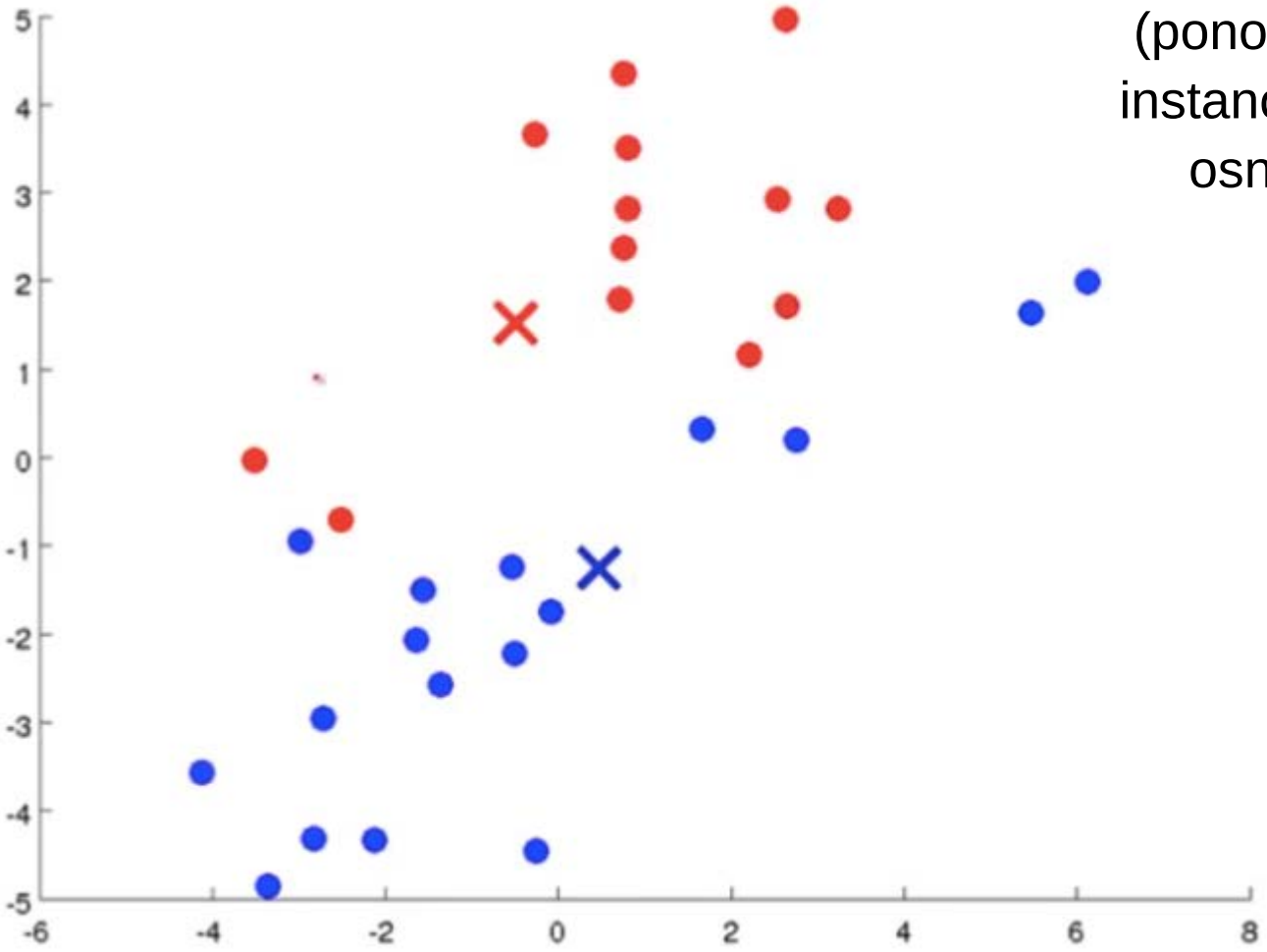
<u>Iteracija 1, korak 1</u>: razvrstavanje instanci po klasterima na osnovu udaljenosti od težišta klastera

K-MEANS: PRIMER



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

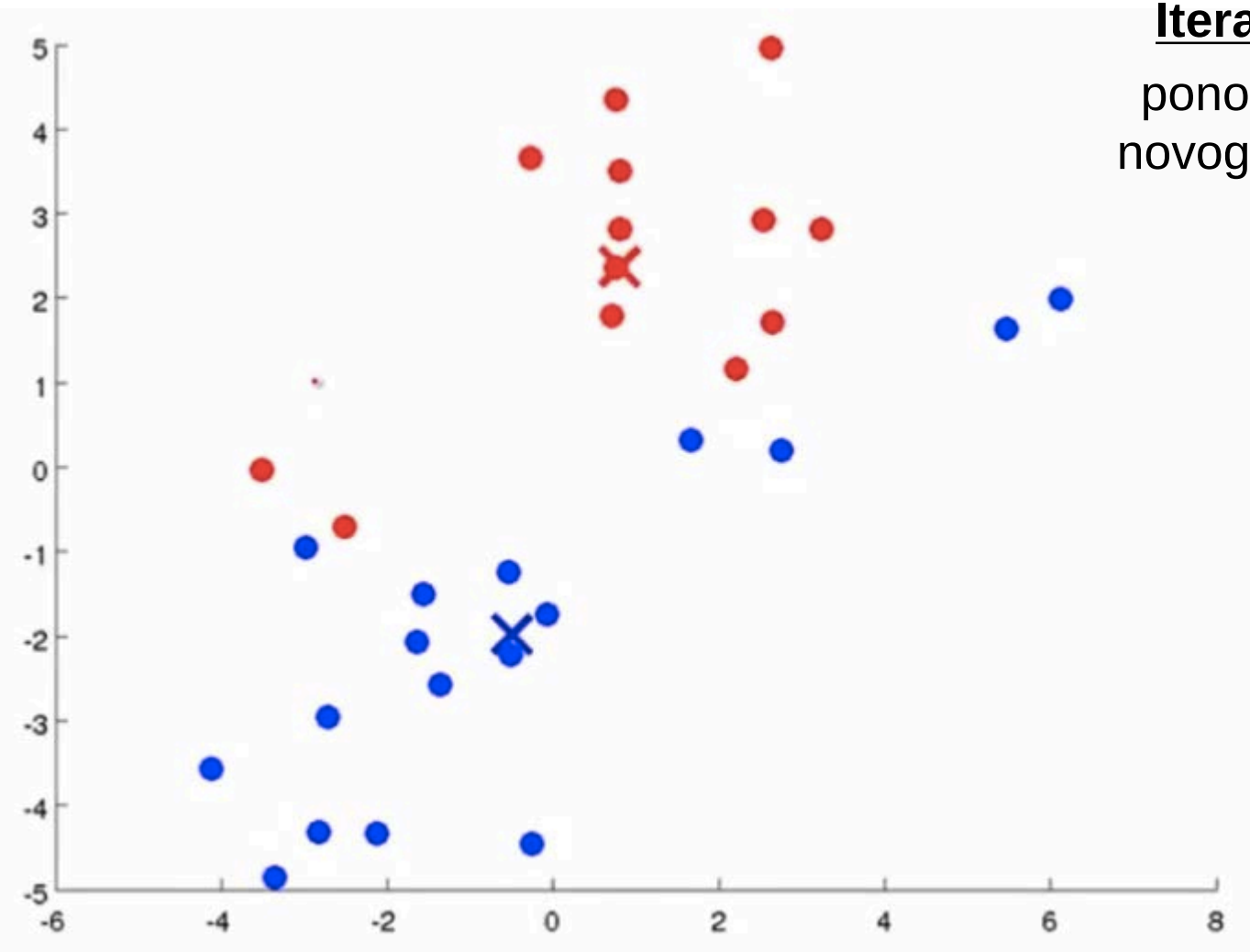
K-MEANS: PRIMER



<u>Iteracija 2, korak 1</u>:

(ponovno) razvrstavanje instanci po klasterima na osnovu udaljenosti od težišta klastera

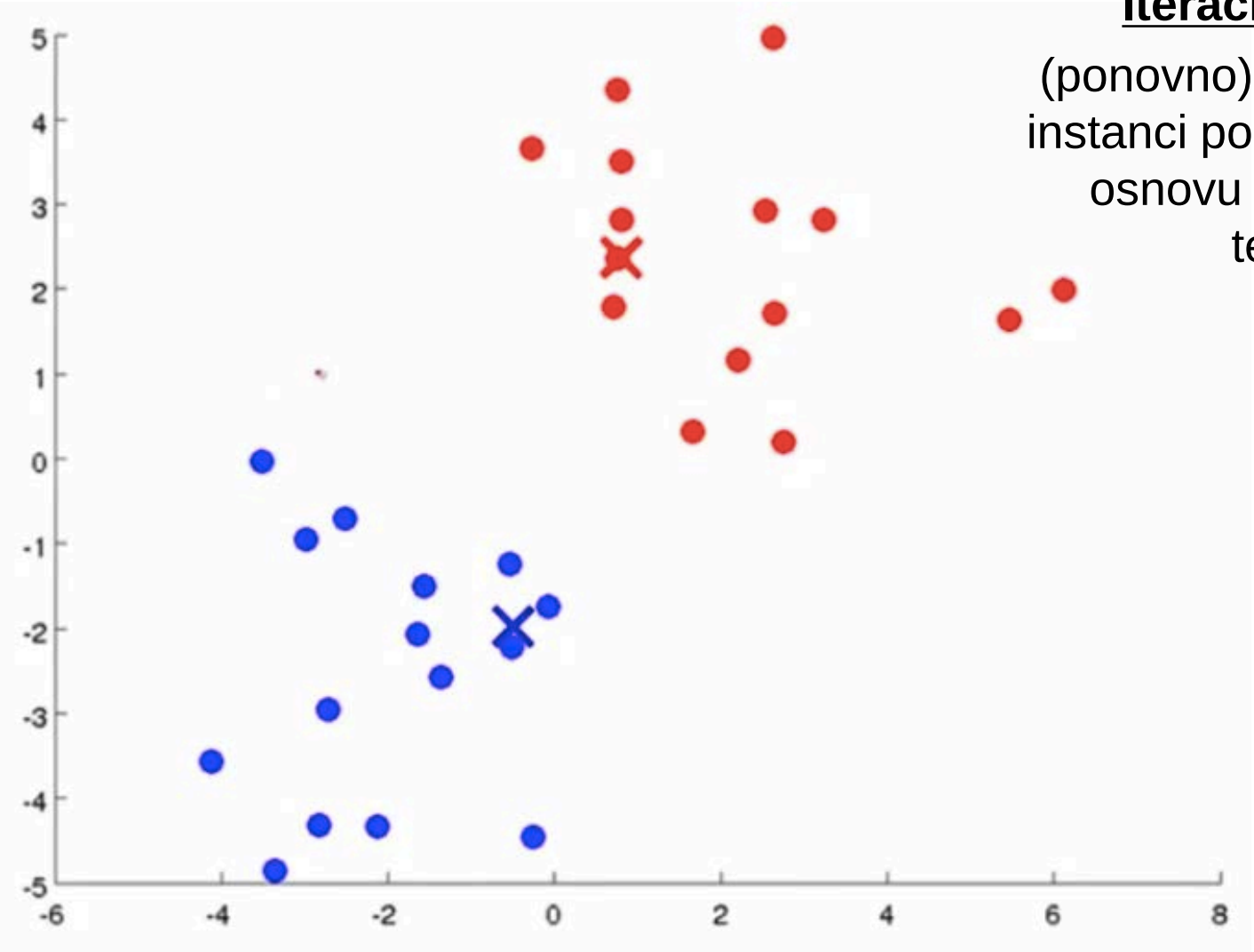
K-MEANS: PRIMER



Iteracija 2, korak 2:

ponovno određivanje novog težišta za svaki klaster

K-MEANS: PRIMER



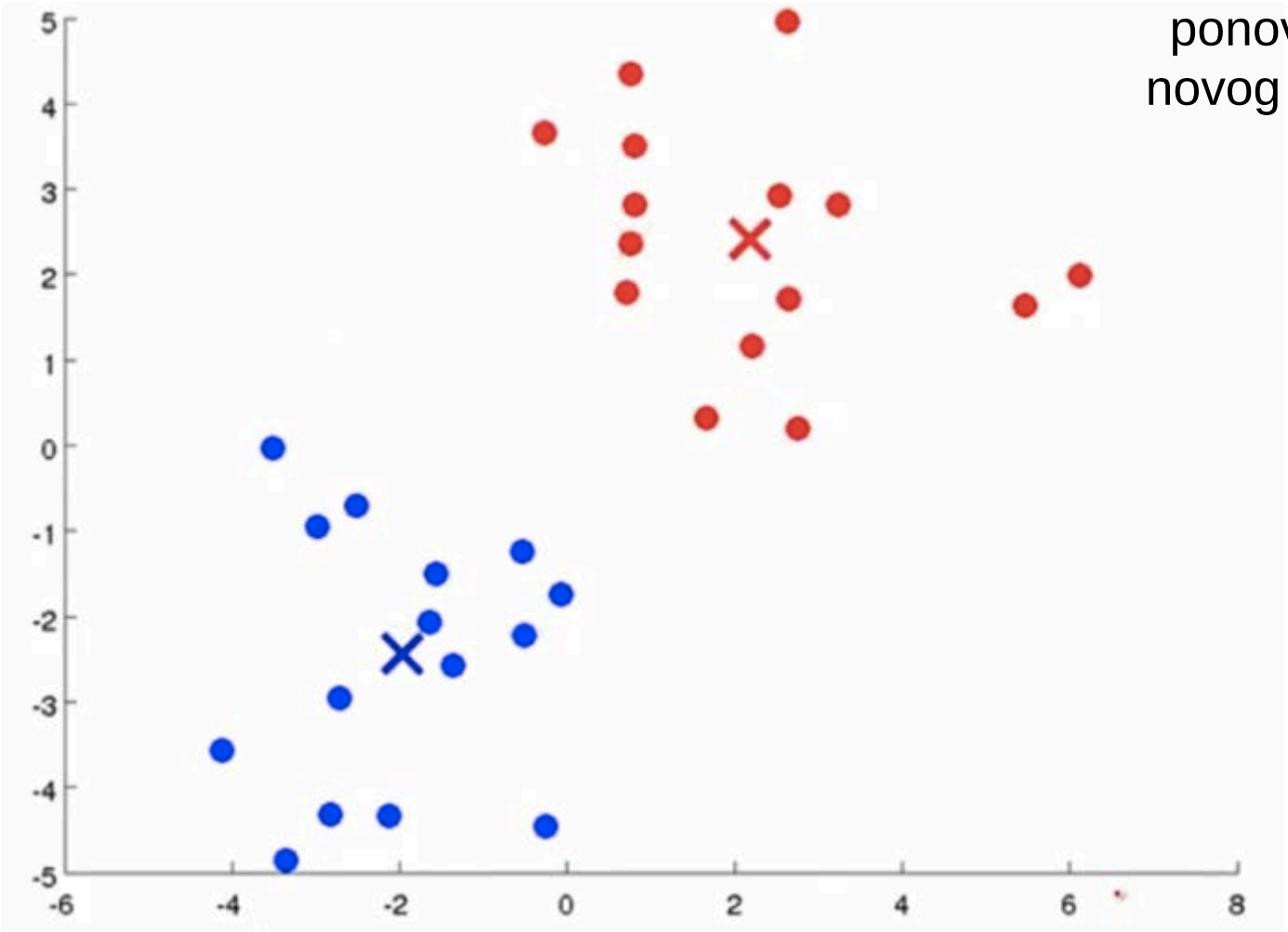
Iteracija 3, korak 1:

(ponovno) razvrstavanje instanci po klasterima na osnovu udaljenosti od težišta klastera

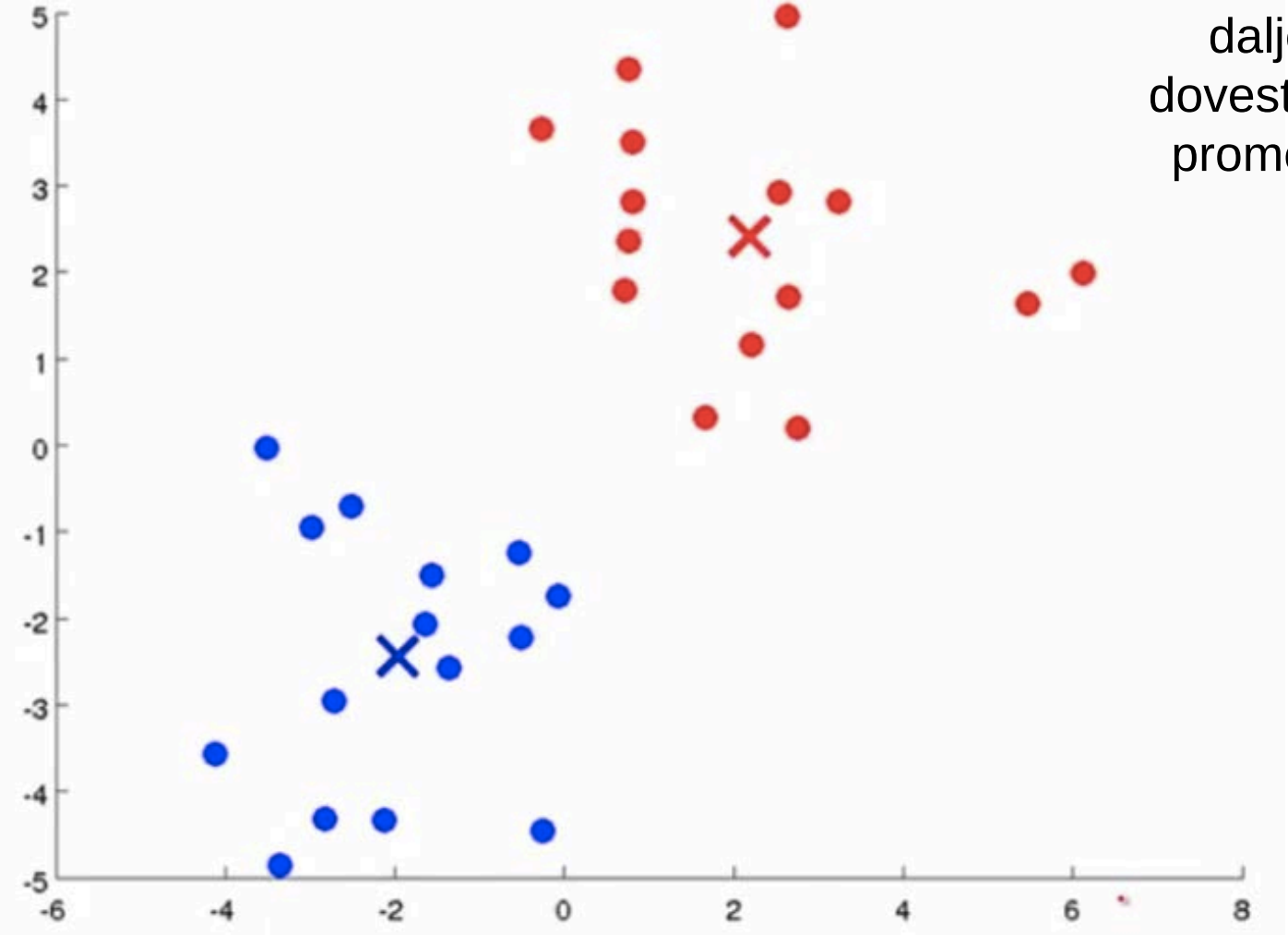
K-MEANS: PRIMER

<u>Iteracija 3, korak 2</u>:

ponovno određivanje novog težišta za svaki klaster



K-MEANS: PRIMER



Algoritam konvergira:

dalje iteracije neće dovesti do značajnijih promena i proces se zaustavlja

K-MEANS: ALGORITAM

Ulaz:

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

K-MEANS: ALGORITAM

Koraci:

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

2) Ponoviti
   1) *Grupisanje po klasterima*: za svaku instancu iz skupa podataka, *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 instanci koje su dodeljene tom klasteru

   dok algoritam ne konvergira ili broj iteracija <= *max*

K-MEANS ALGORITAM: FUNKCIJA KOŠTANJA

Cilj 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, i=1,m

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

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

μ_j – težište klastera j, j=1,K

Ova funkcija se zove i funkcija distorzije (distortion function)

K-MEANS ALGORITAM: FUNKCIJA KOŠTANJA

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

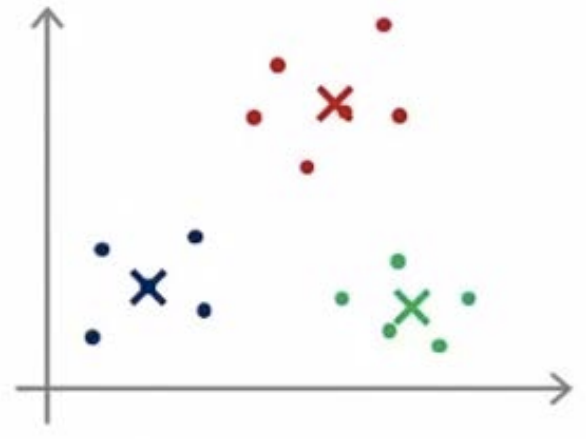
Minimizacija funkcije koštanja ***J*** kroz K-means algoritam:

- faza *Grupisanja po klasterima* minimizuje ***J*** po parametrima $c^{(1)},\ldots,c^{(m)}$, držeći $\mu_1,\ldots,\mu_K$ fiksnim
- faza *Pomeranja težišta* minimizuje ***J*** po parametrima $\mu_1,\ldots,\mu_K$, držeći $c^{(1)},\ldots,c^{(m)}$ fiksnim

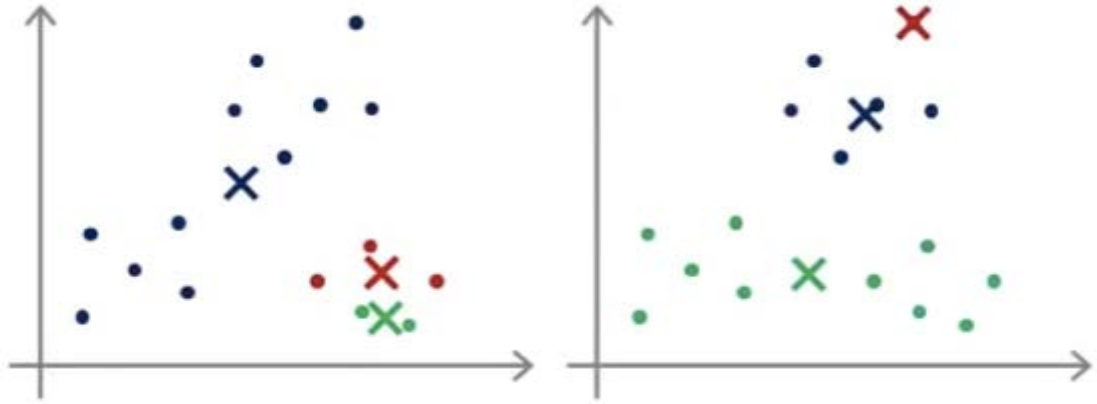
K-MEANS:
PROBLEM INICIJALNOG IZBORA TEŽIŠTA

Zavisno od inicijalnog izbora težišta:

- K-means algoritam može konvergirati brže ili sporije
- Takodje, može “upasti” u lokalni minimum funkcije koštanja i dati loše rešenje



Idealna 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++: BOLJI PRISTUP INICIJALIZACIJE ALGORITMA

Osnovna ideja: inicijalno odabrati *k* težišta koja su što dalje jedna od drugog

Postupak se sastoji u sledećem:

1. Nasumično izabrati prvo težište među instancama
2. Za svaku instancu izračunati njenu udaljenost od prethodno izabranih težišta
3. Nasumično izabrati sledeće težište među instancama koje su najviše udaljene od njima najbližih, prethodno izabranih težišta
4. Ponoviti korake 2 i 3 dok se ne uzorkuje *k* težišta

K-MEANS: KAKO ODREDITI 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*

*Na primer, korišćenjem *within cluster sum of squared errors* metrike

K-MEANS: KAKO ODREDITI K?

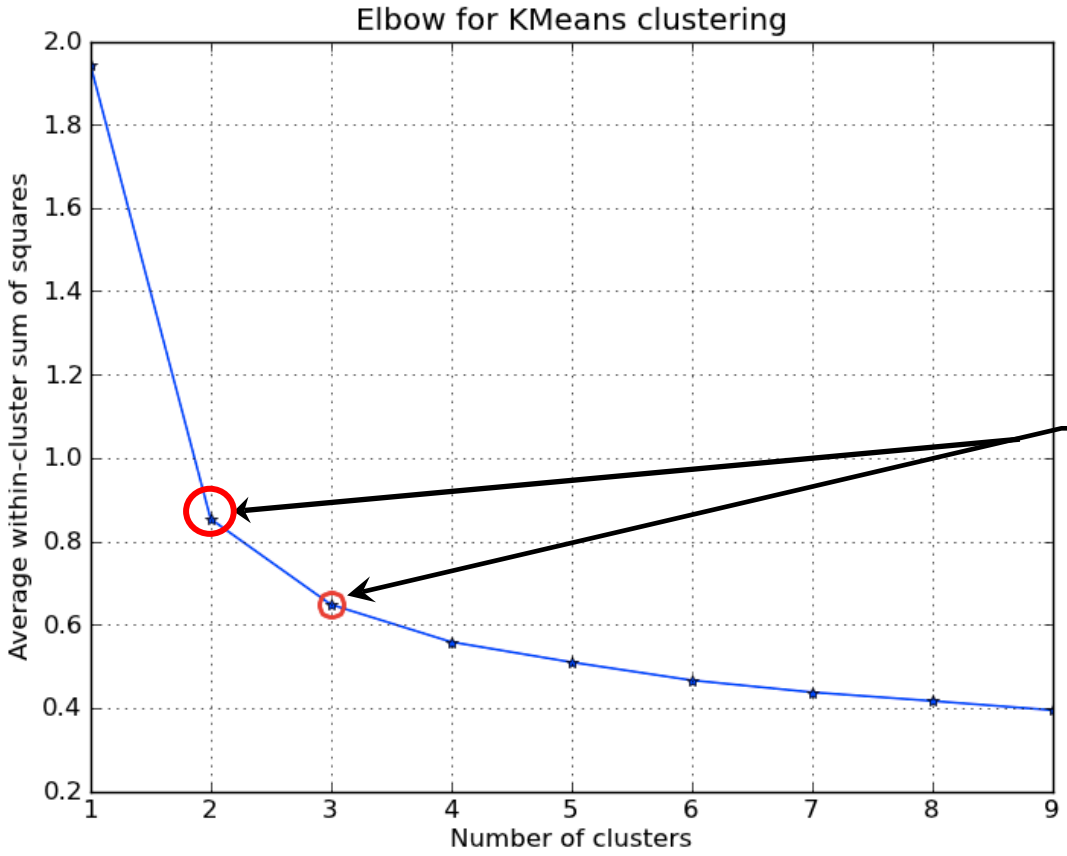
Ukoliko ne posedujemo znanje o fenomenu/pojavi

- 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
- Ova metoda je poznata kao Elbow metoda

*Na primer, korišćenjem *within cluster sum of squared errors* metrike

K-MEANS: KAKO ODREDITI K?

ELBOW METODA



Kandidati za optimalan broj klastera (K)

Izvor slike: http://i.stack.imgur.com/BzwBY.png

KLASTERIZACIJA

JELENA JOVANOVIĆ

jelena.jovanovic@fon.bg.ac.rs