K means klastering

Primenjeni algoritmi

Uvod

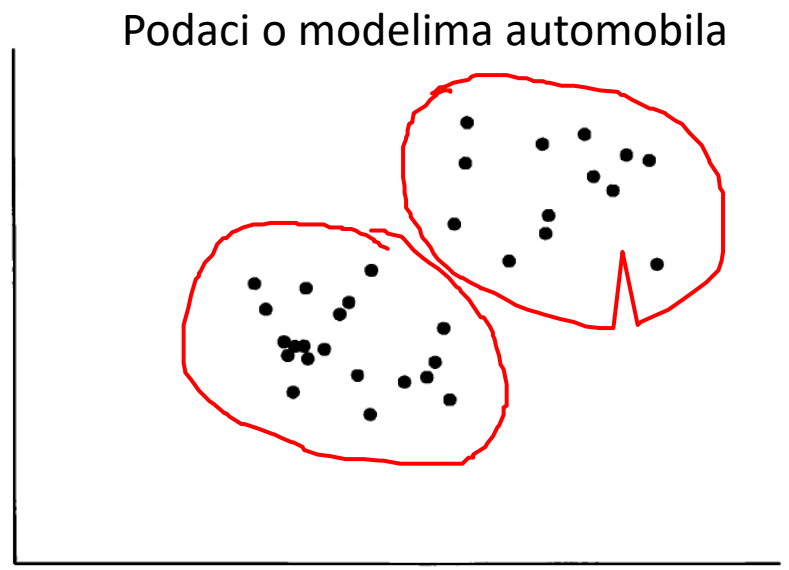
- Nenadgledana tehnika učenja
- Postoji veliki broj tačaka predstavljenih vektorima (elementi vektora su atributi) koje nisu klasifikovane ili označene
- Cilj je pametno grupisati tačke
- Svaka grupa je asocirana svojim centroidom, tj. težištem
- Koliko ima takvih grupa i gde su im težišta?

Primena K means podele

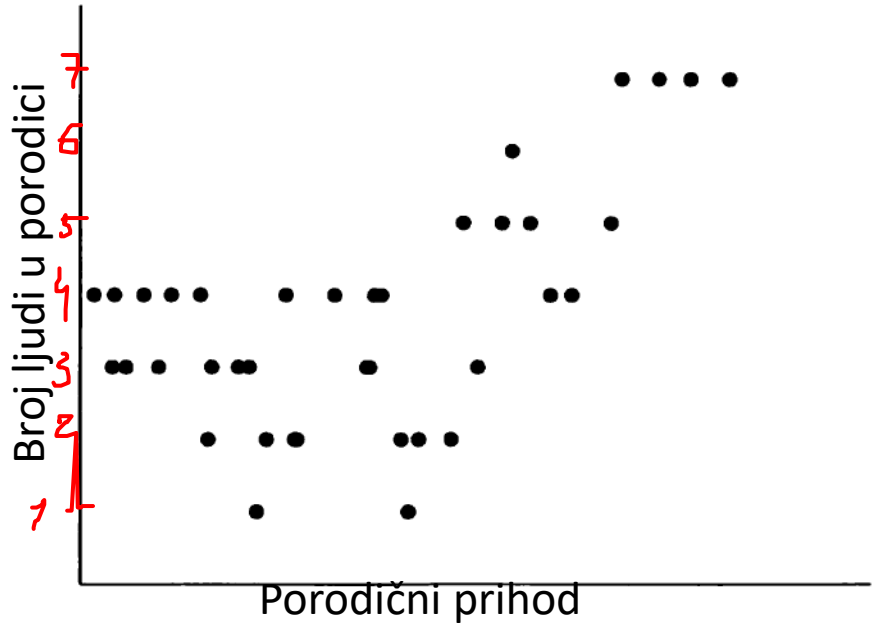
- Grupisanje neobeleženih podataka na osnovu sličnosti njihovih osobina
- Primer: Klasifikacija kupaca prema istoriji kupovine, svaka karakteristika može biti trošak za različite vrste robe
- Primer: Optimalan raspored parkirališta u gradu
- Primer: Optimizacija veličine vrata i dužine ruku košulja
- Primer: Grupisanje sličnih slika, bez prethodnog klasifikovanja

Vrste problema klasifikacije

1. Traženje strukture unutar skupova neklasifikovanih podataka, sa pretpostavkom o kategorijama.
   - Primer: podaci o modelima automobila
     - cene, efikasnost , goriva, veličina točkova, snaga zvučnika itd.



2. Veštačko deljenje podataka čak i ako ne postoji očigledno grupisanje
   - Proizvođač escajga želi da pakuje pibor za jelo – kriterijum: broj viljušaka i noževa i „otmenost“(cena) pribora



K means algoritam

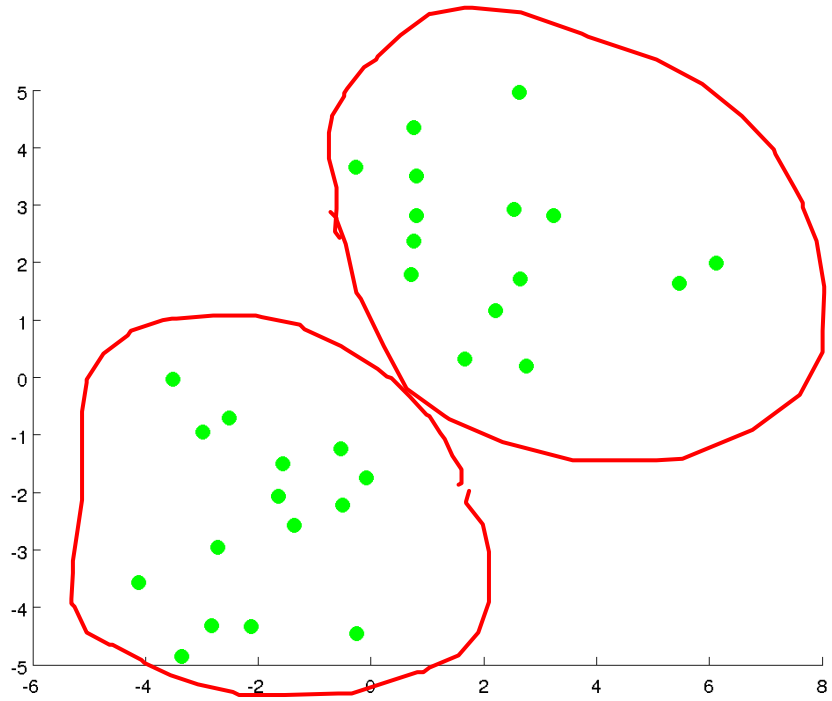
Ulazi:

- K – broj klastera
- Obučavajući skup $\{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$

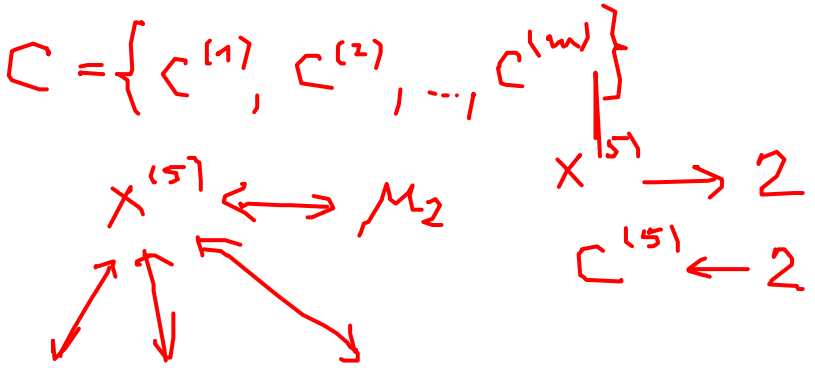
$$x^{(i)} = \begin{bmatrix} x_1^{(i)} \\ x_2^{(i)} \\ \vdots \\ x_n^{(i)} \end{bmatrix}_{n \times 1}, x^{(i)} \in \mathbb{R}^n$$

Izlaz:

- Grupisane tačke $\{c^{(1)}, c^{(2)}, \dots, c^{(m)}\}$



K means algoritam



K-MEANS (X,K)

1 *inicijalizacija*: slučajan izbor K težišta $\mu_1, \mu_2, \dots, \mu_K \in \mathbb{R}^n$

2 **repeat** (do konvergencije težišta)

3 **for** i = 1 **to** *m*

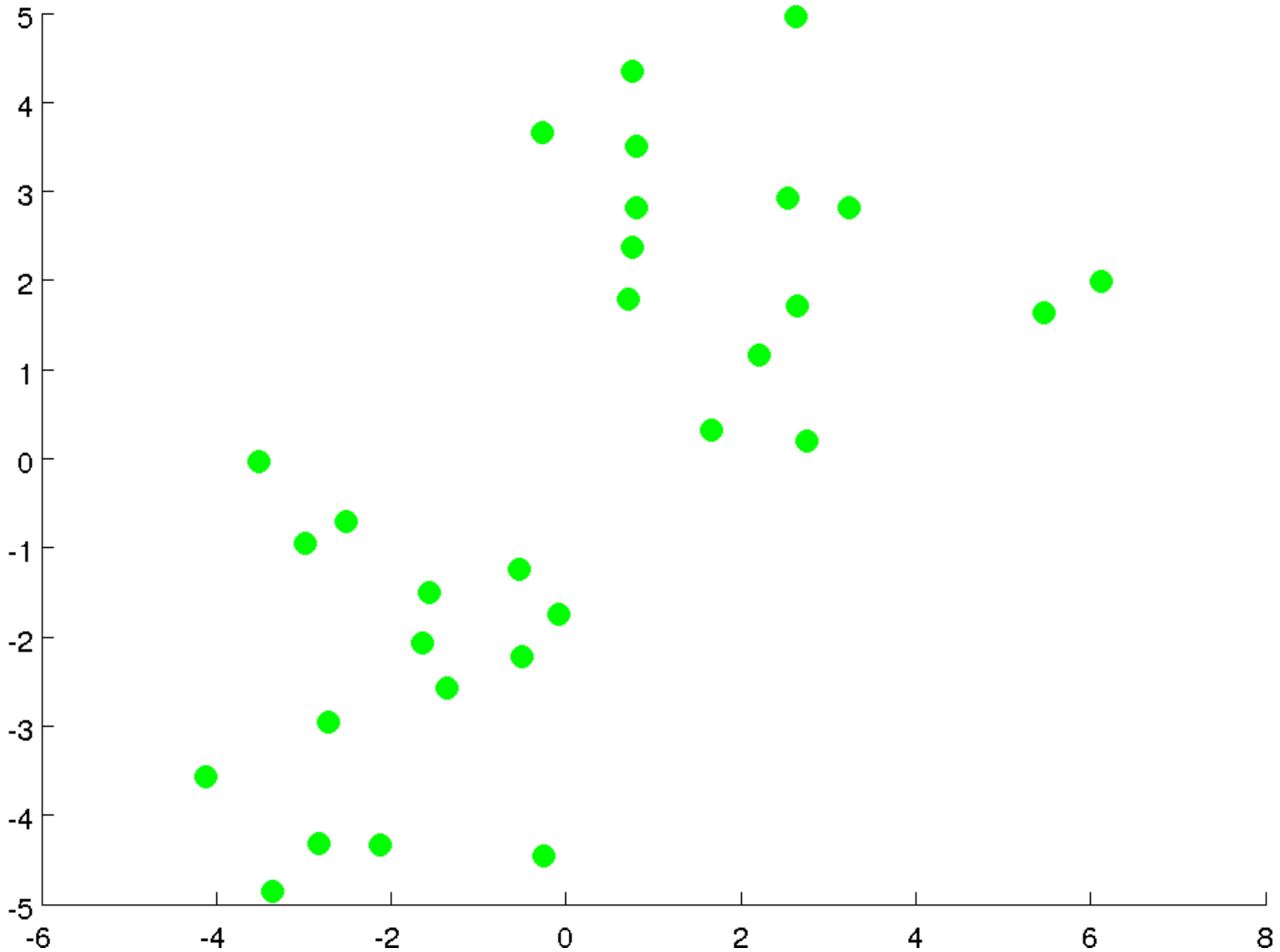
4 $c^{(i)}$ = indeks klastera sa težištem najbližim $x^{(i)}$

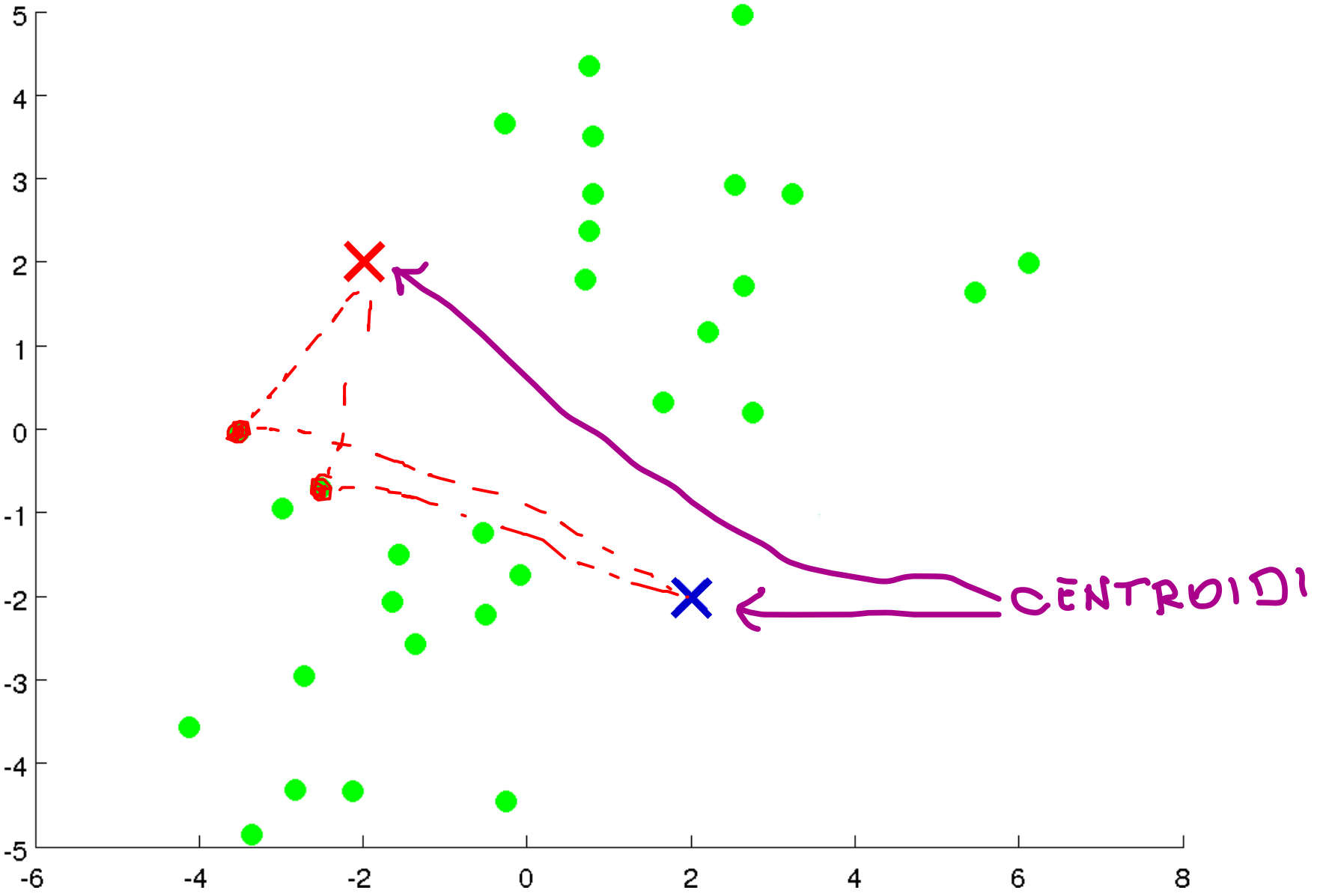
5 **for** k = 1 **to** *K*

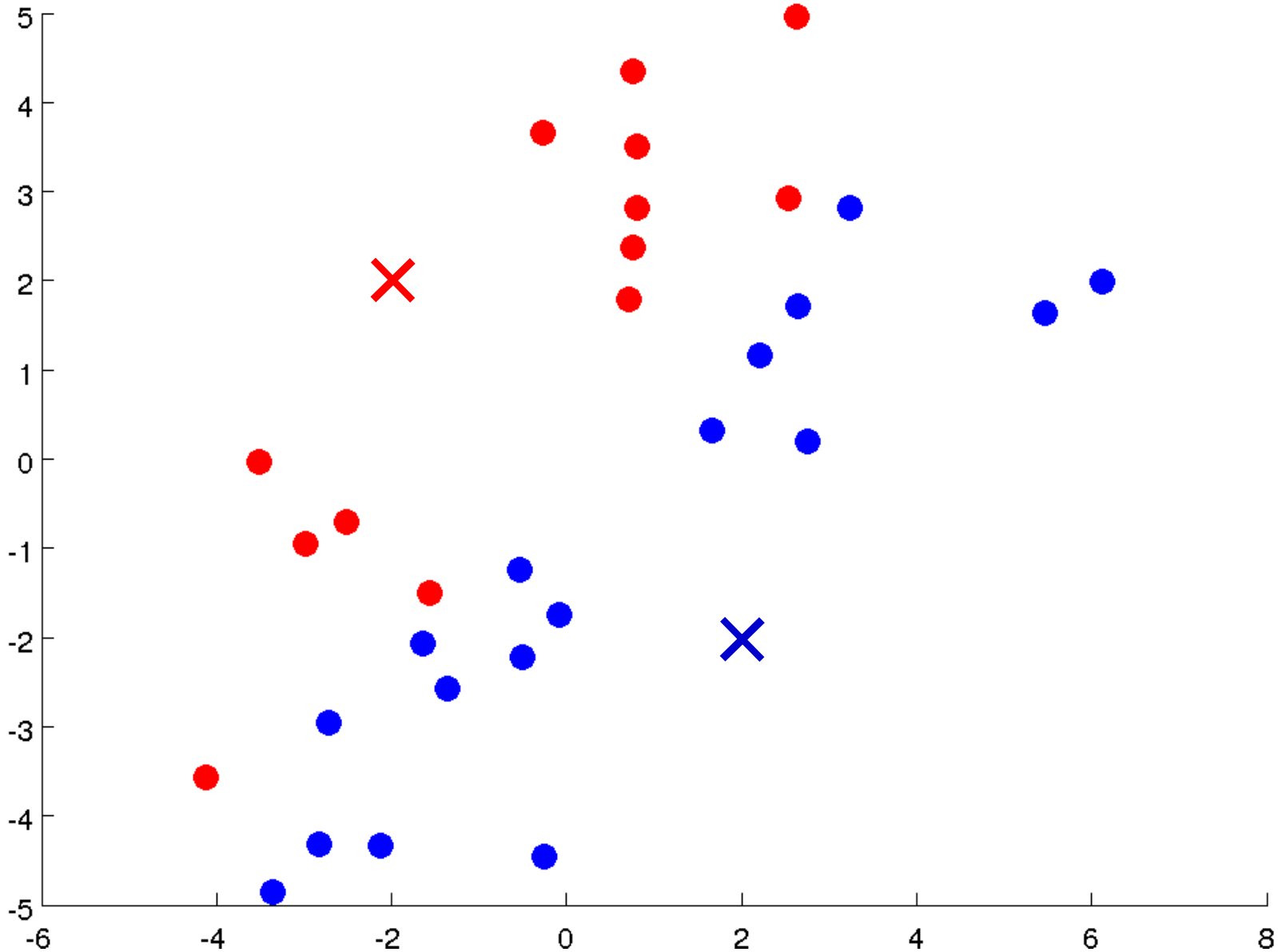
6 μ_k = težište za tačke iz klastera *k*

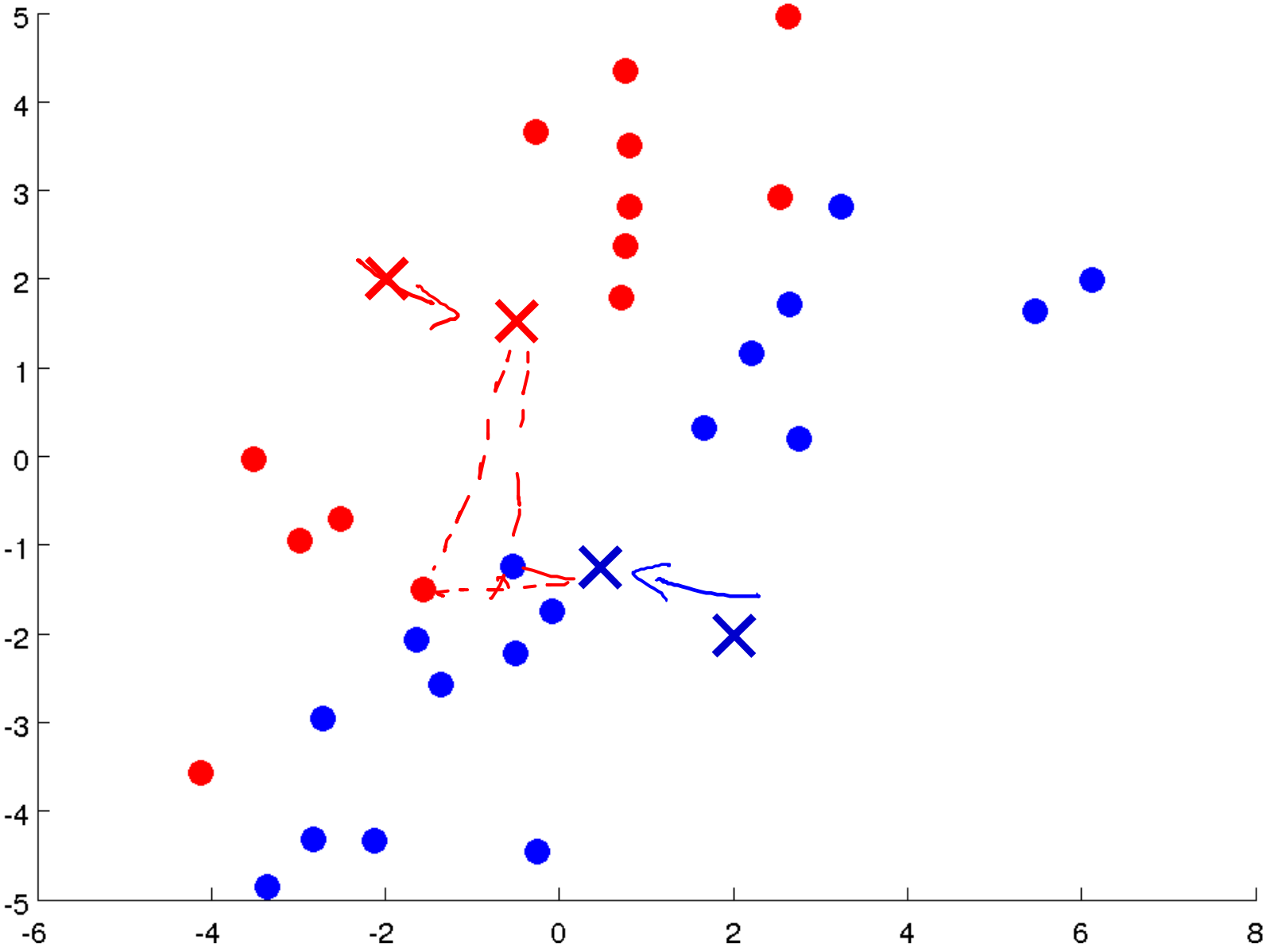
7 **return** $C = \{c^{(1)}, c^{(2)}, \dots, c^{(m)}\}$

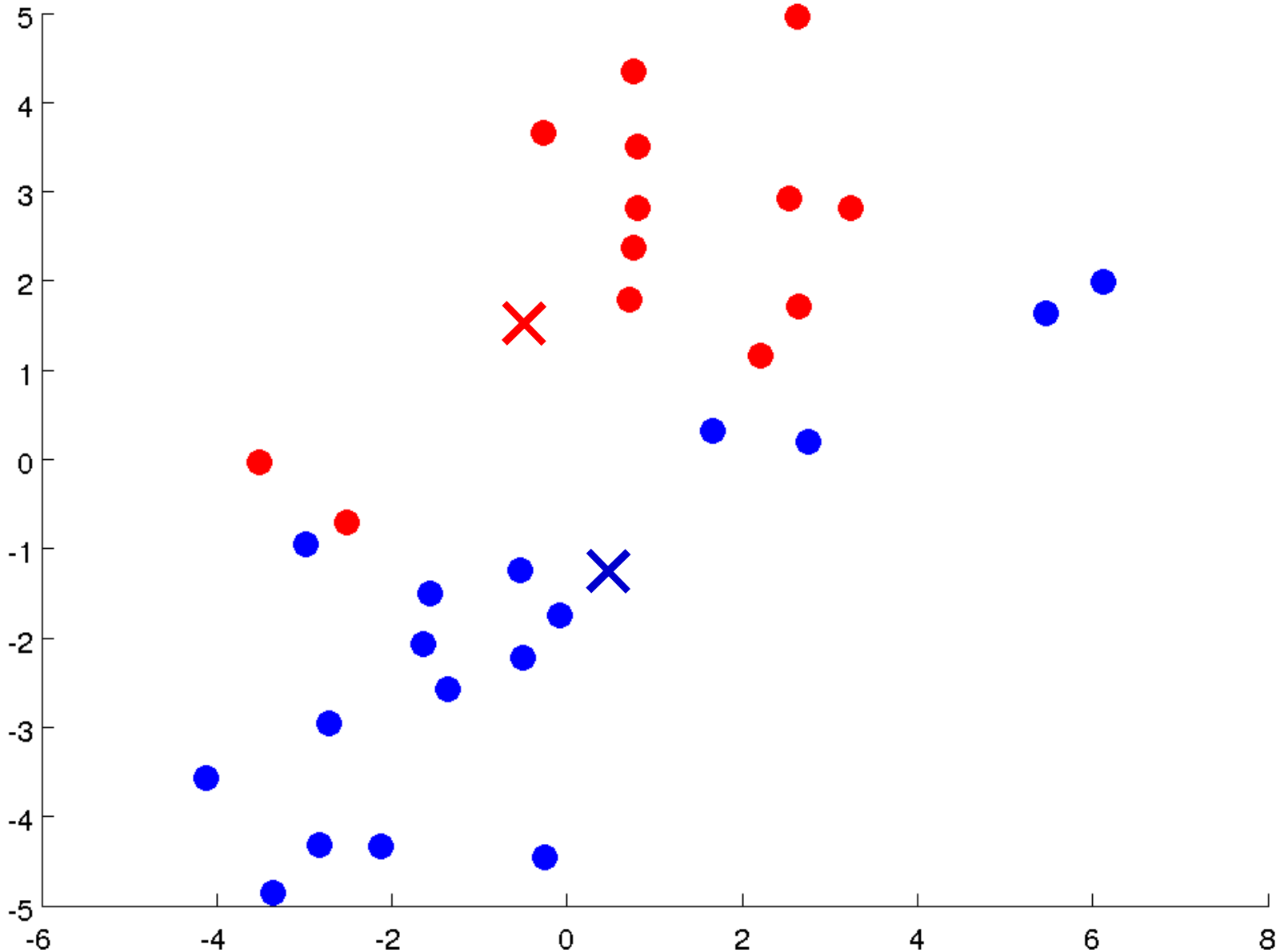
K=2

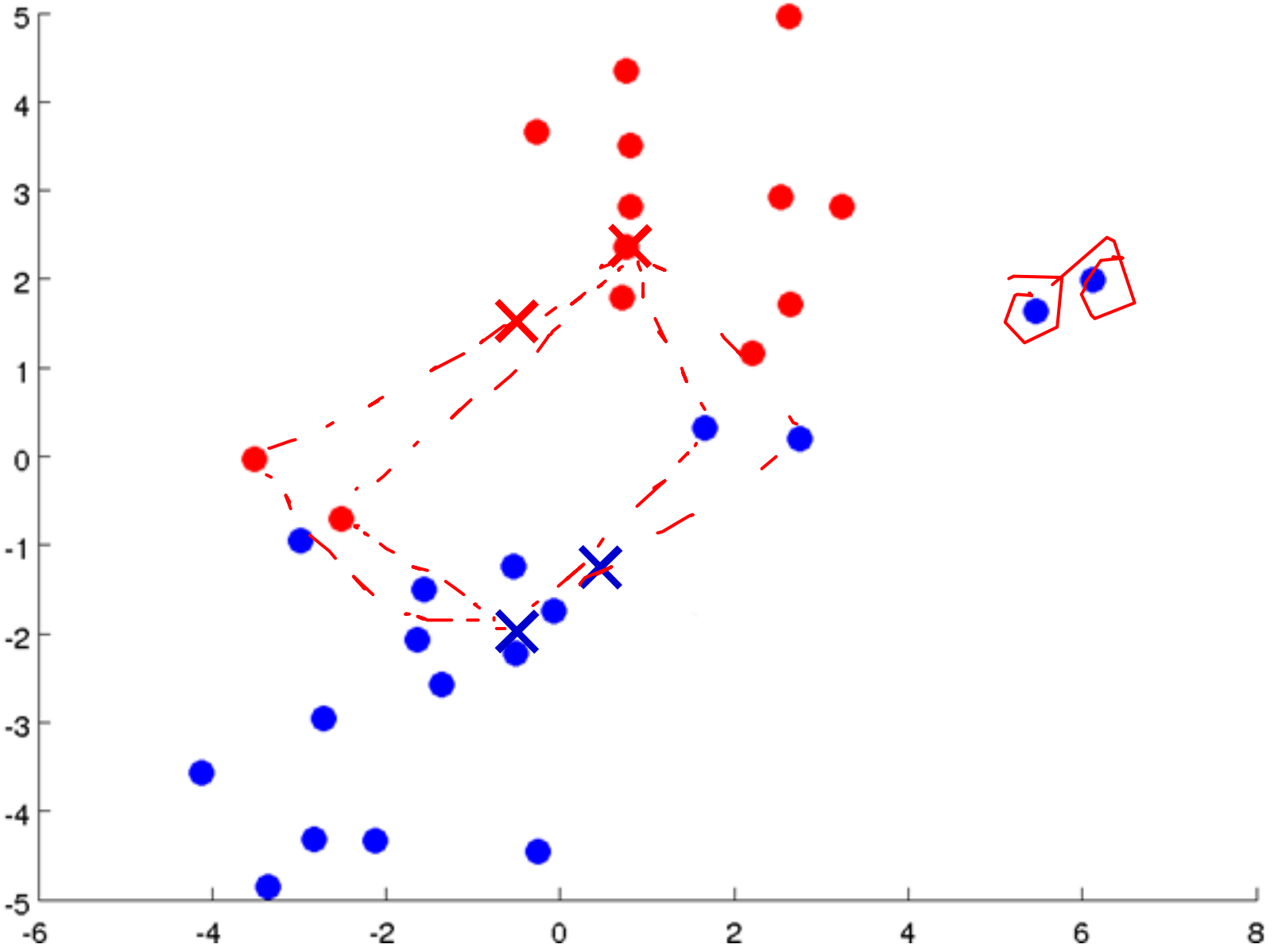


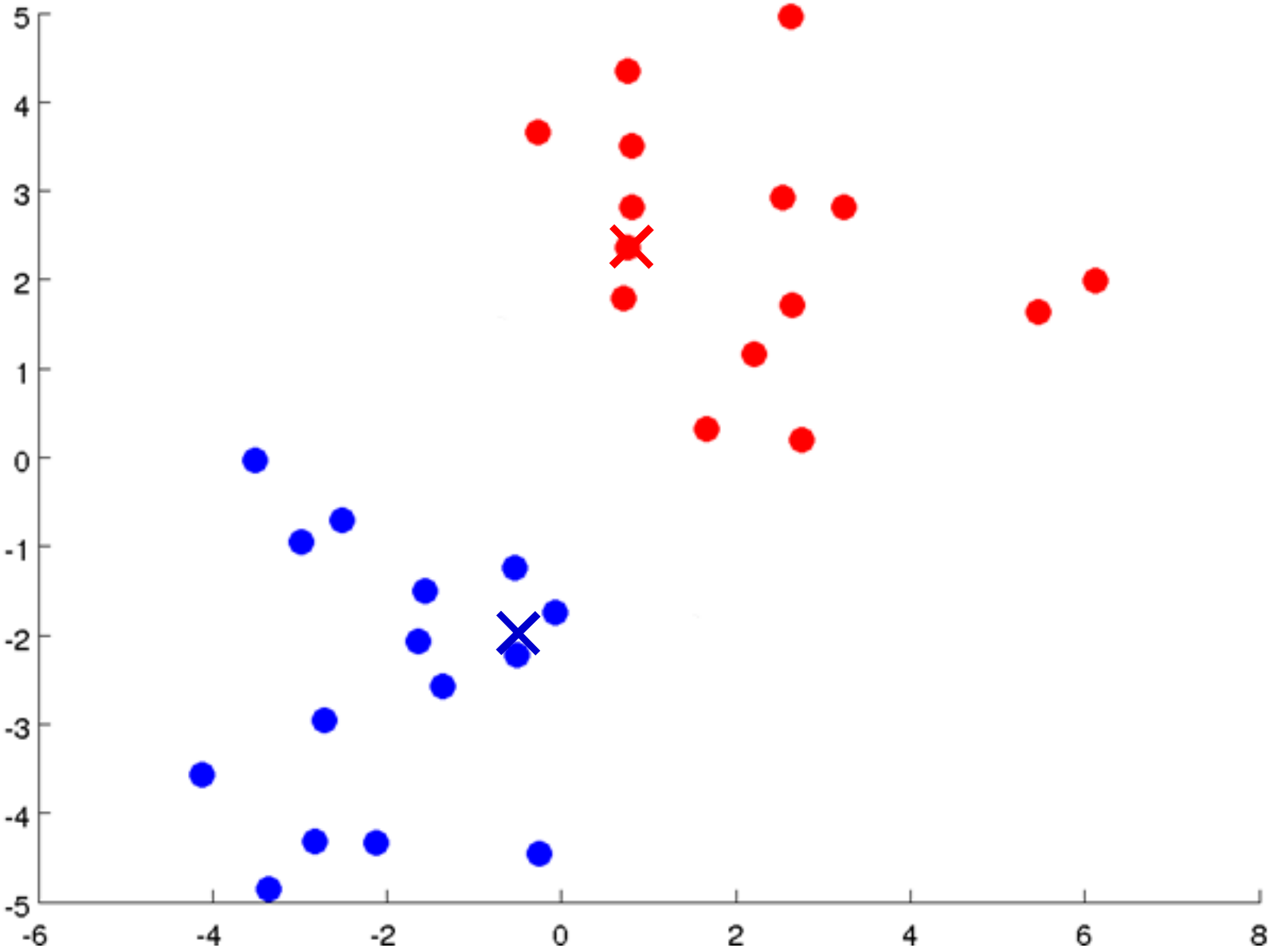

5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8
CENTROIDI

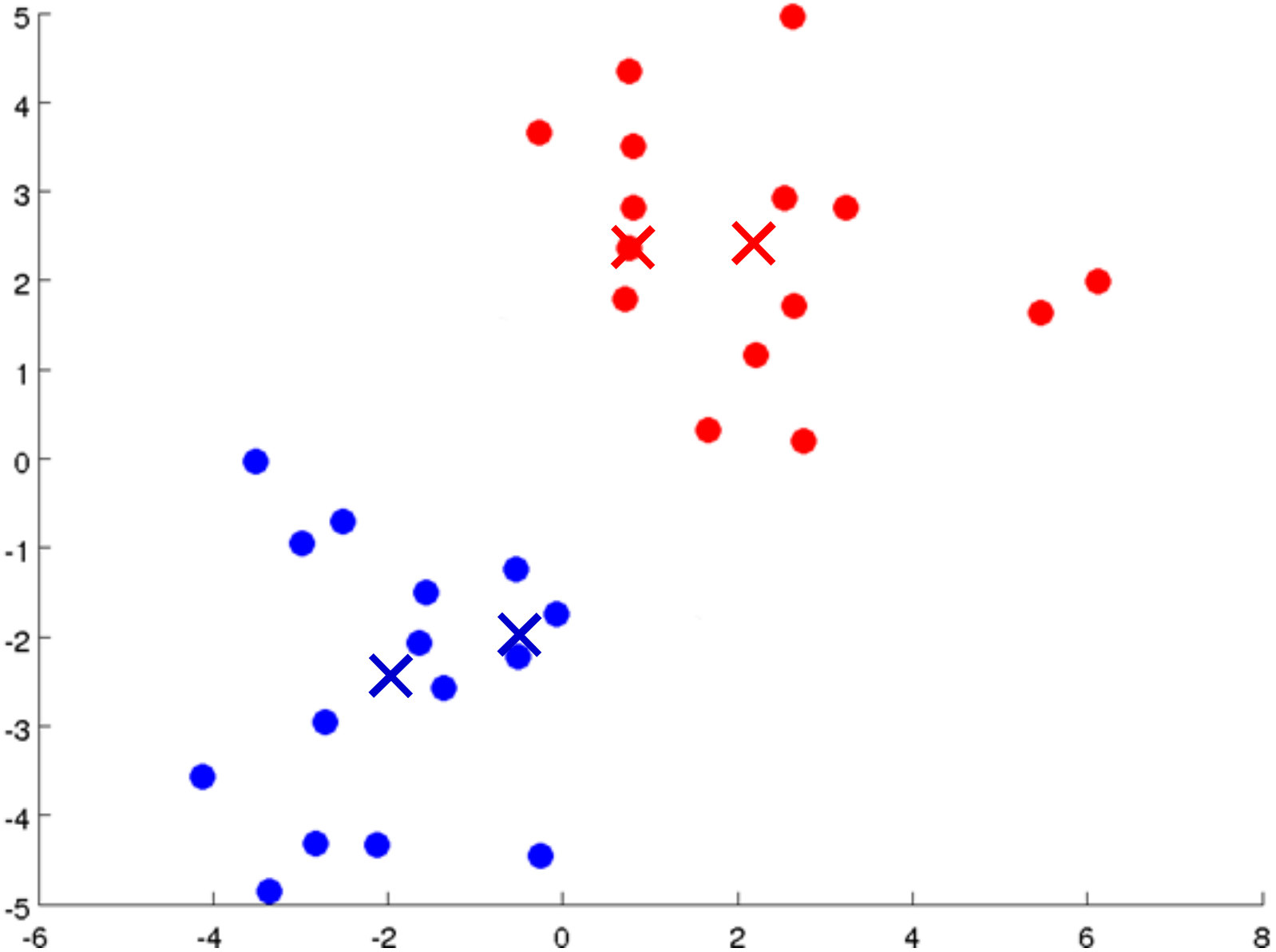

5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8

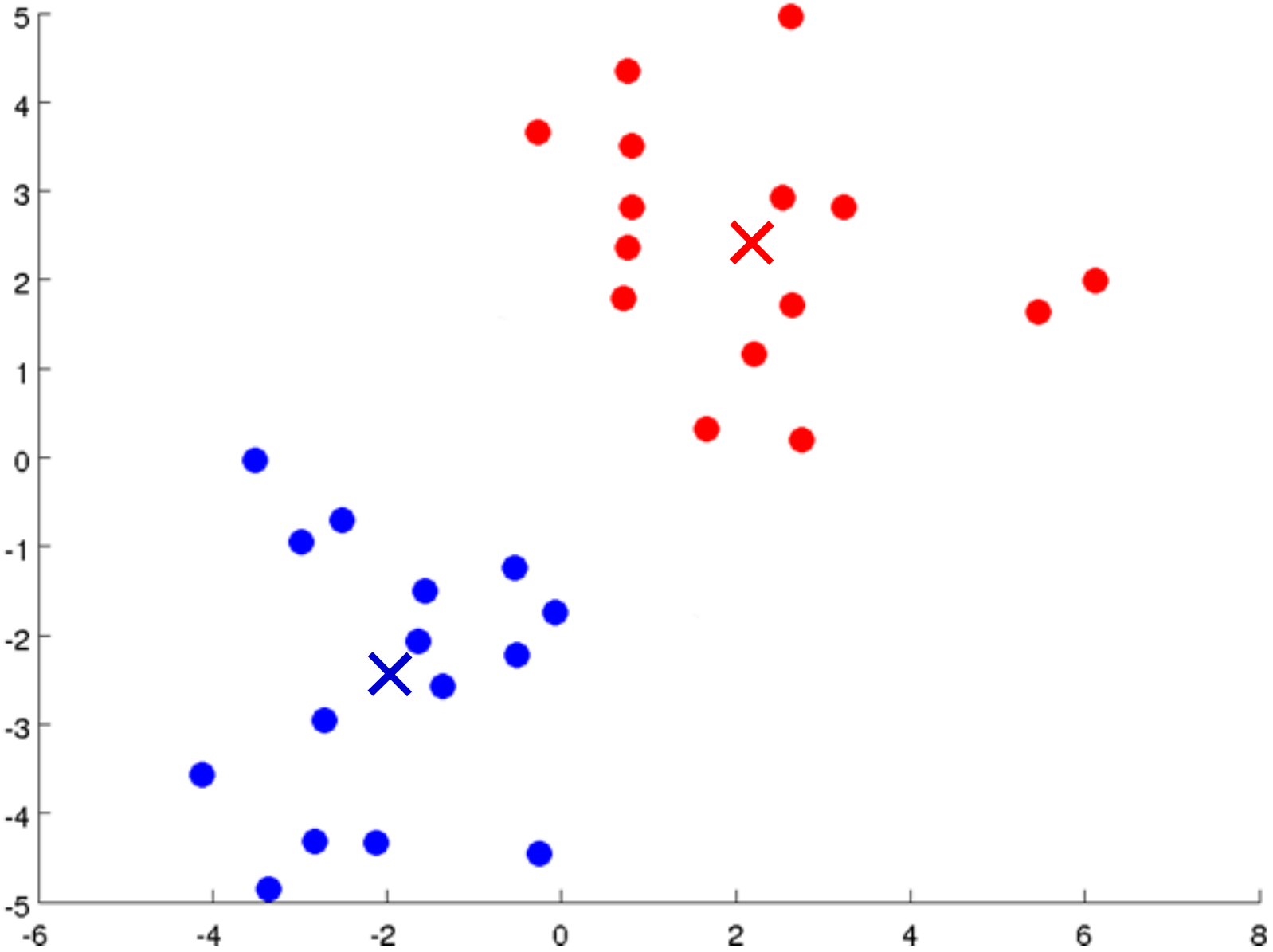

5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8


5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6
8

K means algoritam

- Korak 0: Skaliranje podataka
- Korak 1: Slučajan izbor K težišta $\mu_1, \mu_2, \dots, \mu_K$
- Korak 2: Odrediti udaljenost $D^{(i,k)}$ tačaka (vektora $x^{(i)}$) do težišta svakog klastera *k*

$$D^{(i,k)} = \left\| x^{(i)} - \mu_k \right\| = \sqrt{\sum_{j=1}^{n} \left(x_j^{(i)} - \mu_{j(k)} \right)^2} \qquad k = 1,2, \dots, K$$

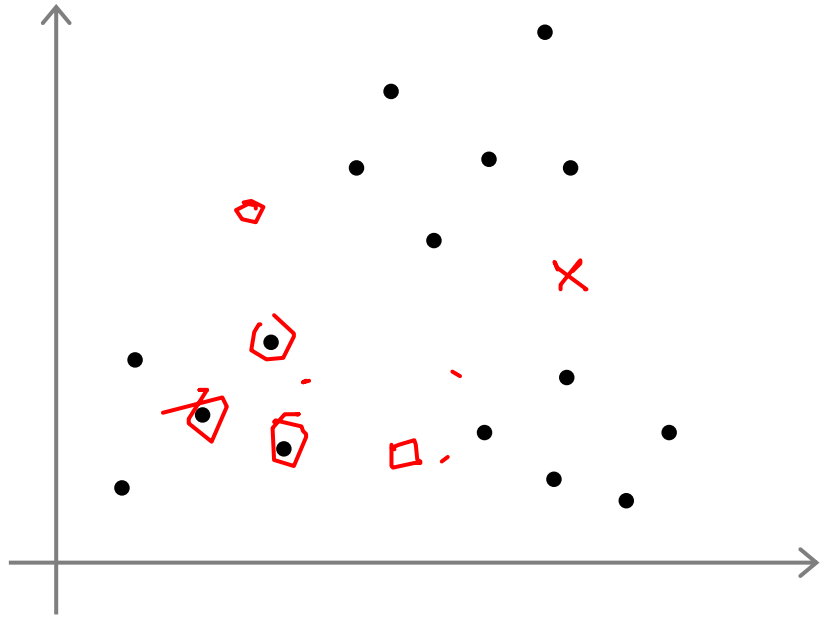
  - Svaka tačka se pridružuje najbližem klasteru

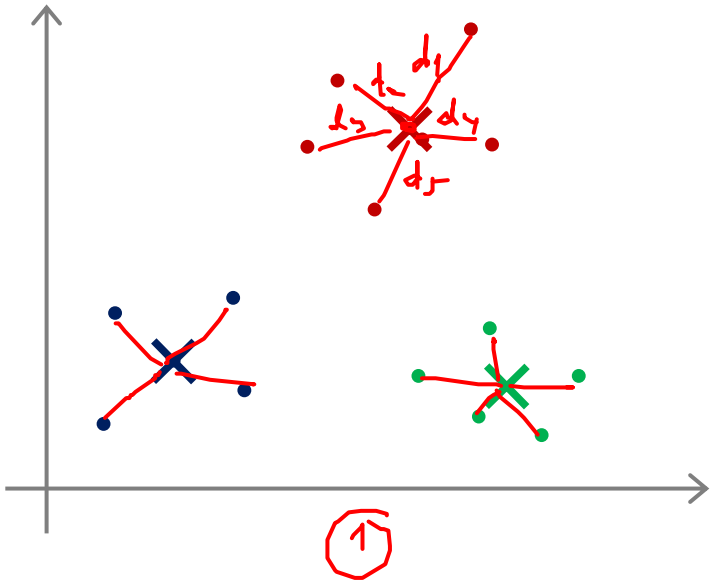
$$c^{(i)} = \underset{k}{\operatorname{argmin}}\, D^{(i,k)}$$

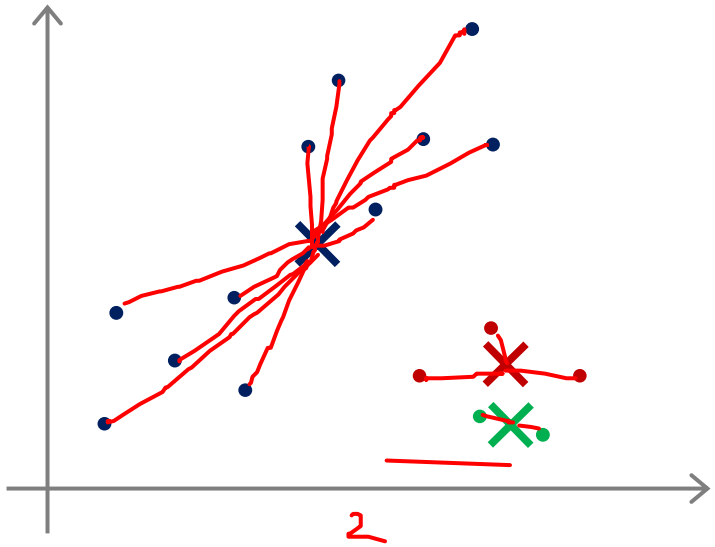
- Korak 3. Pronaći novih K težišta za formirane klastere.
  - Povratak na korak 2 i ponavlja se postupak sve do konvergencije težišta.

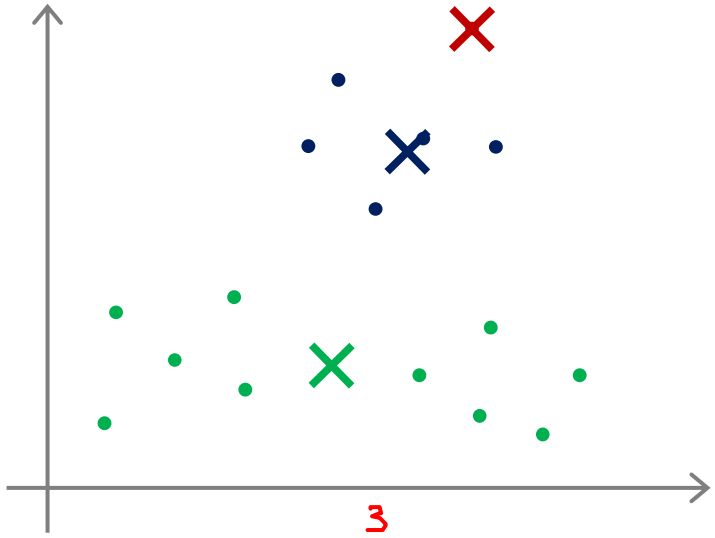
Inicijalizacija – slučajan izbor centroida

- izbor K težišta na slučajan način ($K < m$)
  - Težišta su iz skupa X
  - Težišta su slučajno izabrane tačke iz $\mathbb{R}^n$
    - sa osobinama koje odgovaraju podacima: sr.vred i stdev, min, max)







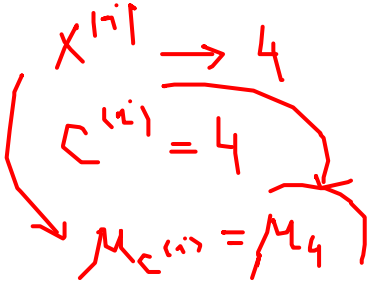


Poređenje rezultata

μ_k – centroid (težište) za tačke iz klastera *k*

$c^{(i)}$ - indeks klastera kome je trenutno pridružen $x^{(i)}$

$\mu_{c^{(i)}}$ - centroid klastera kome je trenutno pridružen $x^{(i)}$



- Kriterijum optimalnosti

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

$$\min_{c,\mu} J$$

$$v = \begin{bmatrix} 2 \\ 1 \\ -4 \\ -2 \end{bmatrix} \qquad \|v\| = \sqrt{\sum_{i=1}^{4} v_i^2} = \sqrt{4+1+16+4} = \sqrt{25} = 5$$

Višestruka slučajna inicijalizacija – izbor optimuma

for i = 1 to 100

→ Randomly initialize K-means.

$c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K$ ← run K-means

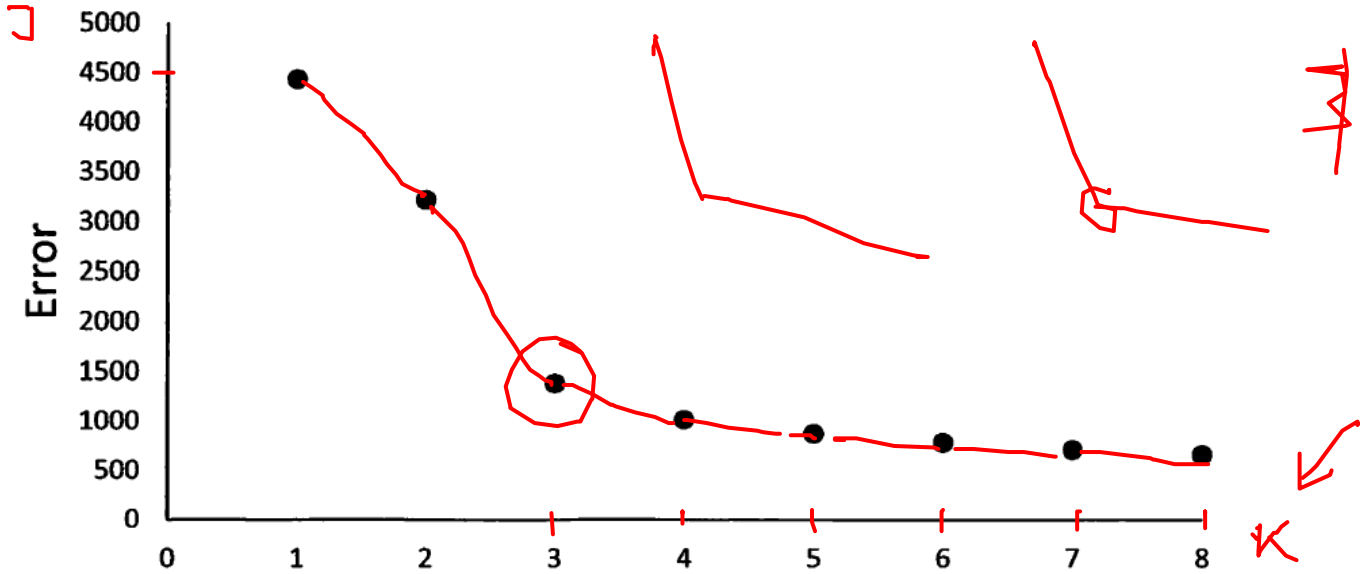
Compute $J\left(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K\right)$

Izabrati podelu koja daje minimalno $J(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_K)$

Primer: Kriminal u Engleskoj

- 13 dimenzija
  - Provala u drugu zgradu osim stana
  - Provala u stan
  - Krivična šteta
  - Prekršaji droge
  - Prevara i falsifikat
  - Prekršaji u vezi vozila
  - Ostali prestupi
  - Ostala krivična dela
  - Pljačka
  - Seksualni prestupi
  - Nasilje nad osobom - sa povredom
  - Nasilje nad osobom - bez povreda

Local Authority	Burglary in a building other than a dwelling	Burglary in a dwelling	Criminal damage	Drug offences	Fraud and forgery	Offences against vehicles	...	Population	Population per Square Mile
Adur	280	120	708	158	68	382	...	58500	3610
Allerdale	323	126	1356	392	79	394	...	96100	198
Alnwick	94	33	215	25	11	71	...	31400	75
Amber Valley	498	367	1296	241	195	716	...	116600	1140
Arun	590	299	1806	471	194	819	...	140800	1651
Ashfield	784	504	1977	352	157	823	...	107900	2543
Ashford	414	226	1144	196	162	608	...	99900	446
Aylesbury Vale	696	377	1490	502	315	833	...	157900	453
Babergh	398	179	991	137	152	448	...	79500	346
Barking & Dagenham	639	1622	2353	1071	1194	3038	...	155600	11862
Barnet	1342	3550	2665	1198	1504	4104	...	331500	9654
Barnsley	1332	860	3450	1220	322	1661	...	228100	1803
Barrow-in-Furness	190	134	1158	179	59	227	...	70400	2339
Basildon	756	1028	1906	680	281	1615	...	164400	3874
Basingstoke & Deane	1728	598	426	930	182	1159	...	147900	605



Primer: Kriminal u Engleskoj

- 3 klastera je već interesantno
  - Klaster 1 – jedna tačka - London

	Cluster 1	Cluster 2	Cluster 3
Number in cluster	1	68	273
Burglary in a building other than a dwelling	0.0433	0.0059	0.0046
Burglary in a dwelling	0.0077	0.0079	0.0030
Criminal damage	0.0398	0.0156	0.0114
Drug offences	0.1446	0.0070	0.0029
Fraud and forgery	0.1037	0.0042	0.0020
Offences against vehicles	0.0552	0.0125	0.0060
Other offences	0.0198	0.0018	0.0009
Other theft offences	0.6962	0.0313	0.0154
Robbery	0.0094	0.0033	0.0004
Sexual offences	0.0071	0.0015	0.0008
Violence against the person - with injury	0.0560	0.0098	0.0053
Violence against the person - without injury	0.0796	0.0128	0.0063
Population per Square Mile	4493	10952	1907

K-means – broj klastera (grupa)