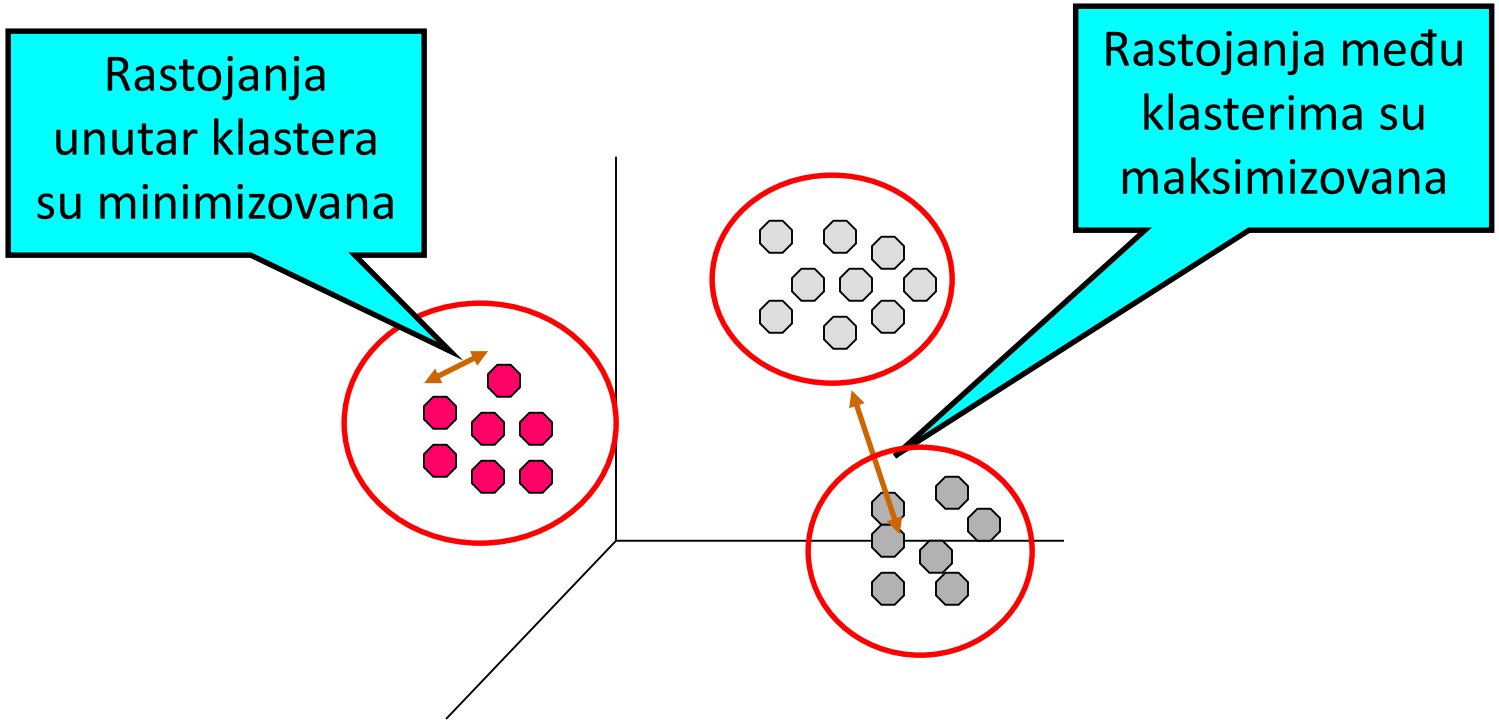
Klaster analiza

Milan M.Milosavljević

Šta je klaster analiza?

- Pronalaženje grupa objekata takvih da su objekti u grupi međusobno slični (ili povezani), i da su objekti u različitim grupama međusobno različiti (ili nepovezani).

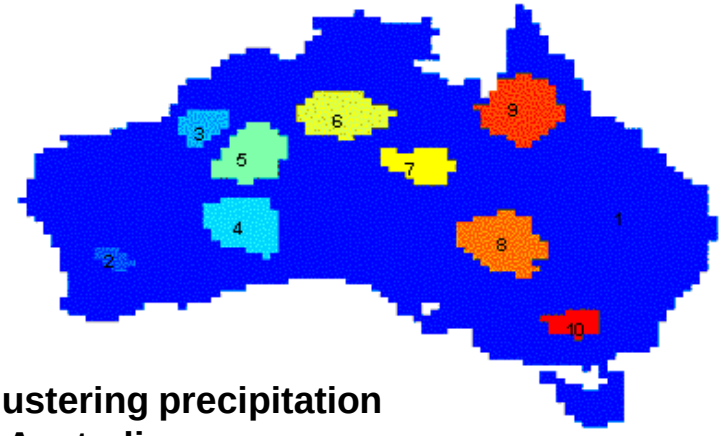


Primena klaster analize

- **Razumevanje**
  - Pregledanje dokumenata iz iste grupe, grupisanje gena i proteina koji imaju iste funkcionalnosti, grupisanje akcija sa sličnim promenama cena

	Discovered Clusters	*Industry Group*
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP	Oil-UP

- **Za dodatnu upotrebu**
  - Sumarizacija
  - Kompresija
  - Efikasno nalaženje najbližih suseda



Clustering precipitation in Australia

Dvosmislenost pojma klastera

Koliko klastera?

Šest klastera

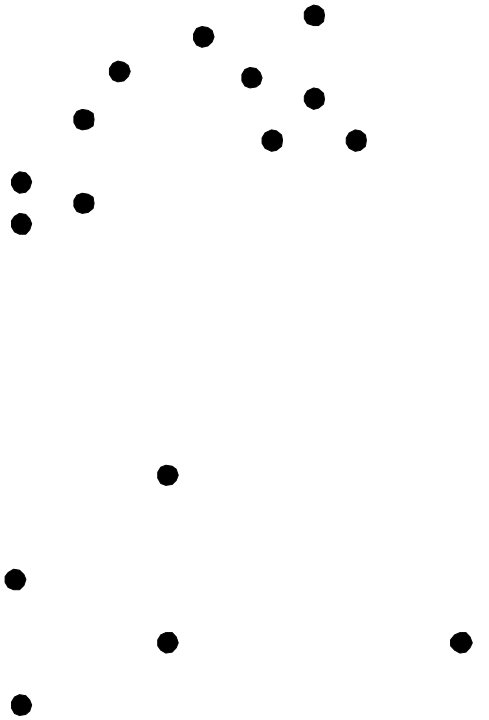
Dva klastera

Četiri klastera

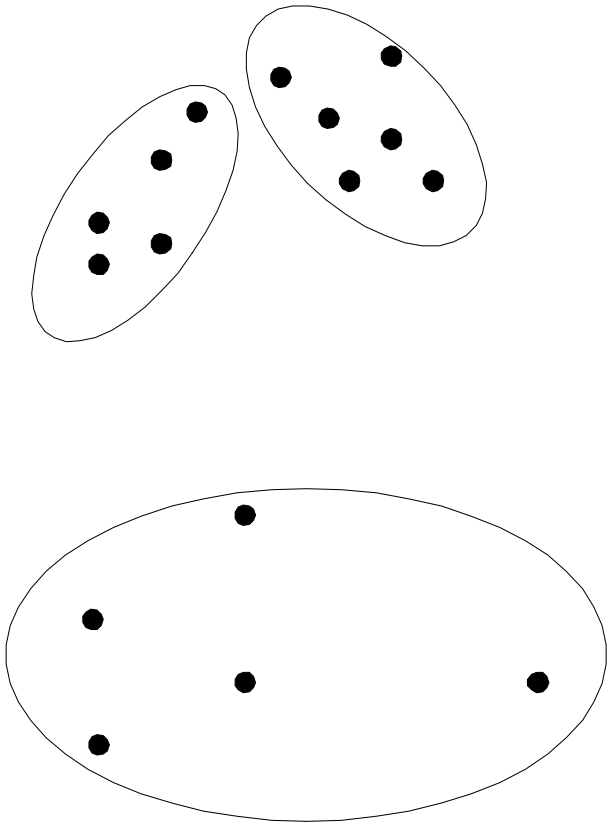
Tipovi klasterovanja

- Klasterovanje nije skup klastera već postupak
- Postoji značajna razlika između hijerarhijskog i particionog skupa klastera
- Particiono klasterovanje
  - Podela skupa podataka u nepreklapajuće podskupove (klastere) takve da je svaki podatak tačno u jednom podskupu
- Hijerarhijsko klasterovanje
  - Skup ugnježdenih klastera organizovan u obliku hijerarhijskog drveta

Particiono klasterovanje

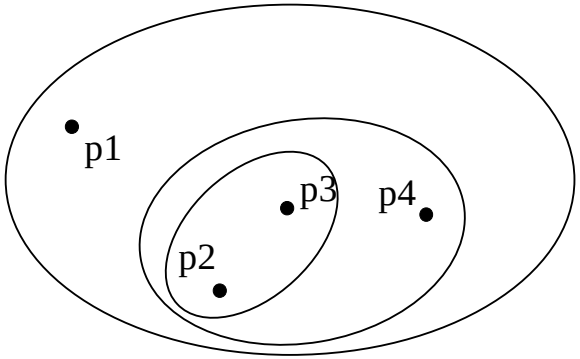


Početni podaci

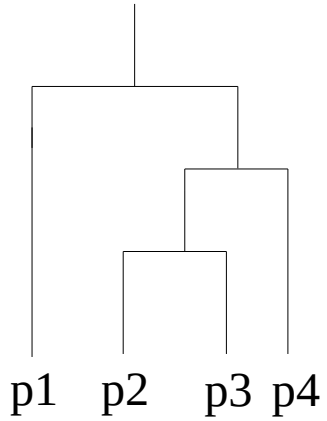


Particiono klasterovanje

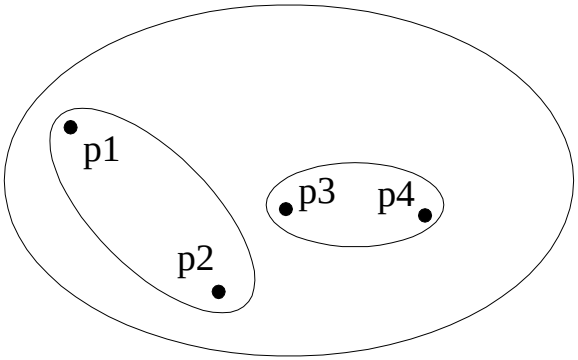
Hijerarhijsko klasterovanje



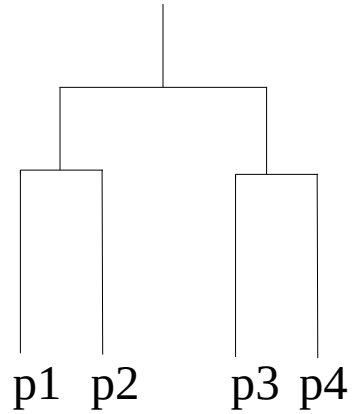
Tradicionalno hijerarhijsko klasterovanje



Tradicionalni dendogram



Netradicionalno hijerarhijsko klasterovanje



Netradicionalni dendogram

Različiti tipovi klasterovanja

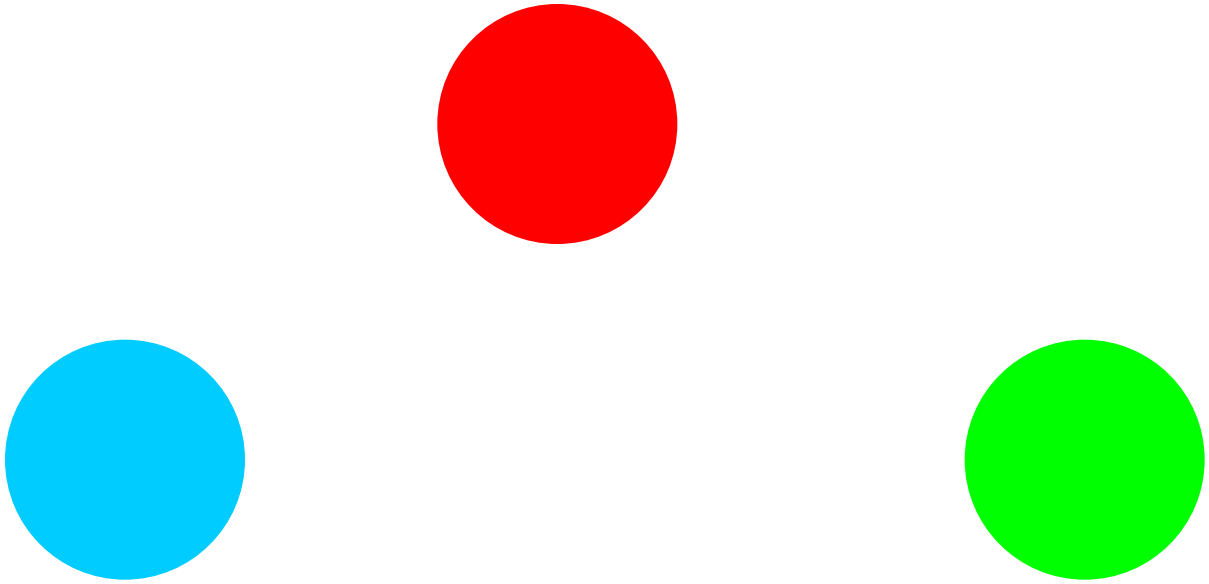
- Ekskluzivno/neekskluzivno klasterovanje
  - U neekskluzivnom klasterovanju taćke mogu da se nalaze u više klasera.
  - Can represent multiple classes or ‘border’ points
- Rasplinuto/nerasplinuto klasterovanje
  - U rasplinutom klasterovanju tačka pripada svakom klasteru sa nekom težinom između 0 i 1
  - Zbor svih težina je jednak 1
  - Slične karakteristike ima verovatnosno klasterovanje
- Delimično/kompletno klasterovanje
  - U nekim slučajevima može se klastervati samo deo podataka
- Heterogeno/homogeno klasterovanje
  - Klasteri različite veličine, oblika i gustine

Tipovi klastera

- Dobro razdvojeni klasteri (eng. *well-separated*)
- Klasteri zasnovani na centru (eng. *center-based*)
- Klasteri zasnovani na grafovima
  - klasteri zasnovani na susedstvu (eng. *contiguous*)
- Klasteri zasnovani na gustini (eng. *density-based*)
- Konceptualni klasteri/klasterovanje na osnovu zajedničkih osobina (eng. *conceptual*)
- Opisani ciljnom funkcijom (eng. *described by an objective function*)

Dobro razdvojeni klasteri

- Dobro razdvojeni klasteri:
  - Klaster je skup tačaka takvih da je bilo koja tačka u klasteru bliže (ili više slična) ostalim tačkama u klaster nego tačkama koje nisu u klasteru



3 dobro razdvojena klastera

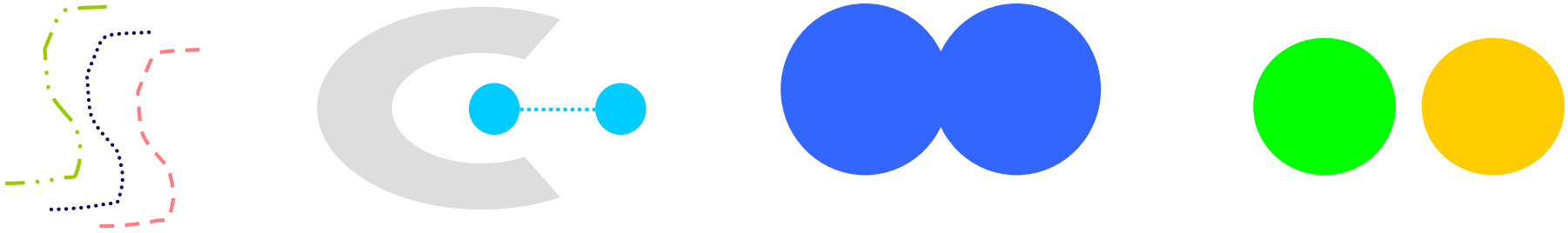
Klasteri zasnovani na centru

- Klasteri zasnovani na centru
  - Klaster je skup objekata takvih da je bilo koji objekat u klasteru bliže (ili više sličan) “centru” klastera u odnosu na centre ostalih klastera
  - Centar klastera je često centroid (prosek svuh tačaka u klasteru) ili medoid (najreprezentativnija tačka u klasteru)

4 klastera zasnovana na centru

Klasteri zasnovani na grafovima

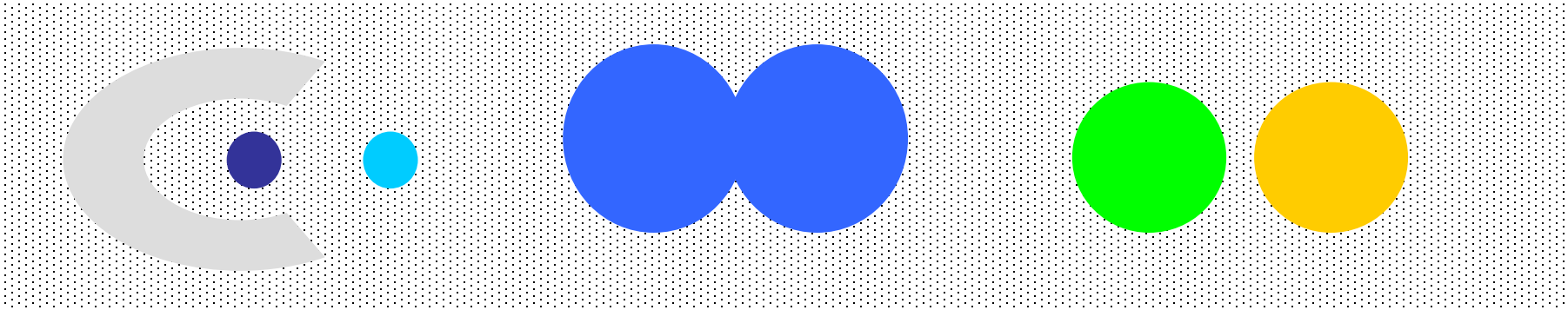
- Klasteri zasnovani na susedstvu(najbliži sused ili tranzitivnost)
  - Klaster je skup tačaka takvih da je tačka u klasteru bliža (ili više slična) jednoj ili više tačaka u klasteru nego bilo kojoj tački koja nije u klasteru



8 susednih klastera

Klasteri zasnovani na gustini

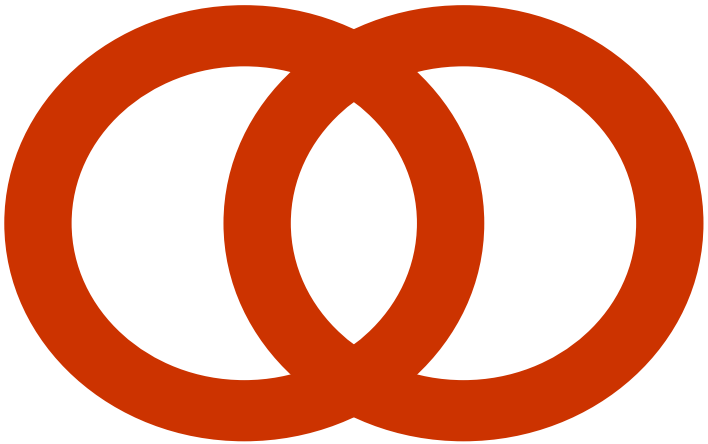
- Klasteri zasnovani na gustini
  - Klasteri su oblasti sa velikom gustinom tačaka koje su razdvojene oblastima sa malom gustinom tačaka
  - Koriste se kada su klasteri nepravilni ili isprepletani, i kada je prisutan šum ili elementi van granica



6 klastera zasnovanih na gustin

Konceptualni klasteri

- Konceptualni klasteri/klasterovanje na osnovu zajedničkih osobina
  - Naći klastere koji dele neku zajedničku osobinu ili predstavljaju pojedinačni koncept



2 konceptualna klastera

Klasteri opisani ciljnom funkcijom

- Klasteri opisani ciljnom funkcijom
  - Naći klastere koji minimizuju/maksimizuju ciljnu funkciju
  - Nabrojati sve moguće načine za podelu tačaka u klastere i izračunati valjanost svakog mogućeg skupa klastera upotrebom date ciljne funkcije (NP kompleksan)
  - Mogu da postoje loklani ili globalni ciljevi
    - Algoritmi hijerarhijskog klasterovanja obično imaju lokalne ciljeve
    - Algoritmi particionog klasterovanja obično imaju globalne ciljeve
  - Varijanta pristupa sa globalnom ciljnom funkcijom je upasovati podatke u parametrizovani model
    - Parametri modela su izvedeni iz podataka
    - Mešani modeli pretpostavljaju da su podaci mešavina nekoliko statističkih raspodela

Klasteri opisani ciljnom funkcijom

- Preslikati problem klasterovanja u različit domen i rešiti ga u tom domenu
  - Matrica sličnosti definiše graf sa težinama u kome su čvorovi tačke koje se klasteruju, dok grane sa težinama predstavljaju sličnosti između tačaka
  - Klasterovanje je ekvivalentno podeli grafa na međusobno povezane komponente koje predstavljaju klastere
  - Cilj je minimizovati težine na granama klastera i maksimizovati težine na granama unutar klastera

Važnost karakterstika ulaznih podataka

- Tip sličnosti ili mera gustine
  - Izvedena mera koja je centralna za klasterovanje
- Raštrkanost
  - Diktira tip sličnosti
  - Utiče na efikasnost
- Tip atributa
  - Diktira tip sličnosti
- Tip podatka
  - Diktira tip sličnosti
  - Ostale karakteristike, npr. autokorelacija
- Dimenzionalnost
- Šum i elementi van granica
- Tip raspodele

Algoritmi klasterovanja

- K-sredine i varijante
- Hijerarhijsko klasterovanje
- Klasterovanje zasnovano na gustinama

Klasterovanje pomoću K-sredina

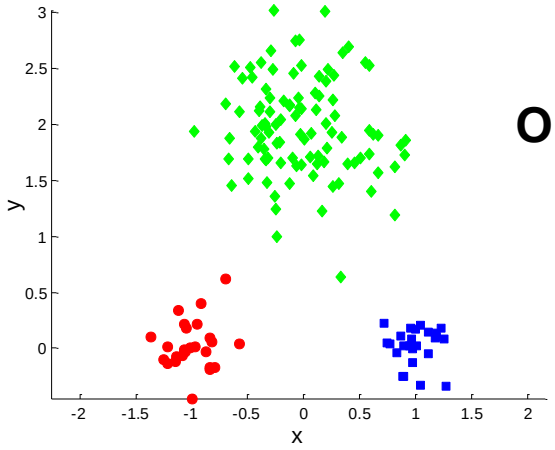
- Pristup particionim klasterovanjem, model sa prototipom
- Svakom klasteru je pridružen centroid (centralna tačka)
- Svaka tačka je dodeljena klasteru sa najbližim centroidom
- K – broj klastera koji mora da se navede
- Osnovni algoritam je vrlo jednostavan

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

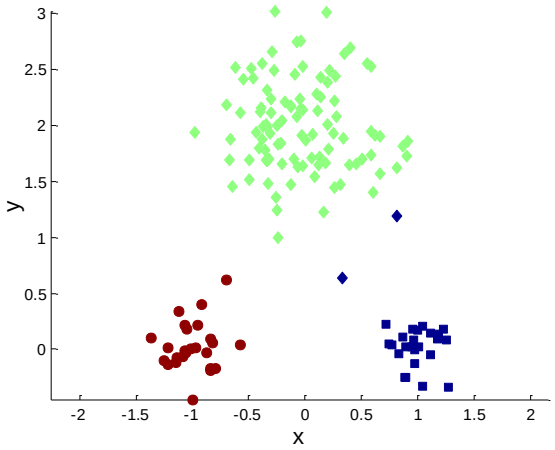
Detalji klasterovanja K-sredinama

- Početni centroid se često bira na slučajan način
  - Dobijeni klasteri mogu da se razlikuju u uzastopnim izvršenjima programa. Rezultati mogu da budu relativno loši
- Uobičajeno je da je centroid srednja vrednost tačaka u klasteru
- 'Najbliže' se meri kao Euklidsko rastojanje, kosinusno rastojanje, korelacija, itd.
- K-sredine konvergiraju ka prethodno pomenutim merama sličnosti
- Najveći deo konvergencije se dešava u prvih nekoliko iteracija
  - Često se uslov zaustavljanja menja na 'sve dok relativno malo tačaka ne promeni klaster'
- Kompleksnost (n = broj tačaka, K = broj klastera, I = broj iteracija, d = broj atributa)
  - Vremenska je reda O(n * K * I * d)
  - Prostorna je reda O((n+K)*d)

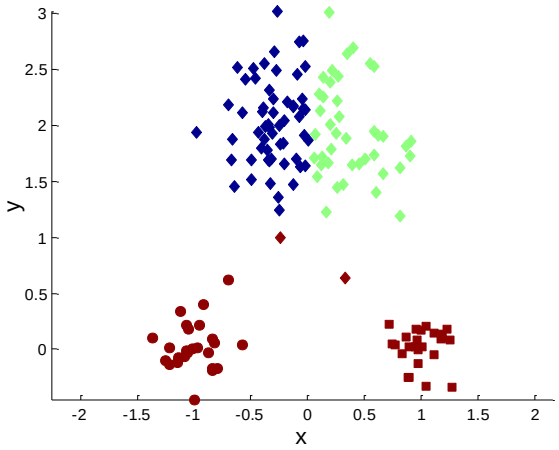
Dva različita klasterovanja K-sredinama



Originalne tačke

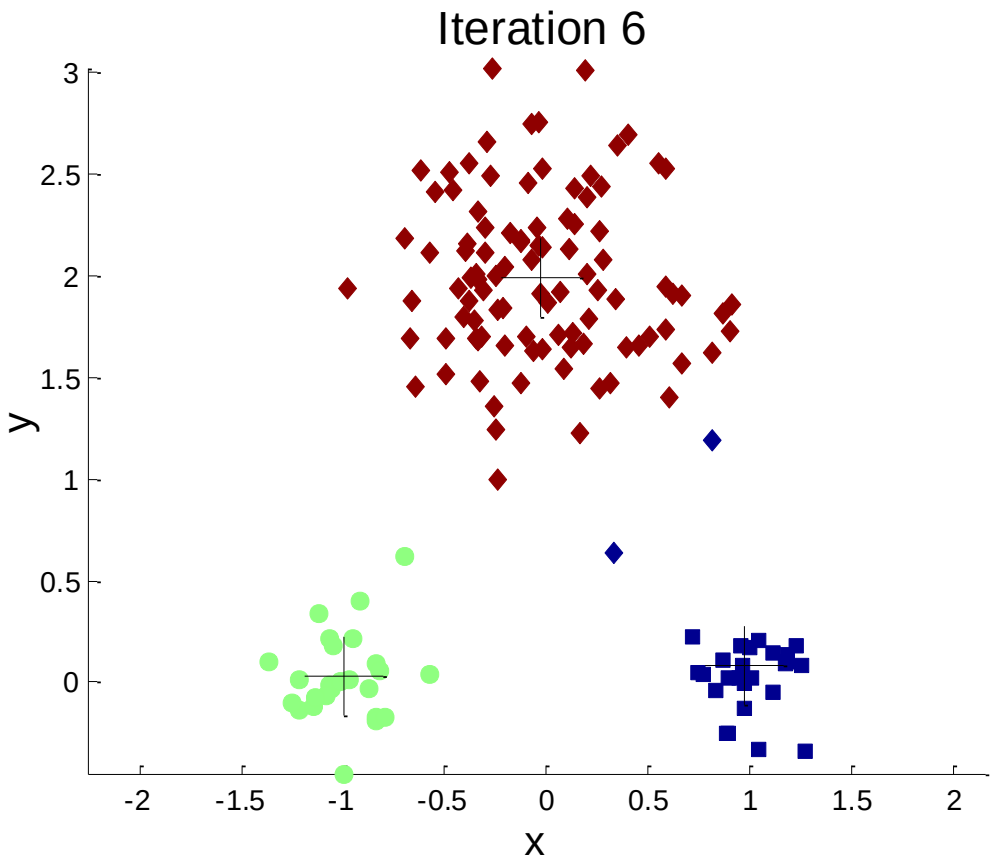


Optimalno klasterovanje

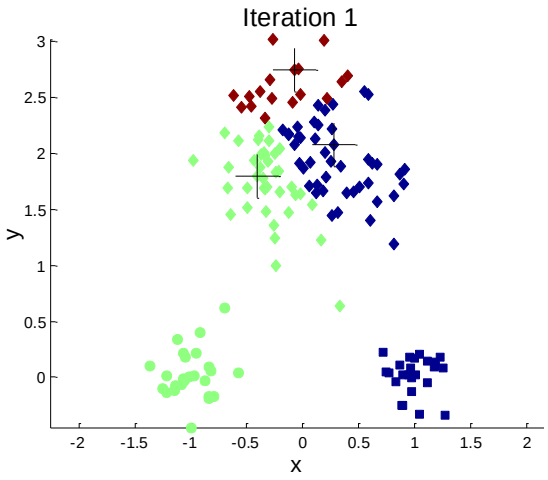


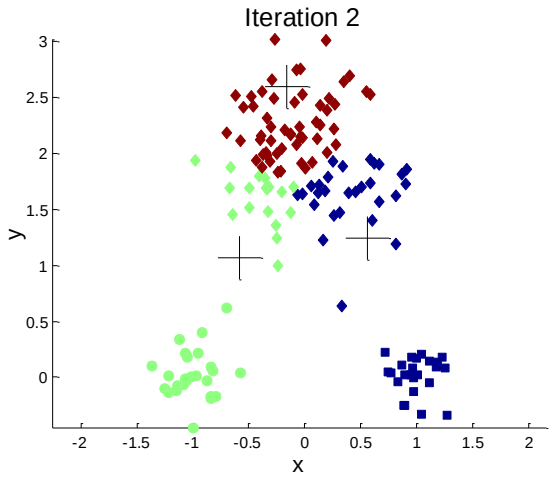
Sub-optimalno klasterovanje

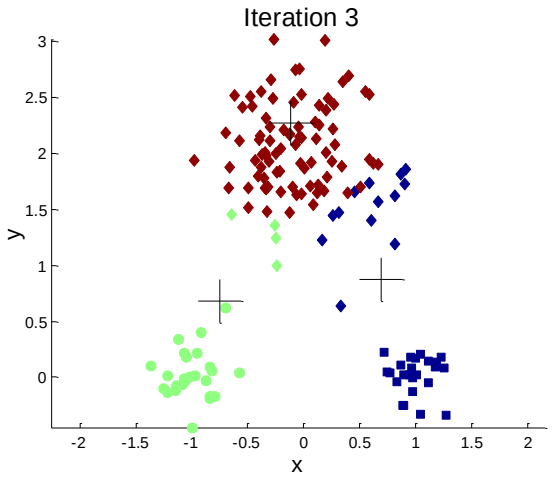
Važnost izbora početnog centroida

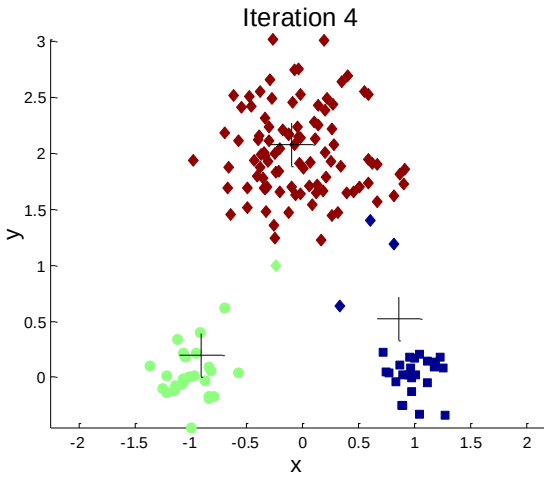


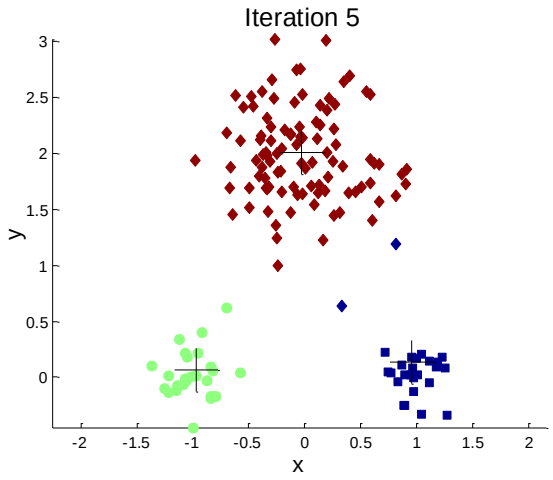
Važnost izbora početnog centroida

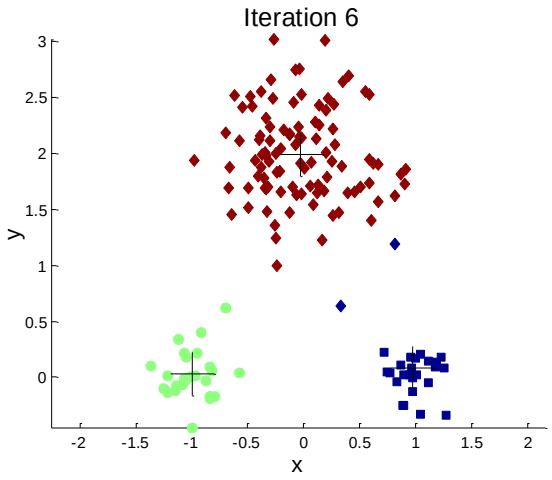












Evaluacija K-sredine klastera

- Za podatke u Euklidskom prostoru se najčešće se kao mera korsti zbir kvadrata grešaka (eng. *sum of squared errors*)
  - Za svaku tačku, greška je rastojanje do najbližeg klastera
  - Dobijene greške se kvadriraju i sabiraju

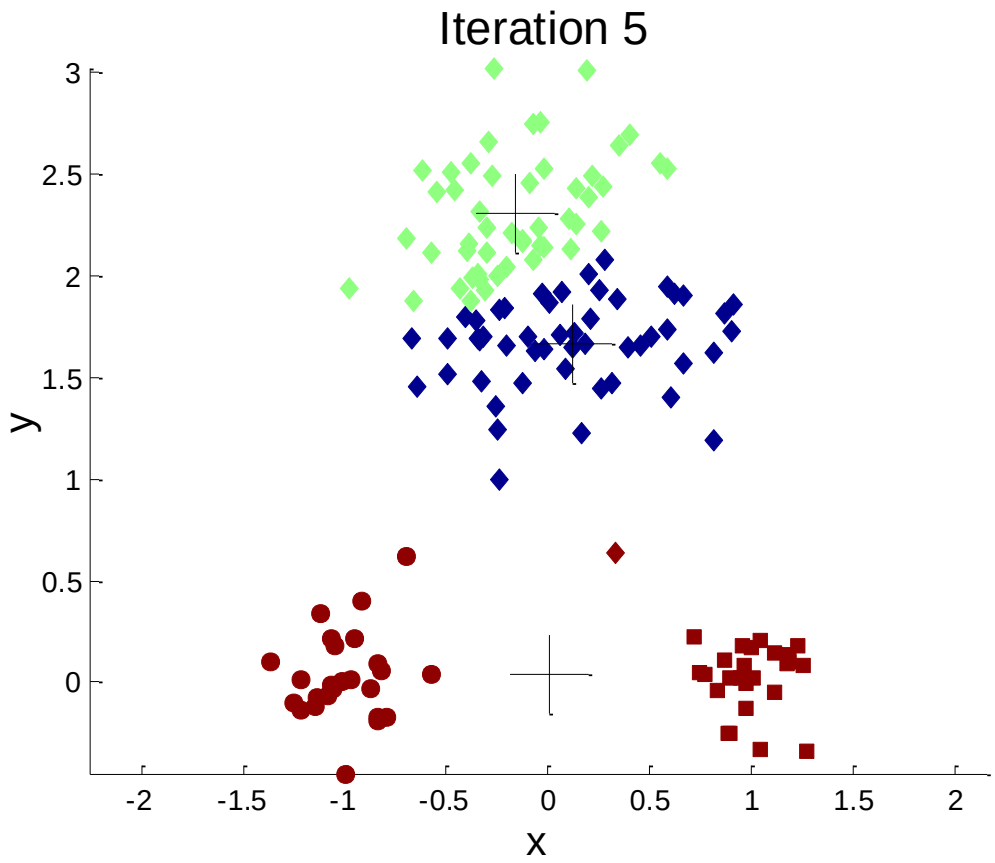
$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

  - x je tačka u klasteru C_i a m_i je reprezentativna tačka u klasteru C_i
    - može se pokazati da m_i odgovara centru (srednjoj vrednosti) klastera
  - Za dva data klastera bira se onaj sa manjom greškom
  - Jedan od načina za smanjenje SSE je povećanje broja klastera K
    - Dobra klasterizacija sa malim K može da ima manju SSE grešku od loše klasterizacije sa velikim K

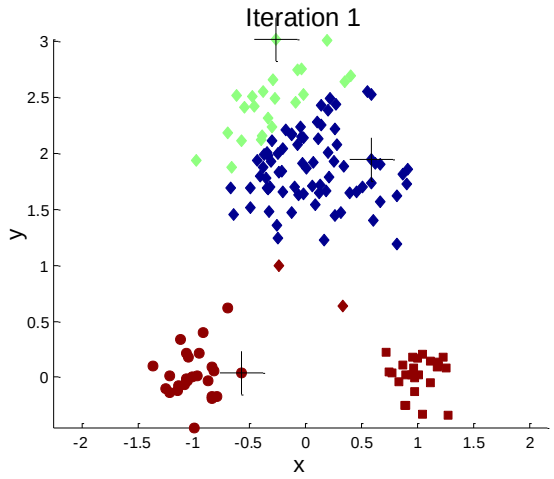
Evaluacija K-sredine klastera

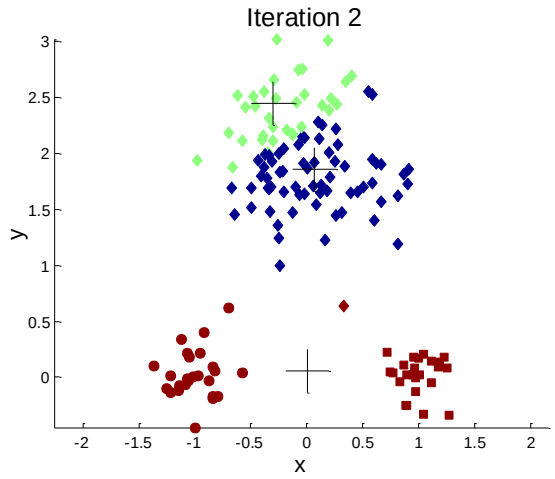
- Za dokumente se kao mera koristi kosinusno rastojanje
  - Podaci se predstavljaju preko matrice termova
  - Kohezija klastera - stepen sličnosti dokumentata u klasteru sa centroidom

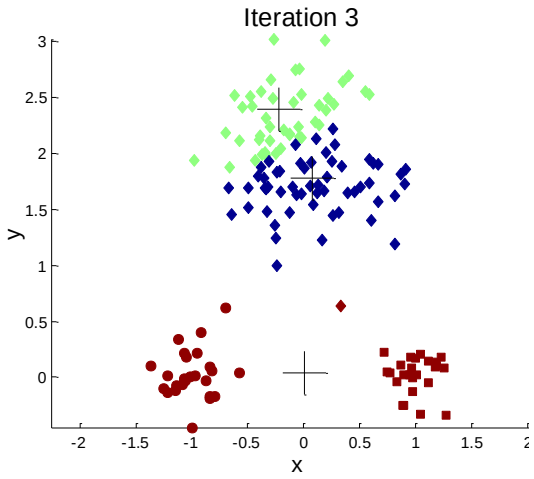
Važnost izbora početnog centroida

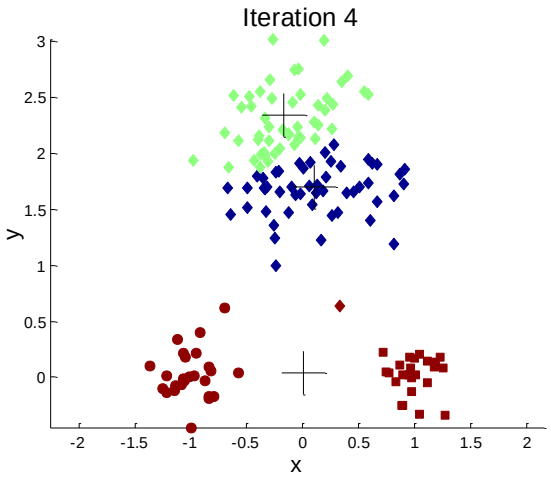


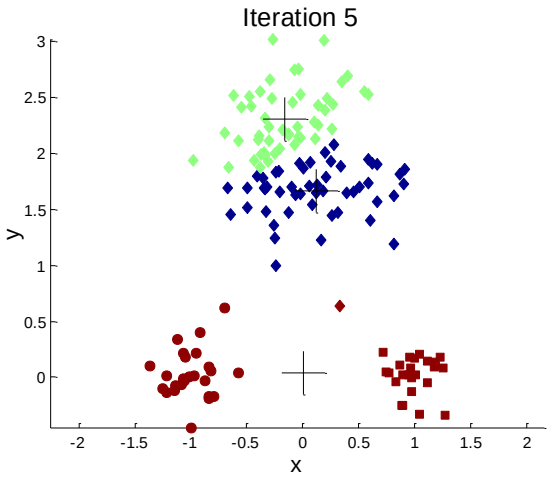
Važnost izbora početnog centroida











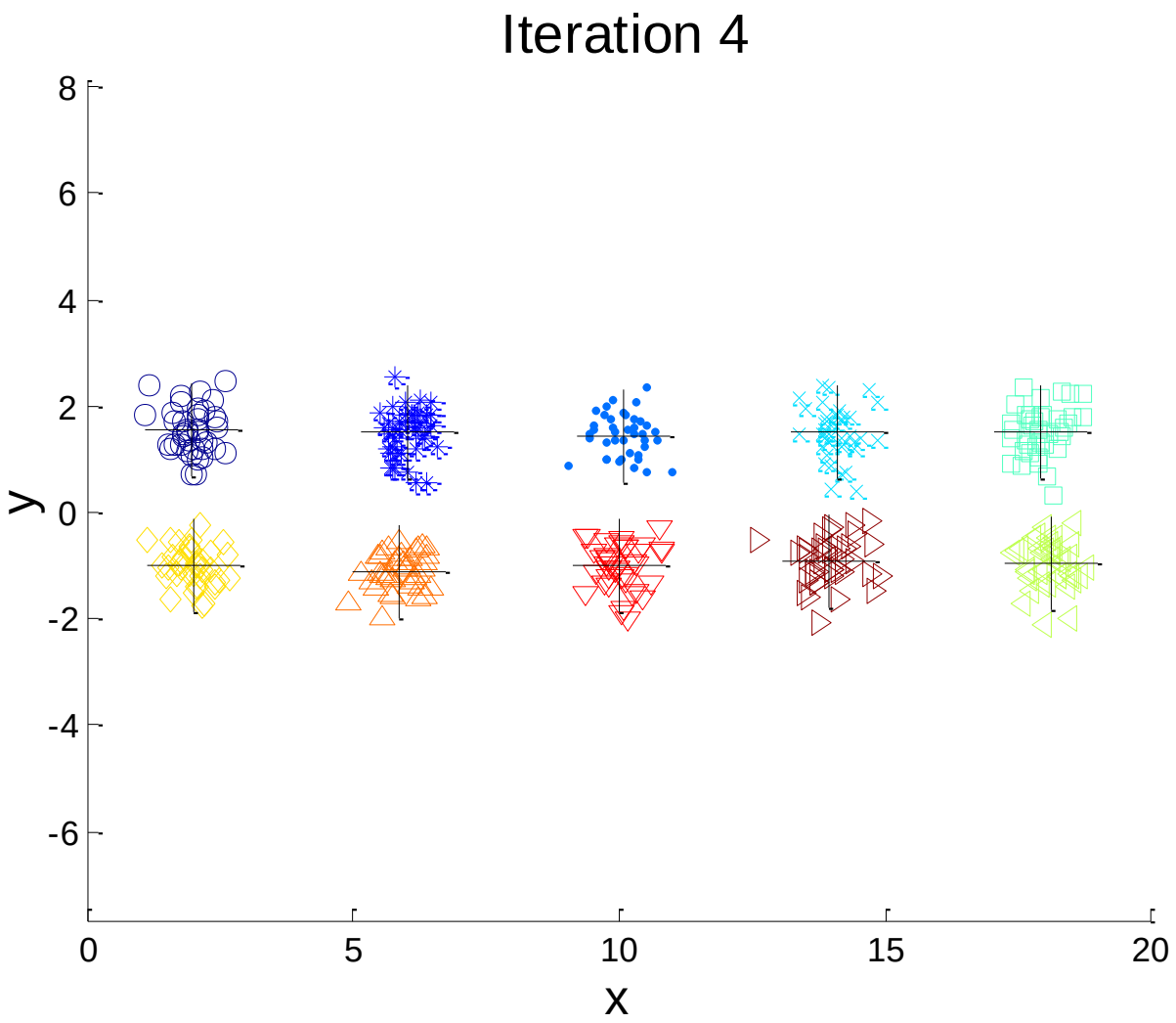
Problem u izboru početnih tačkaka

- Ako postoji K ‘realnih’ klastera tada je šansa da se izabere po jedan centroid u svakom od njih relativno mala
  - Ako je K veliko šansa za dobar izbor je mala
  - Ako klasteri imaju istu veličinu n, tada važi

$$P = \frac{\text{number of ways to select one centroid from each cluster}}{\text{number of ways to select } K \text{ centroids}} = \frac{K!n^K}{(Kn)^K} = \frac{K!}{K^K}$$

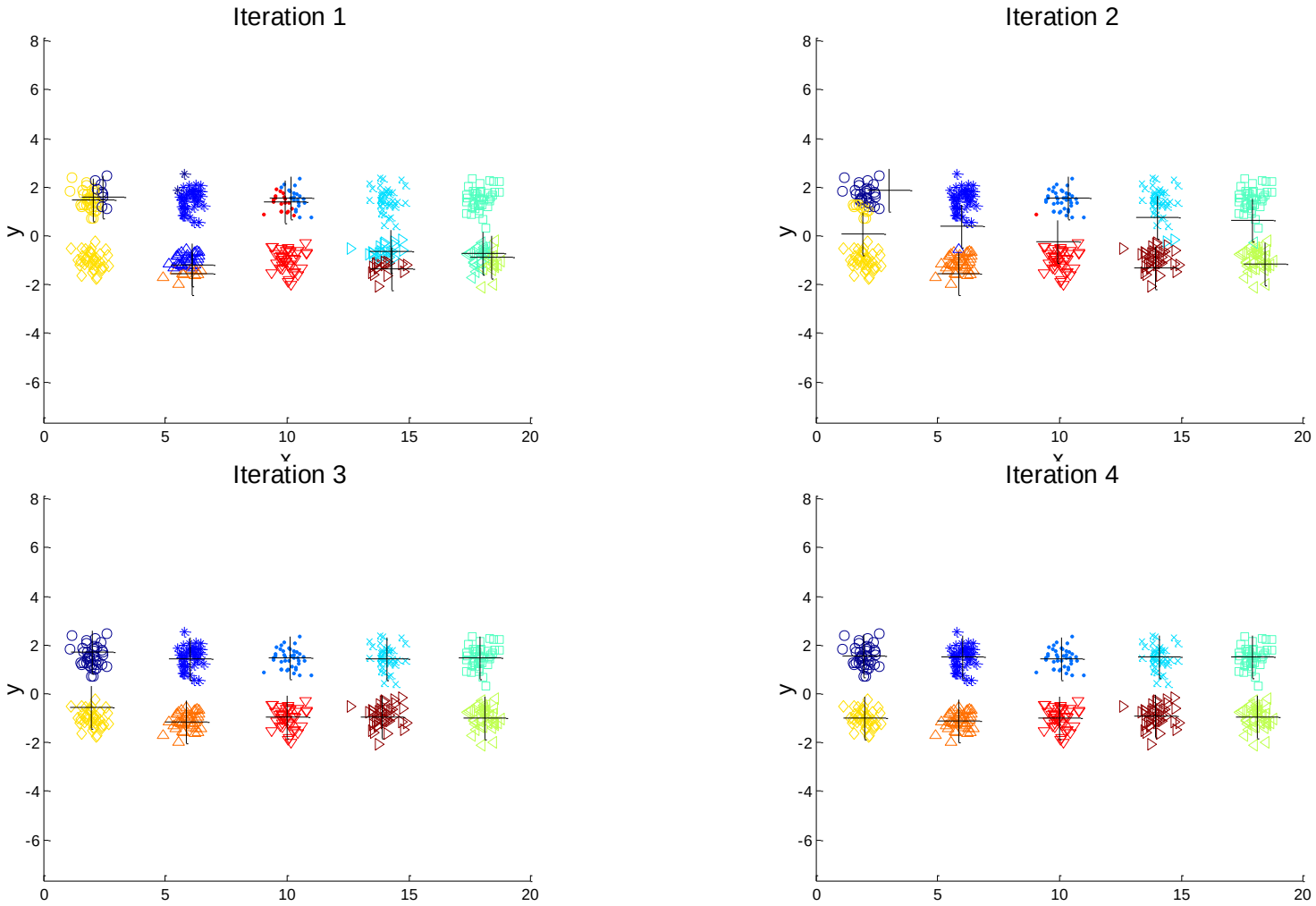
  - Na primer, ako je K = 10, tada je verovatnoća = $10!/10^{10}$ = 0.00036
  - Ponekad se inicijalni centroidi sami poravnaju na ‘pravi’ redosled, a ponekad ne
  - Posmatrajmo primer pet parova klastera

Primer: pet parova klastera



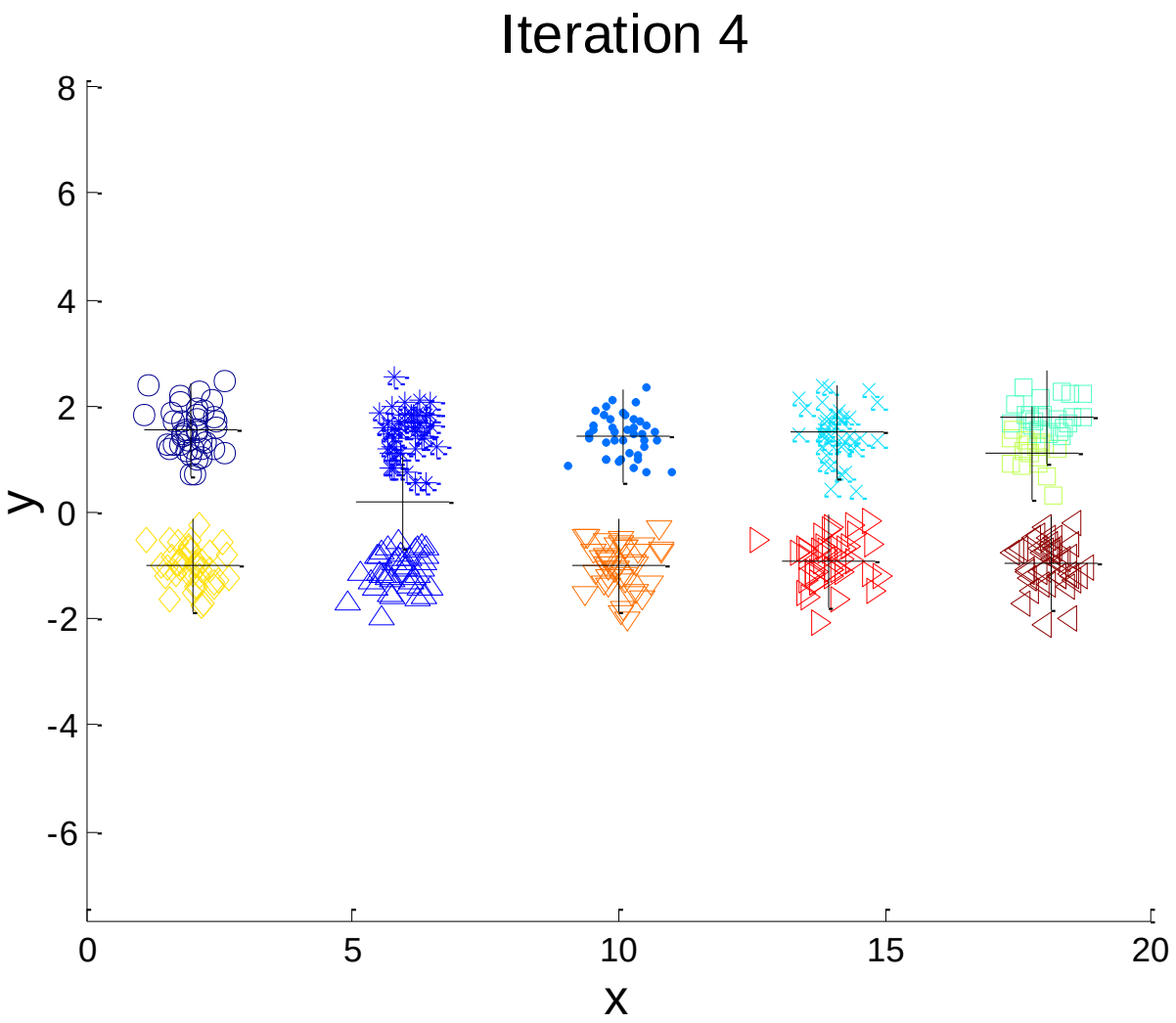
Za svaki par klastera počinje se sa dva centroida u jednom od klastera tog para

Primer: pet parova klastera



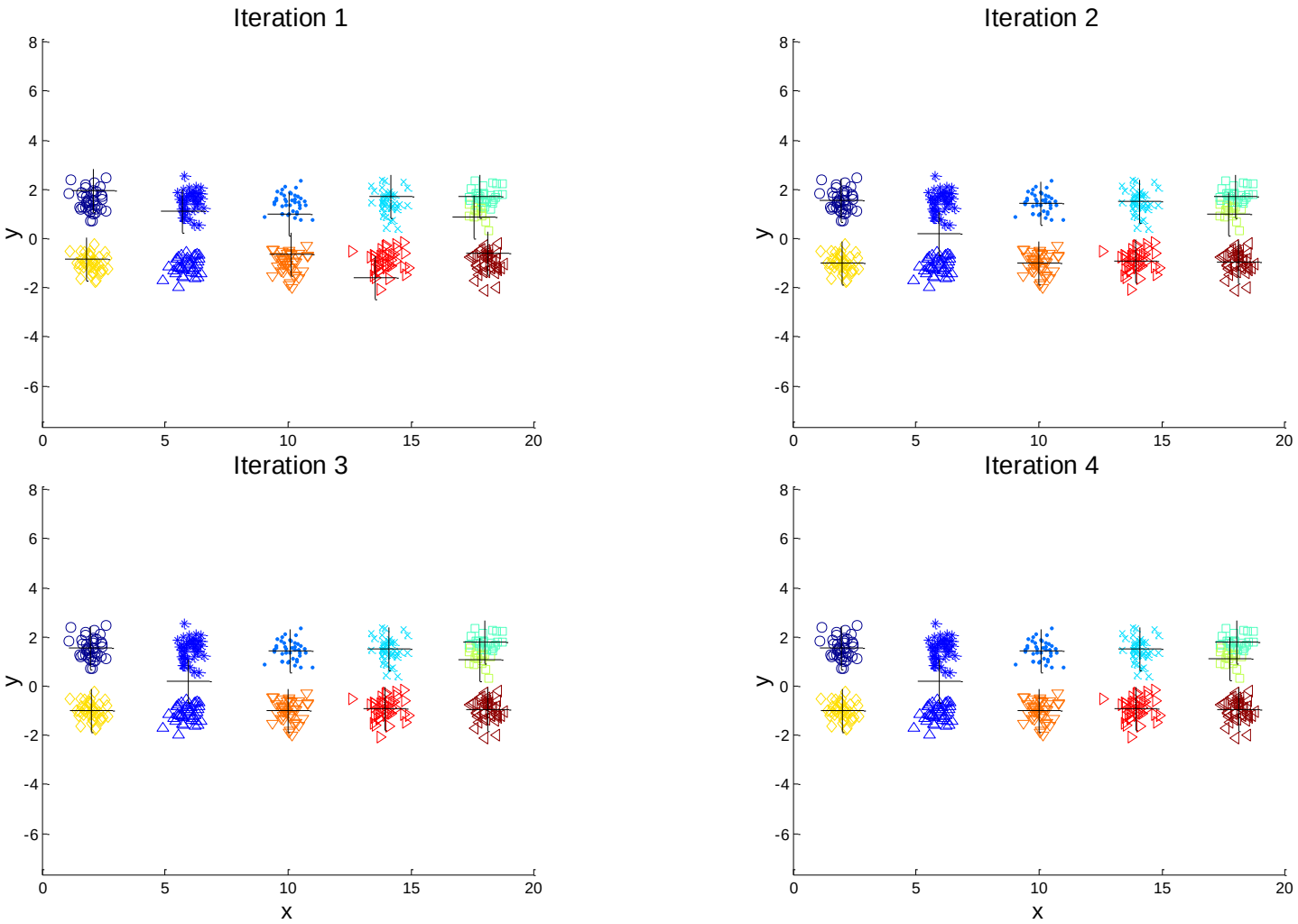
Za svaki par klastera počinje se sa dva centroida u jednom od klastera tog para

Primer: pet parova klastera



U nekim parovima klastera počinje se sa tri početna centroida a u nekima sa samo jednim

Primer: pet parova klastera



U nekim parovima klastera počinje se sa tri početna centroida a u nekima sa samo jednim

Rešenje problema izbora početnih centroida

- Uzastopna izvršavanja
  - Svaki sa npr. slučajno izabranim centroidima
  - između njih se izabere klaster sa najmanjim SSE
- Nad uzorcima se primeni hijerarhijsko klasterovanje i izaberu početni centroidi
- Izabere se više od K početnih centroida i bira se između njih
  - Treba da obuhvate što je moguće širi prostor
- Postprocesiranje
- Bisekcija K-sredina
  - Nije tako osetljiva na incijalne vrednosti

Rad sa praznim klasterima

- Osnovni K-sredina algoritam može da proizvede prazne klastere
- Strategije za eliminaciju
  - Zamenjuje se centroid
    - Izabrati tačku koja najviše učestvuje u SSE
    - Izabrati tačku koja je najdalje od tekućih centroida
    - Izabrati tačku iz klastera sa najvećim SSE. Obično dovodi do deobe klastera
    - Ako ima više praznih klastera ponoviti postupak

Preprocesiranje i postproceisranje

- Preprocesiranje
  - Noramlizacija podataka
  - Eliminacija elemenata van granica (ne važi za svaku aplikaciju, npr. kompresija)
- Postprocesiranje
  - Eliminacija malih klastera sa elementima van granica
  - Podela ‘izgubljenih’ klastera, tj. klastera sa visokim SSE
  - Integracija klastera koji su ‘blizu’ i imaju relativno mali SSE
  - Ovi koraci se mogu koristiti u procesu klasterizacije

Bisekcija K-sredina

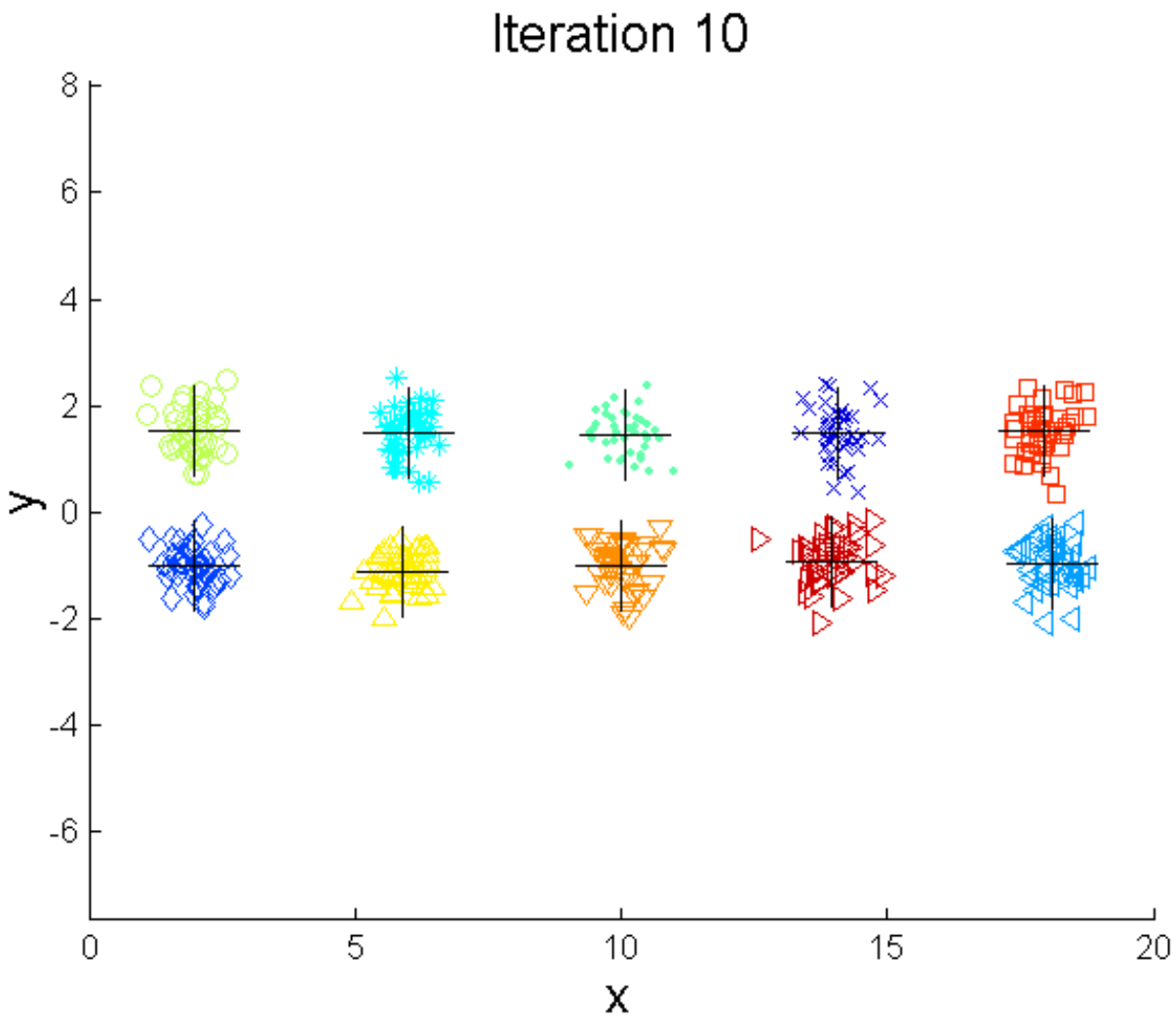
- Algoritam bisekcije K-sredina
  - Varijanta K-sredine koja može da proizvede particiono ili hijerarhijsko klasterovanje
- Osnovna ideja:
  - Za dobijanje K klastera podeli se skup svih tačaka u dva klastera, izabere se jedan od njih za podelu, uz ponavljanje postupka sve dok se ne dobije K klastera.
- Različiti načini podele:
  - najveći klaster
  - klaster sa najvećim SSE
  - kriterijum zasnovan i na veličini klastera i na veličini SSE-a
- Često se dobijeni centroidi koriste za ulaz u osnovni K-sredina algoritam klasterovanja

Bisekcija K-sredina

- Algoritam bisekcije K-sredina

1: Initialize the list of clusters to contain the cluster containing all points.
2: **repeat**
3: Select a cluster from the list of clusters
4: **for** $i = 1$ to *number_of_iterations* **do**
5: Bisect the selected cluster using basic K-means
6: **end for**
7: Add the two clusters from the bisection with the lowest SSE to the list of clusters.
8: **until** Until the list of clusters contains K clusters

Primer bisekcije K-sredina

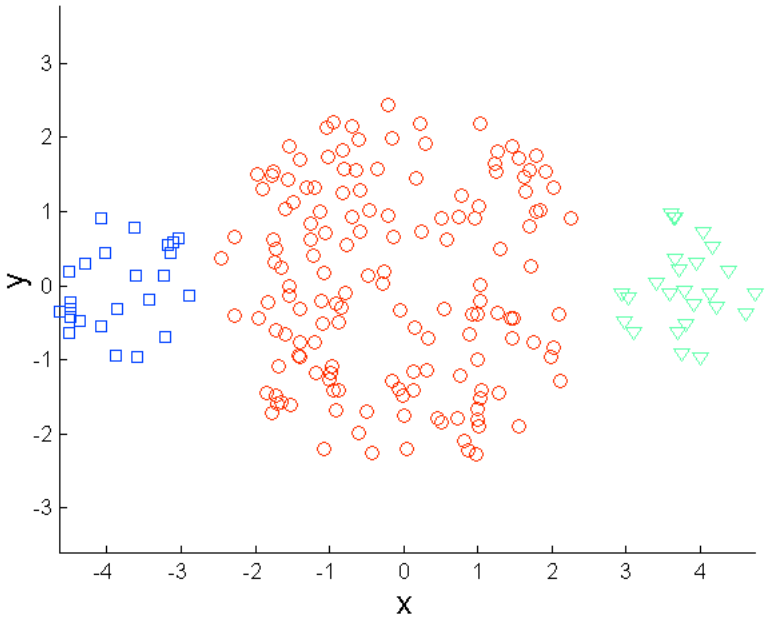


Ograničenja K-sredina

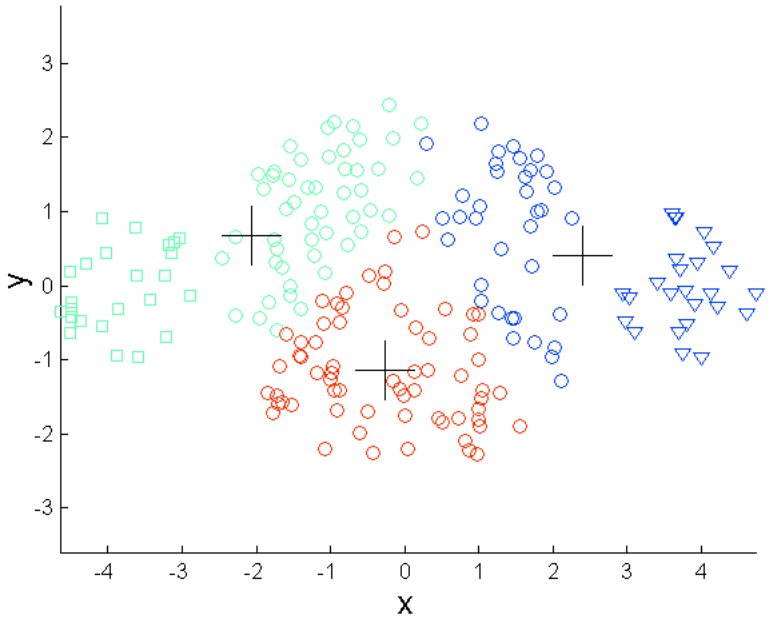
- Metoda za klasterizaciju K-sredinama ima problem u klasterovanju različitih
  - Veličina
  - Gustina
  - Neglobularnih oblika

- Metoda za klasterizaciju K-sredinama ima problem kada postoje elementi van granica

Ograničenja K-sredina : Različite veličine

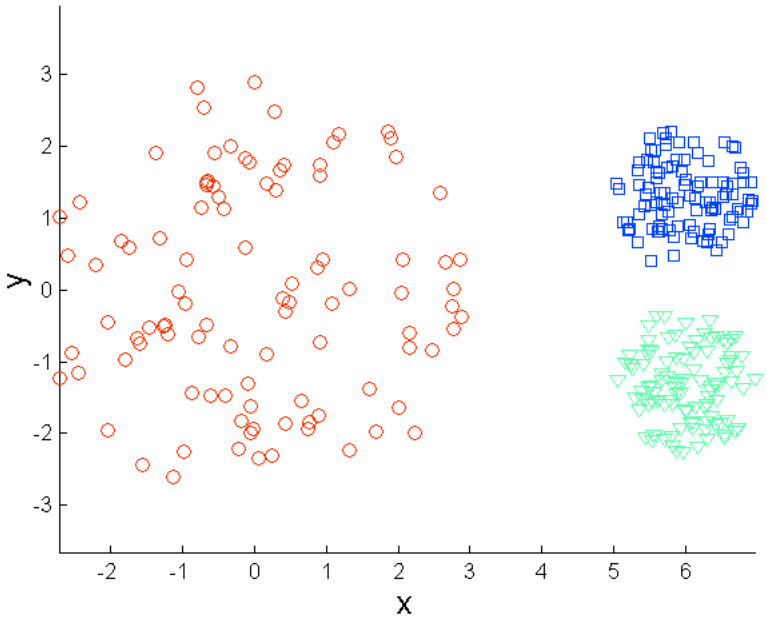


Originalne tačke

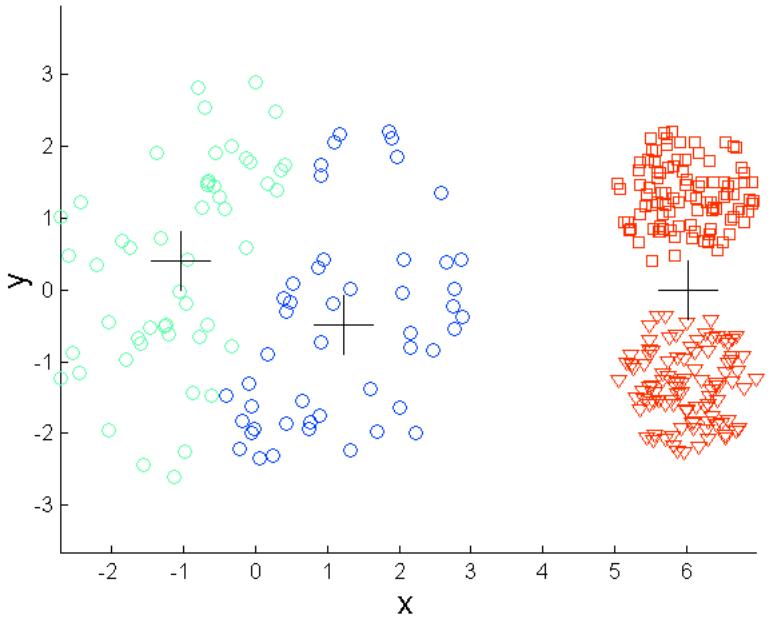


K-sredine (3 klastera)

Ograničenja K-sredina : Različite gustine

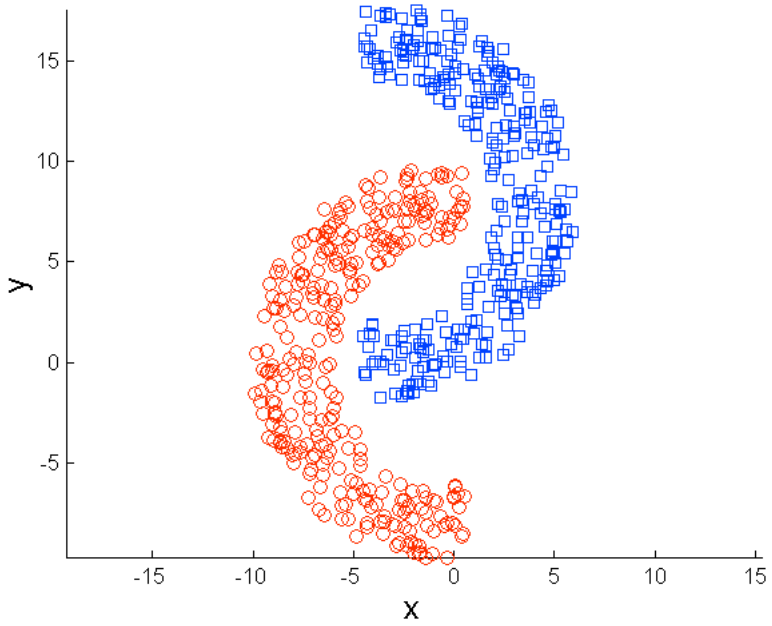


Originalne tačke

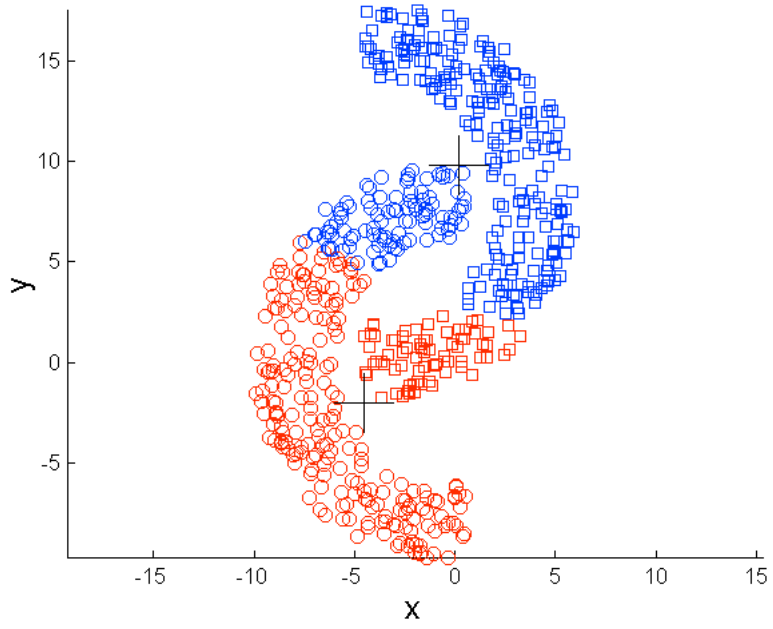


K-sredine (3 klastera)

Ograničenja K-sredina : neglobularni oblici

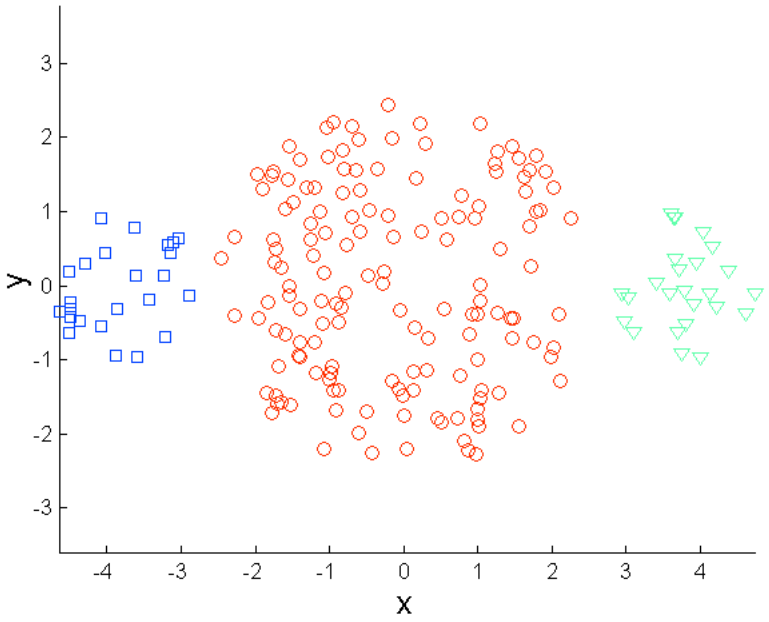


Originalne tačke

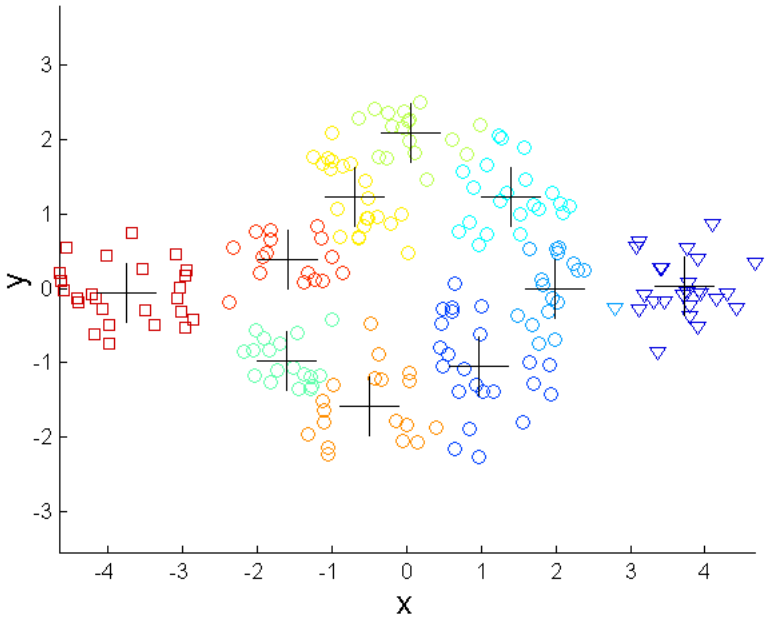


K-sredine (2 klastera)

Prevazilaženje ograničenja K-sredina



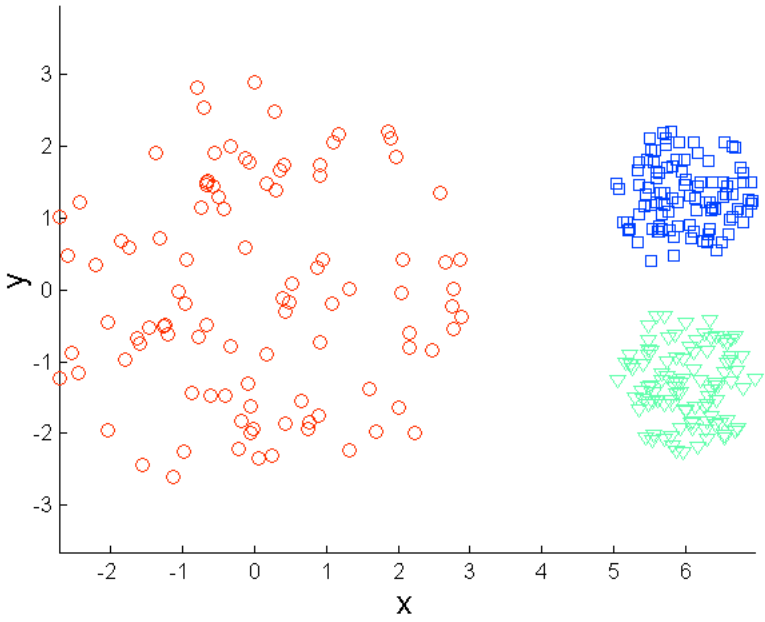
Originalne tačke



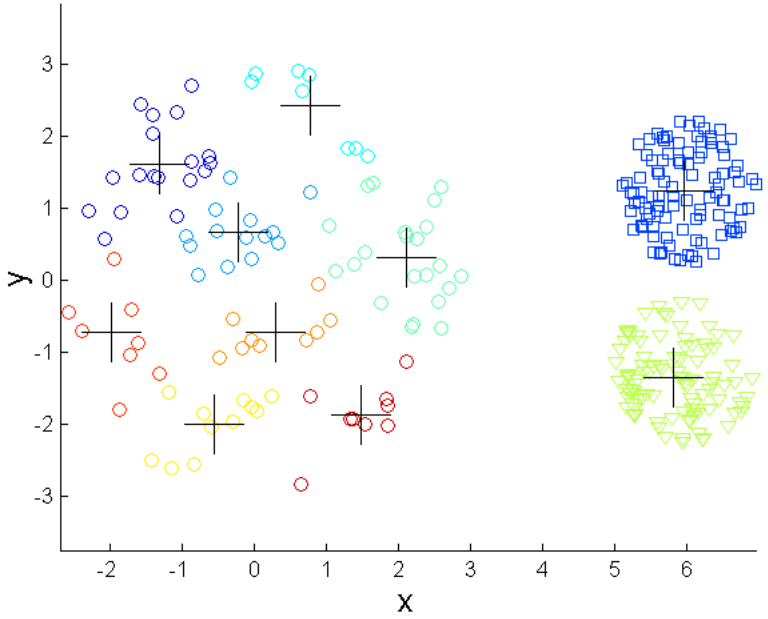
K-sredine klasteri

Jedno od rešenja je upotreba više klastera. Naći delove klastera koje zatim treba smestiti zajedno.

Prevazilaženje ograničenja K-sredina

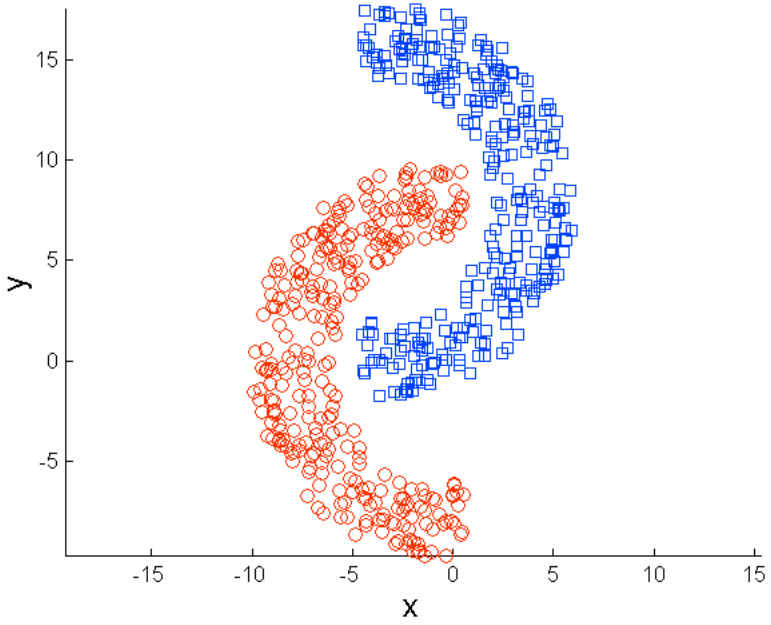


Originalne tačke

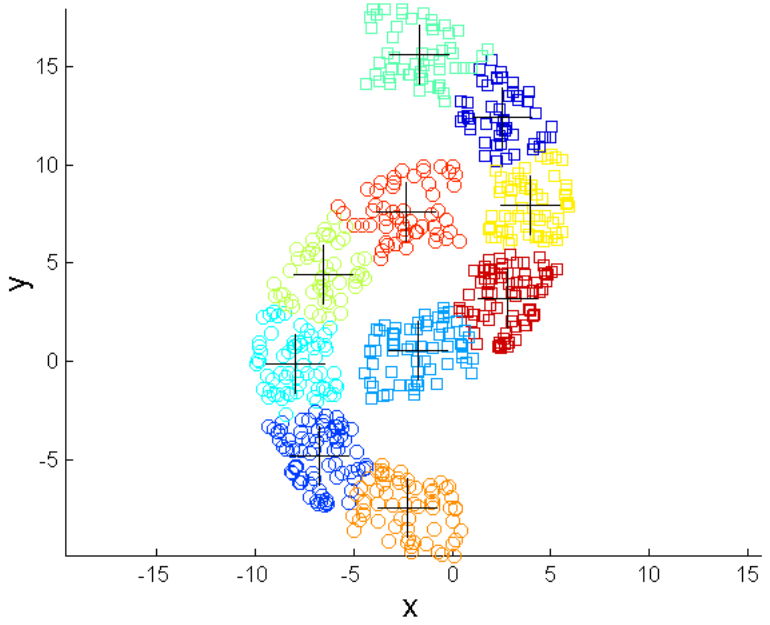


K-sredine klasteri

Prevazilaženje ograničenja K-sredina



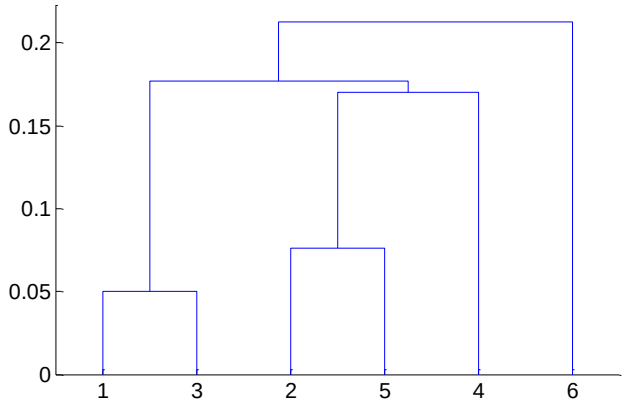
Originalne tačke

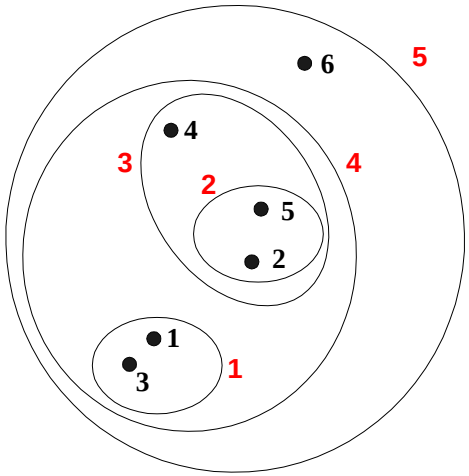


K-sredine klasteri

Hijerarhijsko klasterovanje

- Formira skup ugneždenih klastera organizovanih u obliku drveta
- Vizualizuje se u obliku dendrograma





Prednosti hijerarhijskog klasterovanja

- Nije potrebno davati pretpostavke o broju klastera
  - Željeni broj klastera se dobija skraćivanjem dendograma na odgovarajući nivo
- Mogu da imaju značenje u taksnonomijama
  - Na primer, u biologiji (e.g., filogeneza, ...)

Hijerarhijsko klasterovanje

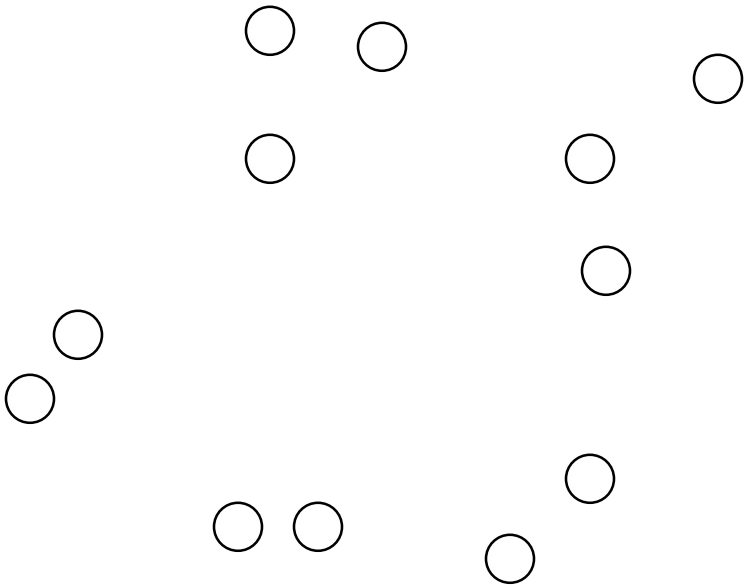
- Postoje dva glavna tipa
  - Sakupljajuće (eng. *agglomerative*):
    - U početku je svaka tačka jedan klaster
    - U svakom koraku se sakuplja najbliži par klastera u novi klaster sve dok ne ostane jedan (ili k) klastera
  - Razdvajajuće (eng. *divisive*):
    - Počinje se sa jednim klasterom koji uključuje sve tačke
    - U svakom koraku se klaster deli sve dok se ne dođe do toga da svaki klaster sadrži samo jednu tačku ili dok se ne javi k klastera
- Tradicionalni hijerarhijski algoritmi koristie matrice sličnosti ili matrice rastojanaj
  - Dele ili spajaju po jedan klaster u jednom koraku

Algoritmi sakupljajućeg klasterovanja

- Osnovni algoritam
  1. Izračunati matricu sličnosti
  2. Neka je svaka tačka klaster
  3. **Repeat**
  4. Sakupi dva najbliža klastera
  5. Ažuriraj matricu sličnosti
  6. **Until** dok ne ostane samo jedna klaster
- Ključna operacija je izračunavanje sličnosti dva klastera
  - Različiti algoritimi su posledica različitih pristupa u definisanju rastojanja između klastera

Početno stanje

- Početi sa pojedinačnim tačkama kao klasterima i sa matricom sličnosti

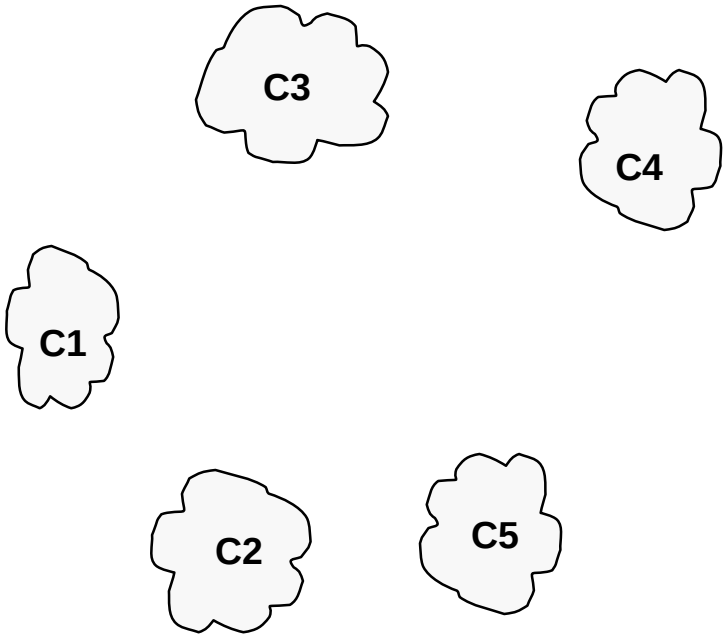


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Matrica sličnosti

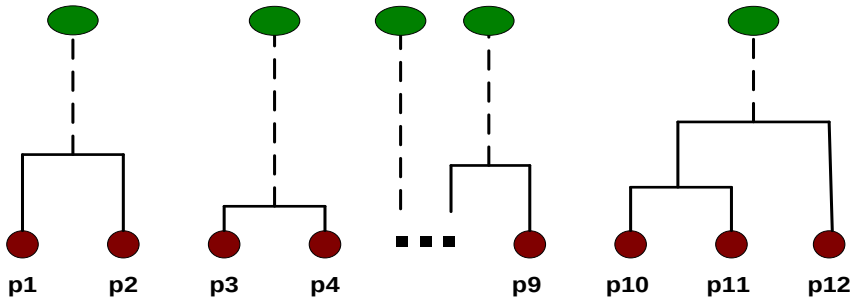
Stanje u sredini postupka

- Posle nekoliko sakupljanja javllaju se sledeći klasteri



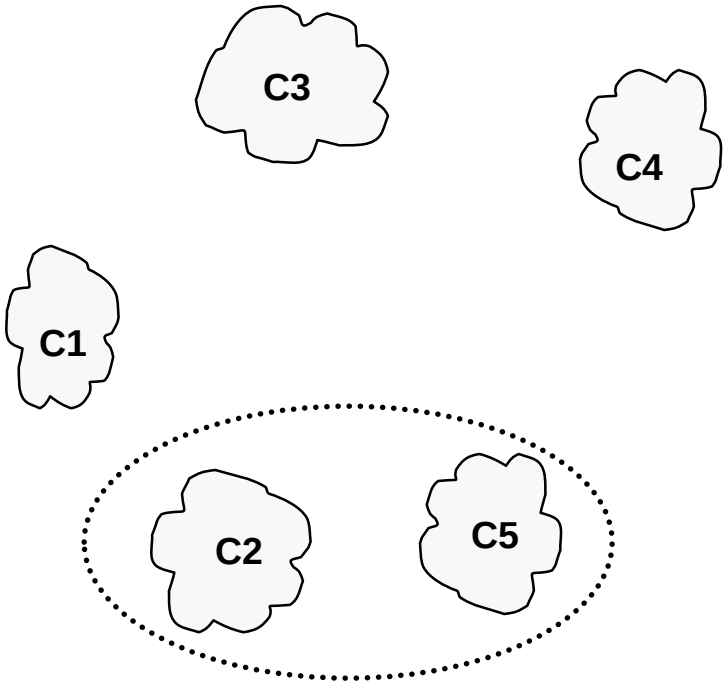
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Matrica sličnosti



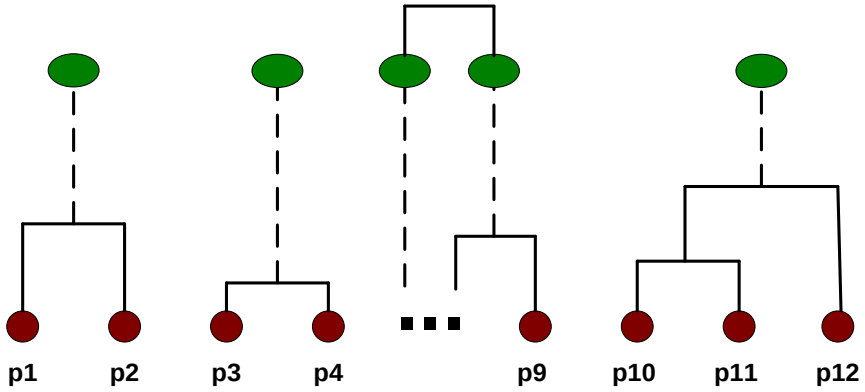
Stanje u sredini postupka

- Želimo da skupimo dva najbliža klastera (C2 i C5) i zatim da ažuriramo matricu sličnosti



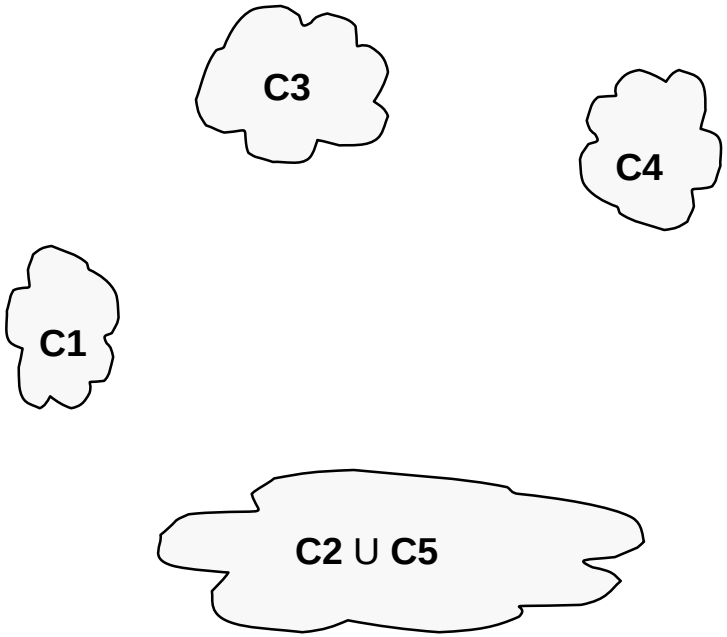
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Matrica sličnosti



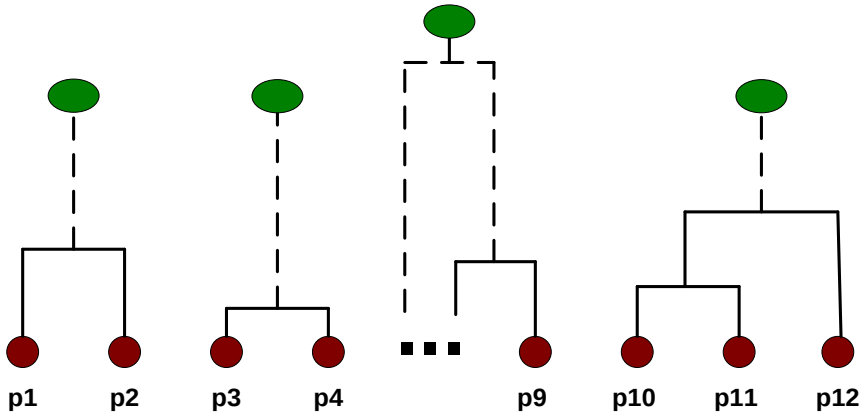
Posle sakupljanja

- Pitanje: "Kako ažurirati matricu sličnosti?"

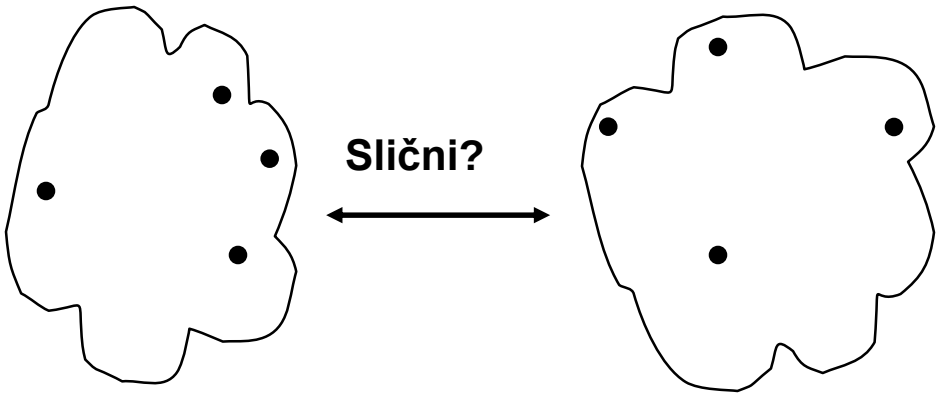


	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Matrica sličnosti



Kako definisati sličnost među klasterima

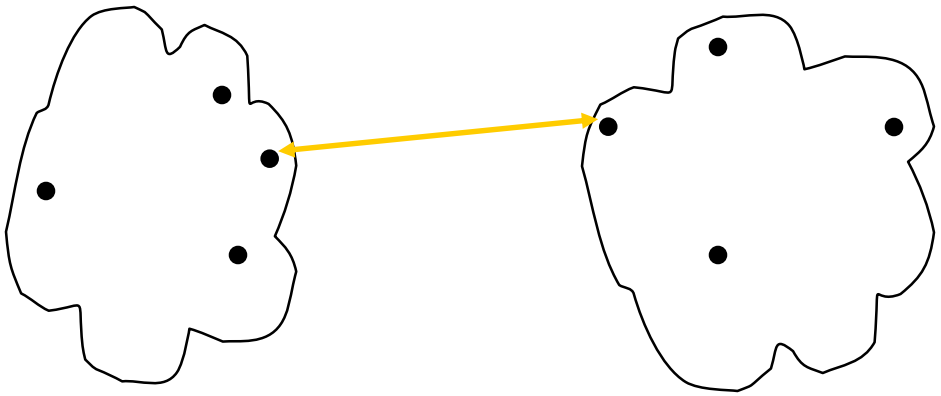


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Matrica sličnosti

- MIN
- MAX
- Prosek grupe
- Rastojanje između centroida
- Ostale metode definisane ciljnim funkcijama
  - Ward-ov metod sa kvadratom greške

Kako definisati sličnost među klasterima

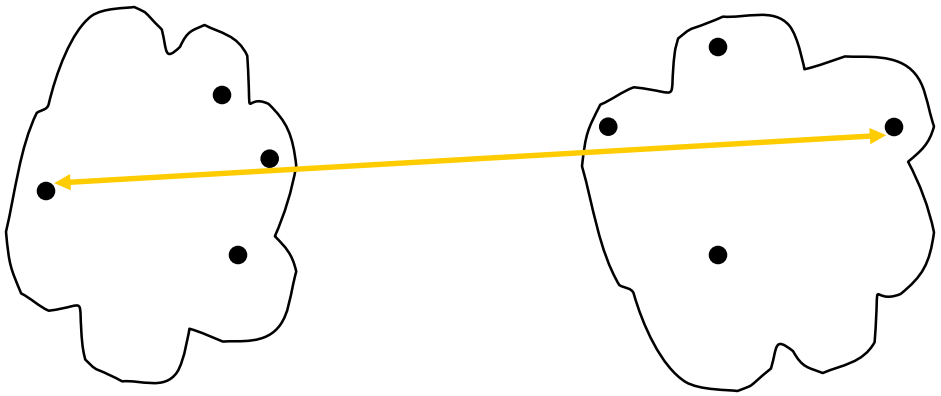


- MIN
- MAX
- Prosek grupe
- Rastojanje između centroida
- Ostale metode definisane ciljnim funkcijama
  - Ward-ov metod sa kvadratom greške

	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Matrica sličnosti

Kako definisati sličnost među klasterima



- MIN
- MAX
- Prosek grupe
- Rastojanje između centroida
- Ostale metode definisane ciljnim funkcijama
  - Ward-ov metod sa kvadratom greške

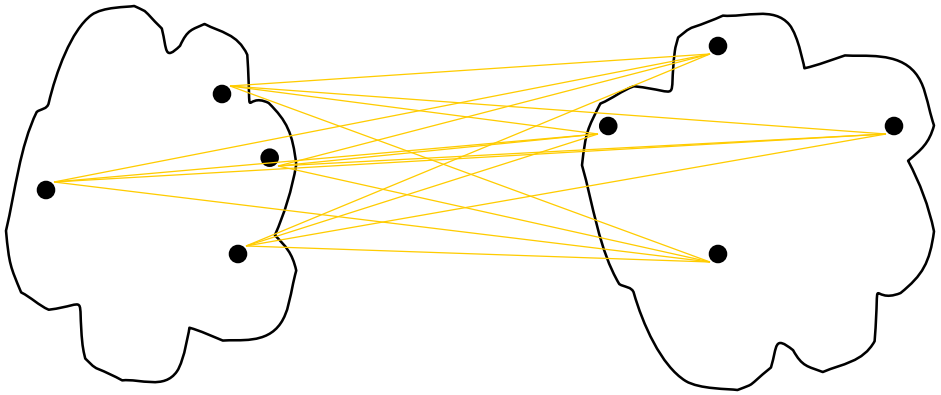
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						

.

.

Matrica sličnosti

Kako definisati sličnost među klasterima

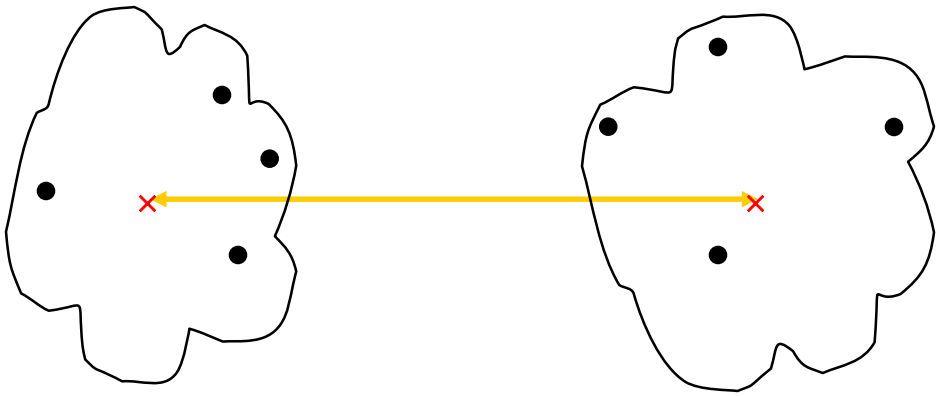


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Matrica sličnosti

- MIN
- MAX
- Prosek grupe
- Rastojanje između centroida
- Ostale metode definisane ciljnim funkcijama
  - Ward-ov metod sa kvadratom greške

Kako definisati sličnost među klasterima



- MIN
- MAX
- Prosek grupe
- Rastojanje između centroida
- Ostale metode definisane ciljnim funkcijama
  - Ward-ov metod sa kvadratom greške

	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						

.

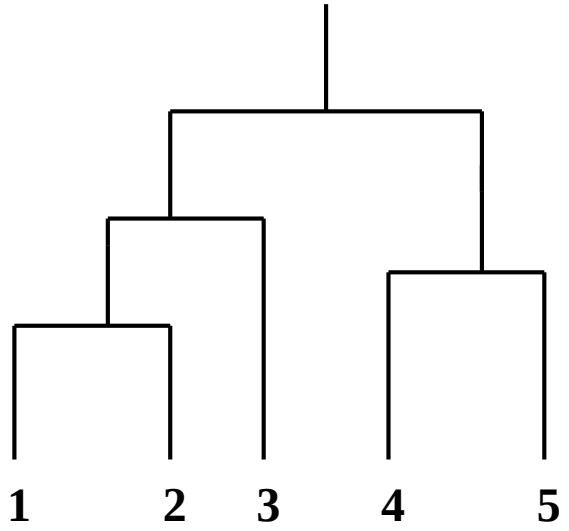
.

Matrica sličnosti

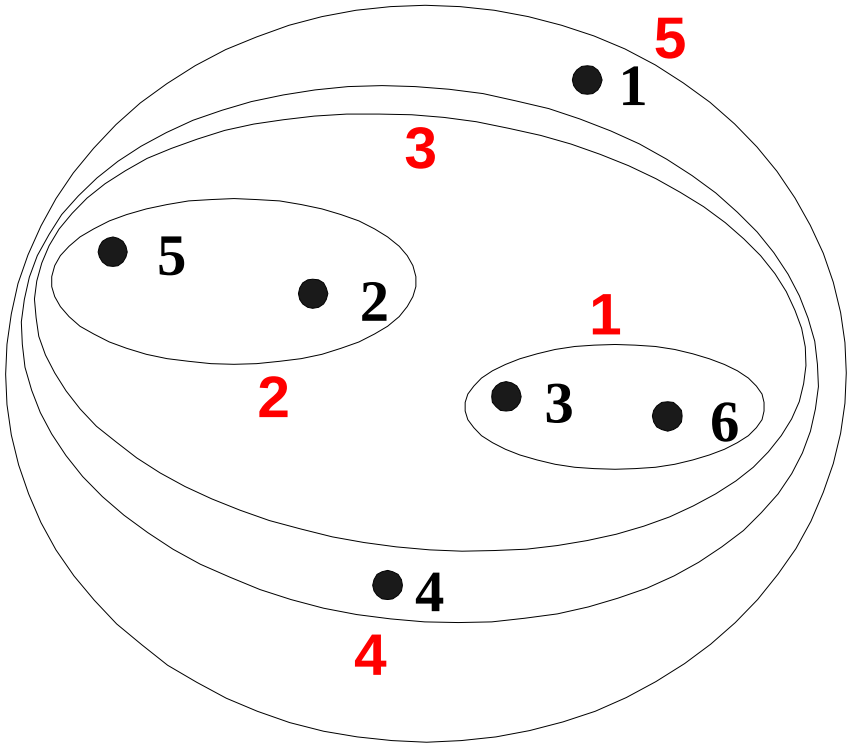
Sličnost klastera: MIN ili jedna veza

- Sličnost između dva klastera se određuje na osnovu dve najsličnije (najbliže) tačke u različitim klasterima
  - Određuje se jednim parom tačaka, tj. jednom vezom na grafu sličnosti

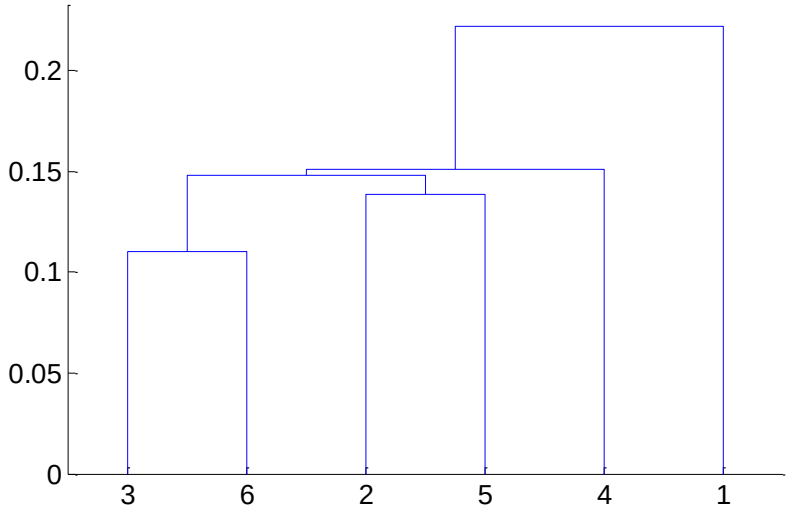
	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00



Hijerarhijsko klasterovanje: MIN

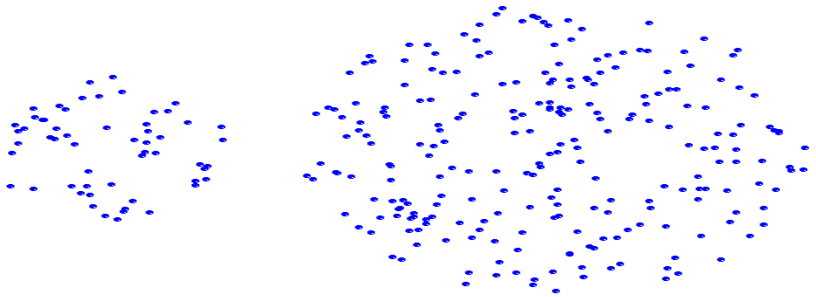


Ugneždeni klasteri

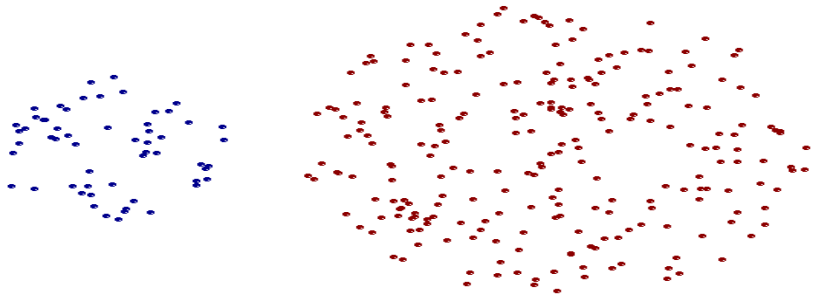


Dendrogram

Prednosti MIN



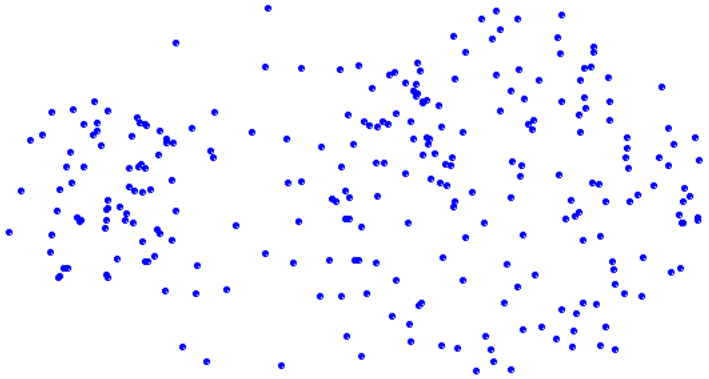
Originalne tačke



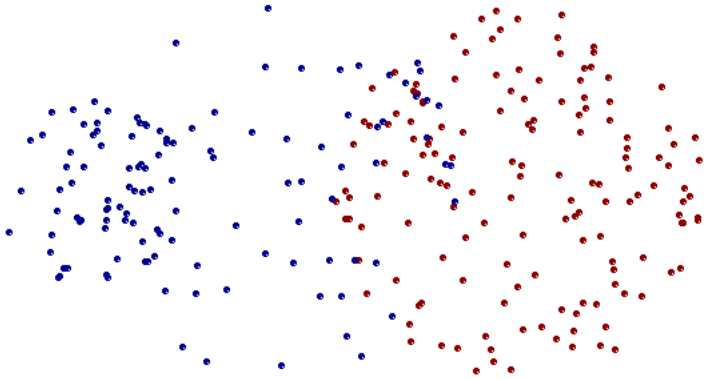
Dva klastera

- **Može da radi sa neeliptičkim oblicima**

Nedostaci MIN



Originalne tačke



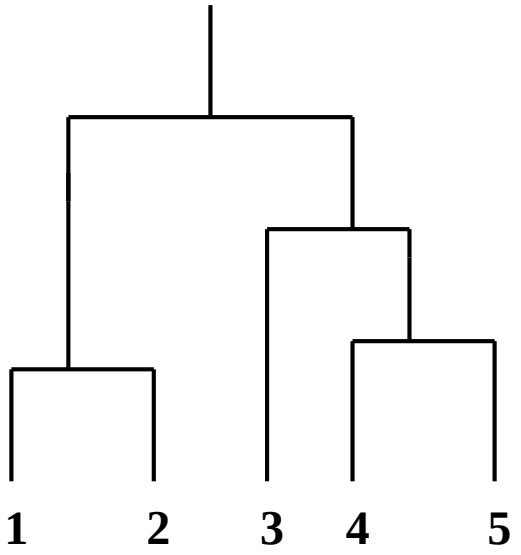
Dva klastera

- **Osetljivost na šum i elemente van granica**

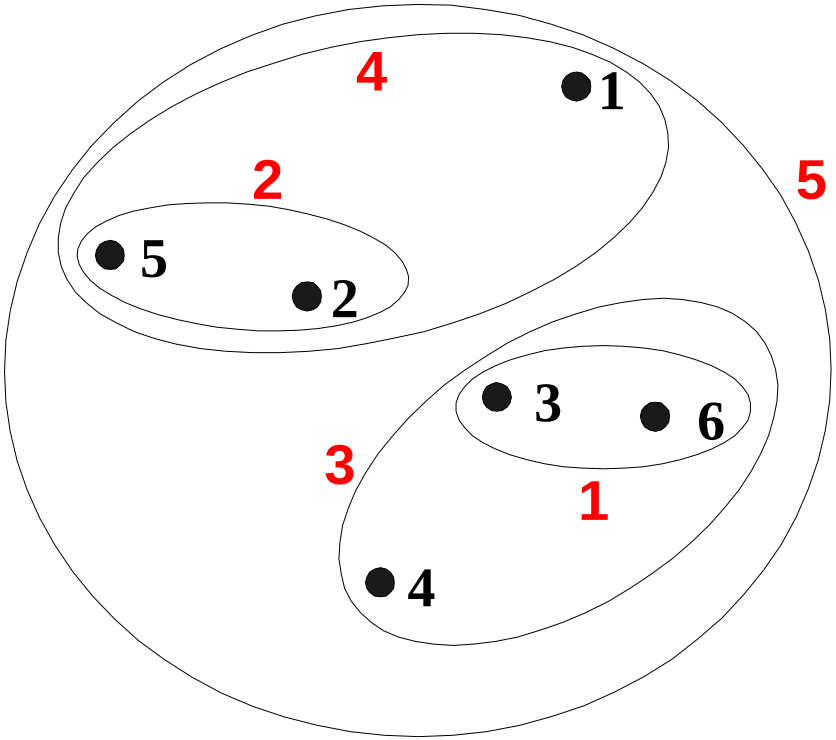
Sličnost klastera: MAX ili kompletna veza

- Sličnost između dva klastera se određuje na osnovu dve najmanje slične (najdalje) tačke u različitim klasterima
  - Određuju se parovi svih tačaka u različitim klasterima

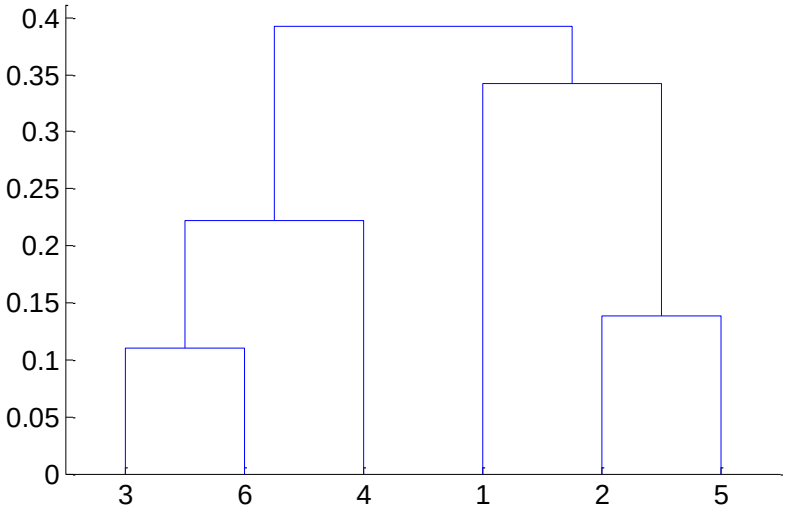
	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00



Hijerarhijsko klasterovanje: MAX

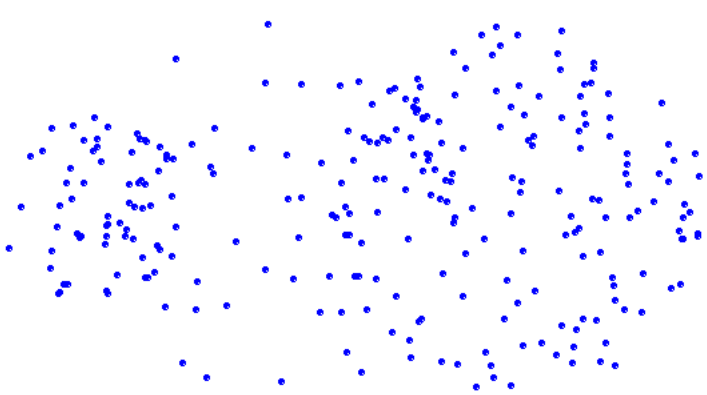


Ugneždeni klasteri

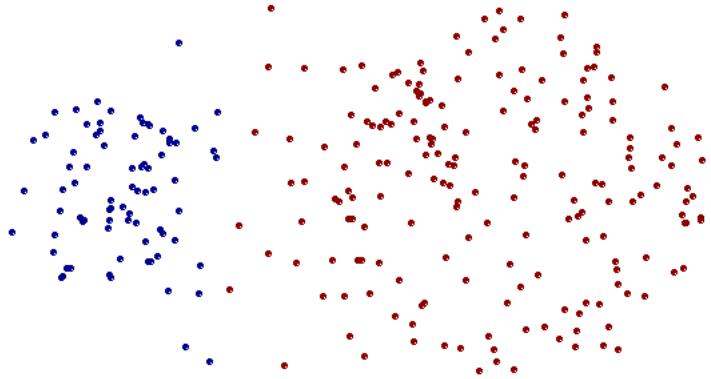


Dendrogram

Prednosti MAX



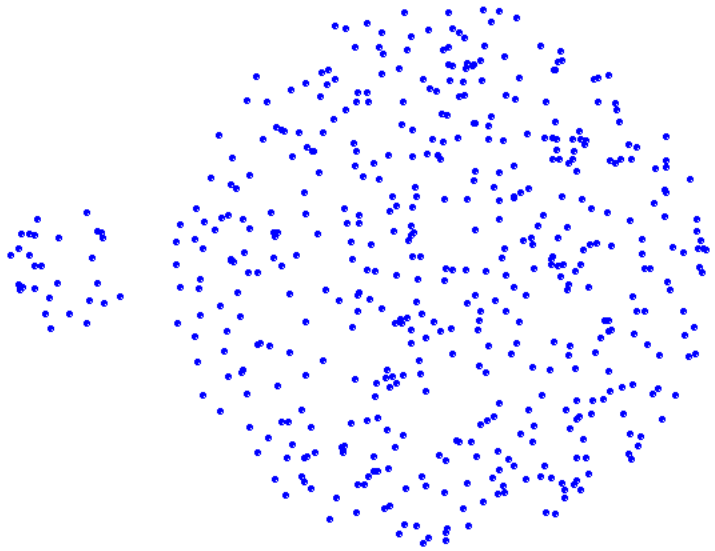
Originalne tačke

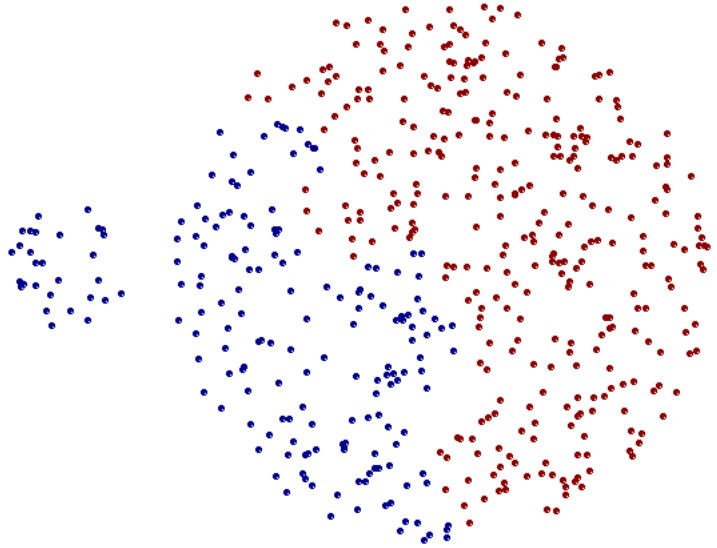


Dva klastera

- **Manje je osetljiva na šum i i elemente van granica**

Ograničenja MAX





Originalne tačke **Dva klastera**

- **Tendencija je da se razbijaju veliki klasteri**
- **Naklonost ka globularnim klasterima**

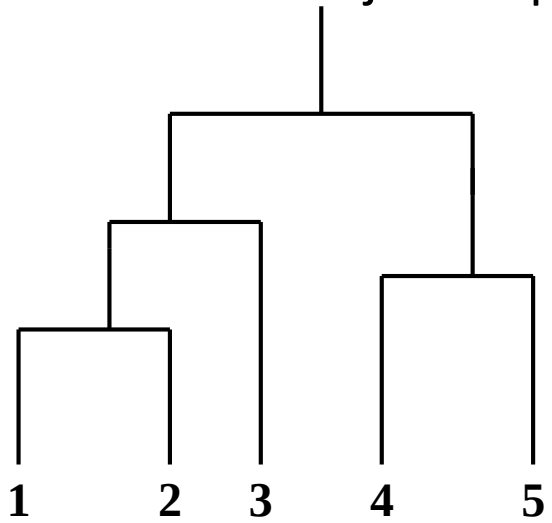
Sličnost klastera: prosek grupe

- Sličnost dva klastera je prosečna vrednost sličnosti parova tačaka u dva klastera

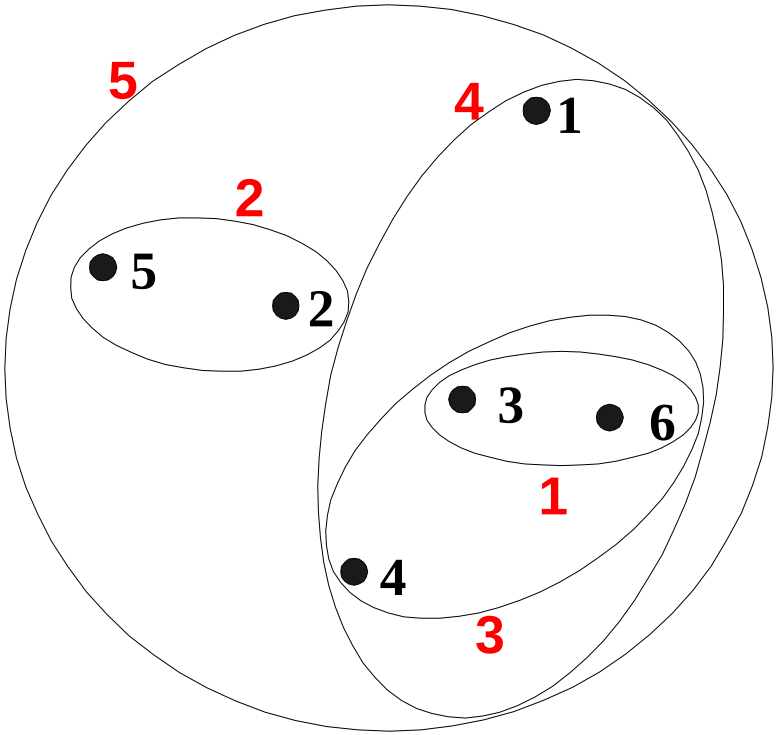
$$\text{proximity}(\text{Cluster}_i, \text{Cluster}_j) = \frac{\sum_{\substack{p_i \in \text{Cluster}_i \\ p_j \in \text{Cluster}_j}} \text{proximity}(p_i, p_j)}{|\text{Cluster}_i| * |\text{Cluster}_j|}$$

- Potrebno je koristiti prosečnu povezanost za skalabilnost jer ukupna sličnost favorizuje veće klastere

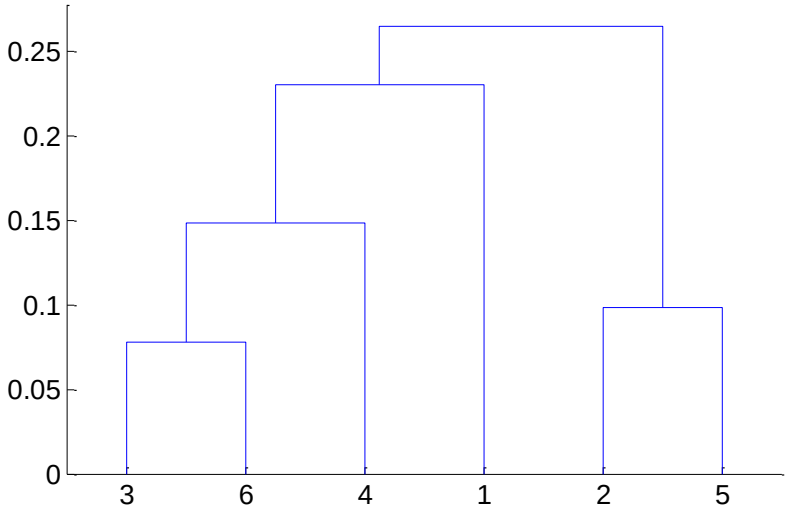
	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00



Hijerarhijsko klasterovanje: prosek grupe



Ugneždeni klasteri



Dendrogram

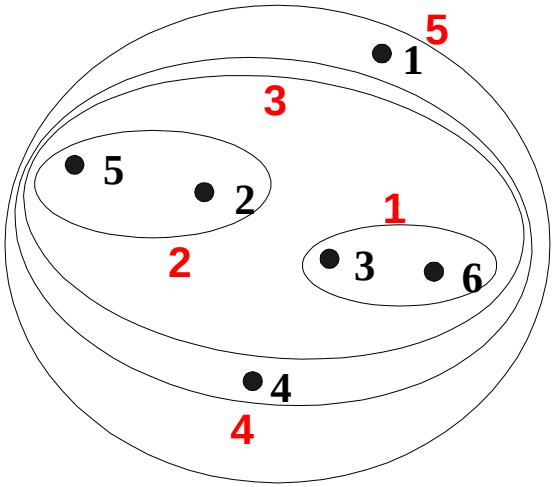
Hijerarhijsko klasterovanje: prosek grupe

- Kompromis između jedne veze i kompletne veze
- Prednosti
  - Manja osetljivost na šum i elemente van granica
- Nedostaci
  - Naklonost ka globularnim klasterima

Sličnost klastera: Ward-ov metod

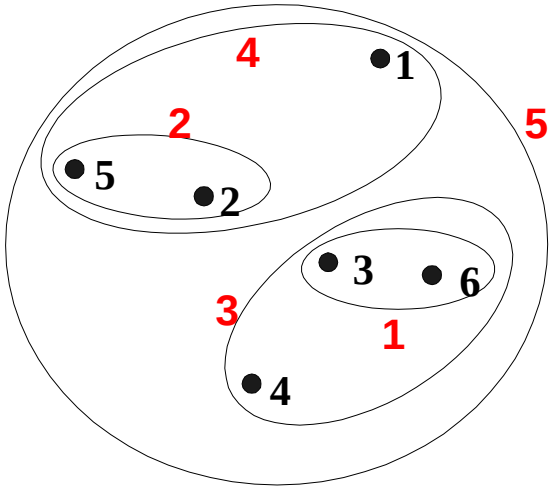
- Sličnost dva klastera je zasnovana na povećanju kvadrata greške pri sakupljanju dva klastera
  - Slično kao prosek grupe kod koga se rastojanje meri kao kvadrat rastojanja tačaka
- Manja osetljivost na šum i elemente van granica
- Naklonost ka globularnim klasterima
- Hijerarhijski analogon K-sredina
  - Može da se koristi za inicijalizaciju K-sredina

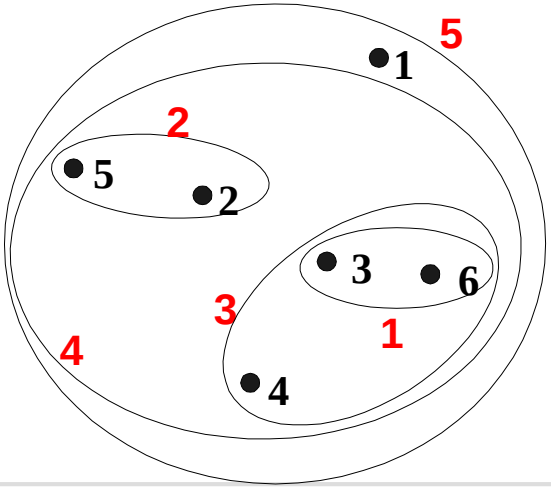
Hijerarhijsko klasterovanje: poređenje



MIN

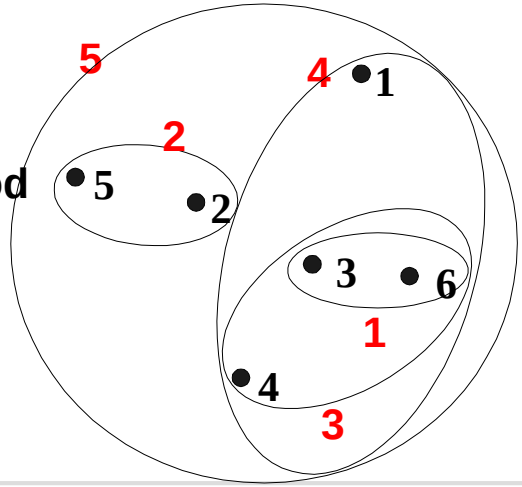
MAX





Prosek grupe

Ward-ov metod



Hijerarhijsko klasterovanje: prostorni i vremenski zahtevi

- Zahteva $O(N^2)$ prostora jer koristi matricu sličnosti (N je broj tačaka)

- Zahteva $O(N^3)$ vremena u najvećem broju slučajeva
  - Postoji N koraka; u svakom od njih se matrica sličnosti ažurira i pretražuje
  - Koristeći modifikovan pristup kompleksnost vremenskih zahteva može da se redukuje na $O(N^2 \log(N))$

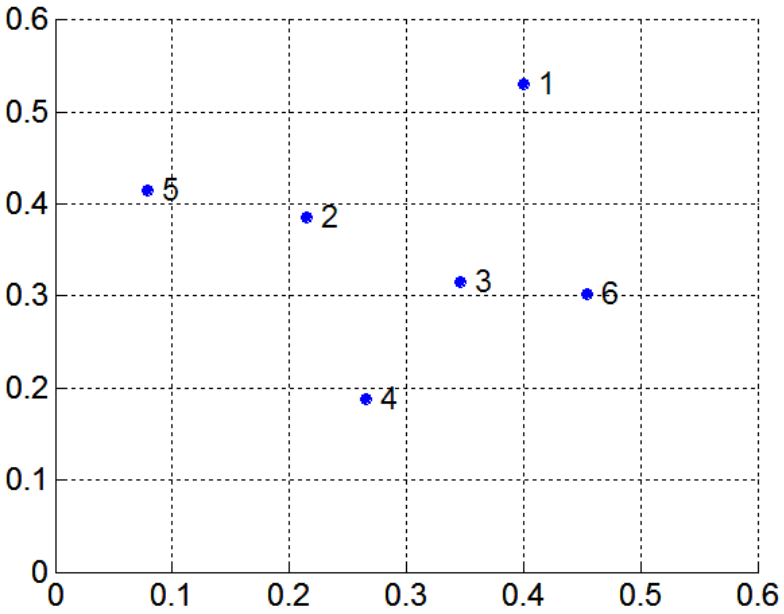
Hijerarhijsko klasterovanje: problemi i ograničenja

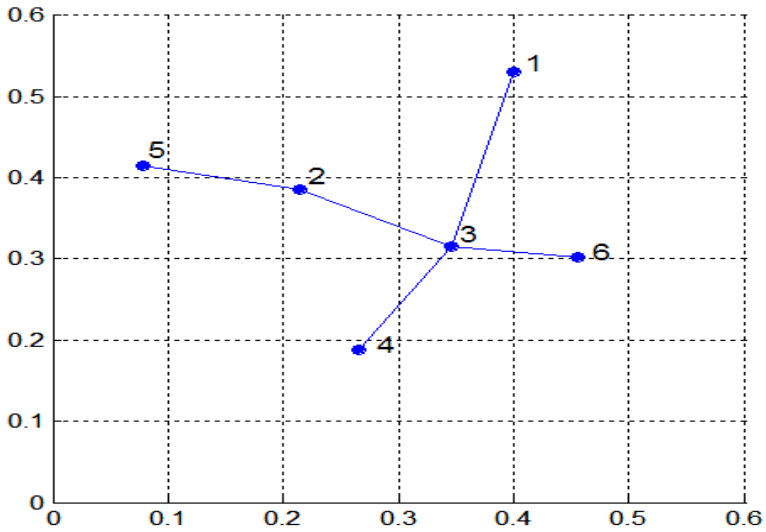
- Jednom spojeni kalsteri ne mogu više da se razdvoje; doneta odluka ne može da se opozove
- Nema ciljne funkcije koja zahteva minimizaciju

- Različite sheme imaju jedan ili više problema sa
  - Osetljivošću na šum i elemente van granica
  - Rukovanjem sa klasterima različite veličine
  - Radom sa klasterima koji imaju konkveksan oblik
  - Razbijanjem velikih klastera

MST: razdvajajuće hijerarhijsko klasterovanje

- Izgradnja MST (eng. *Minimum Spanning Tree*)
  - Početi sa drvetom koje sadrži neku tačku
  - U uzastopnim koracima tražiti parove najbližih tačaka oblika (p, q) tako da je jedna tačka (p) u tekućem drvetu dok druga (q) nije
  - dodati q u drvo i povezati p i q jednom granom





MST: razdvajajuće hijerarhijsko klasterovanje

- Opotreba MST za formiranje hijerahije klastera

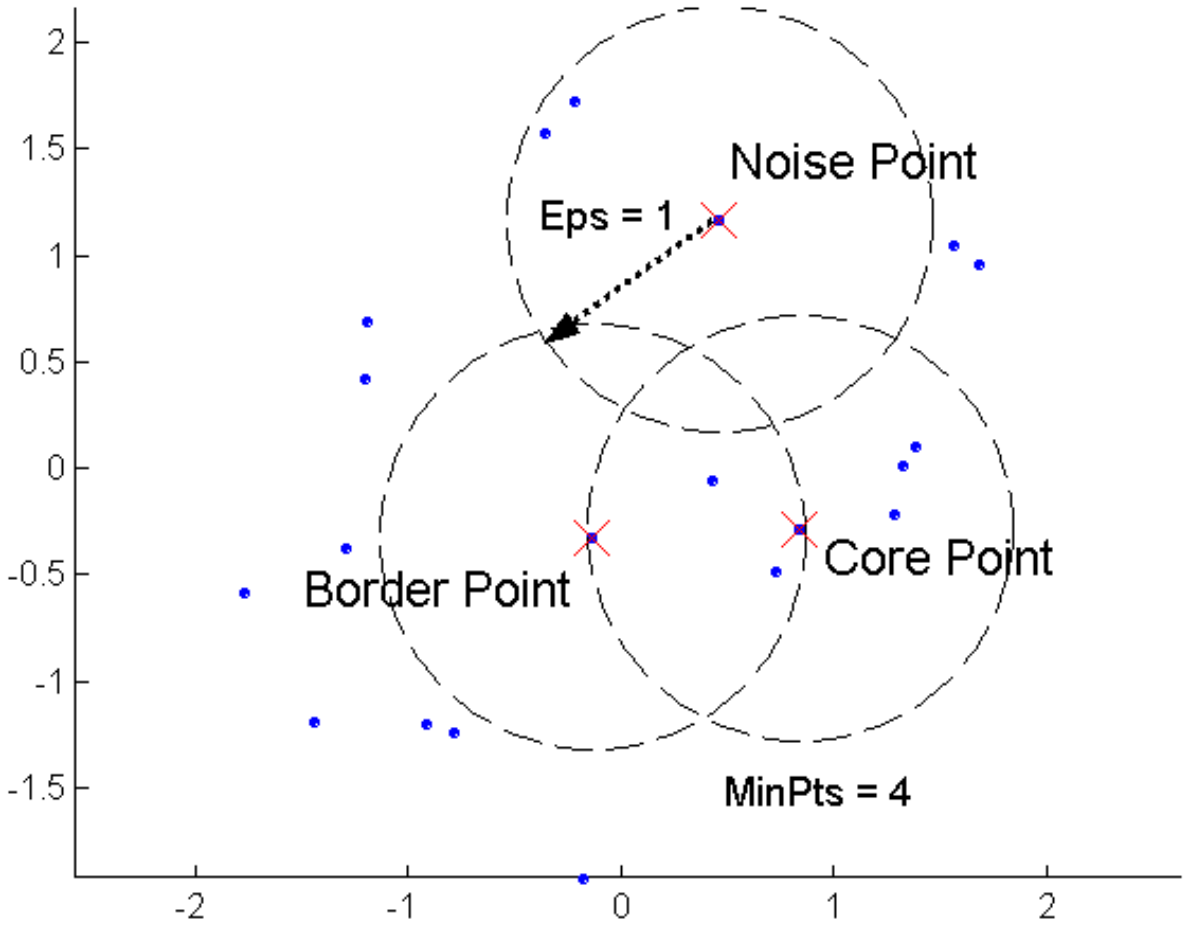
Algorithm 7.5 MST Divisive Hierarchical Clustering Algorithm

1: Compute a minimum spanning tree for the proximity graph.
2: **repeat**
3: Create a new cluster by breaking the link corresponding to the largest distance (smallest similarity).
4: **until** Only singleton clusters remain

DBSCAN

- DBSCAN (Density-Based Spatial Clustering od Application with Noise)
- DBSCAN je algoritam zasnovan na gustini
  - Gustina = broj tačaka unutar određenog poluprečnika (Eps)
  - Tačka A je tačka u jezgru ako je više od zadatog broja tačaka (MinPts) unutar Eps
    - U pitaju su tačke unutar klastera
  - Tačka A je tačka na granici ako ima manje od MinPts unutar Eps, ali je susedna sa tačkom u jezgru
  - Tačka A je šum ako nije ni tačka u jezgru ni tačka na granici

DBSCAN: Tačke u jezgru, na granici i šum



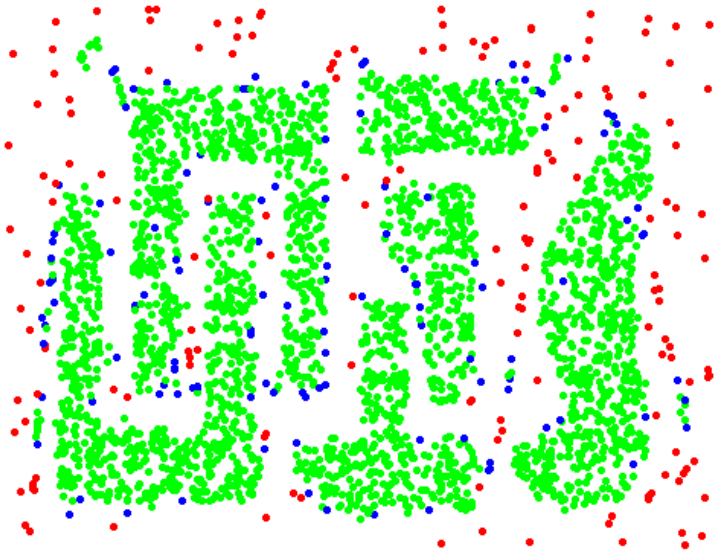
DBSCAN Algoritam

- Eliminisati tačke koje su šum
- Izvršiti klasterovanje na preostalim tačkama

```
current_cluster_label ← 1
for all core points do
  if the core point has no cluster label then
    current_cluster_label ← current_cluster_label + 1
    Label the current core point with cluster label current_cluster_label
  end if
  for all points in the Eps-neighborhood, except i^th the point itself do
    if the point does not have a cluster label then
      Label the point with cluster label current_cluster_label
    end if
  end for
end for
```

DBSCAN: Tačke u jezgru, na granici i šum

Originalne tačke

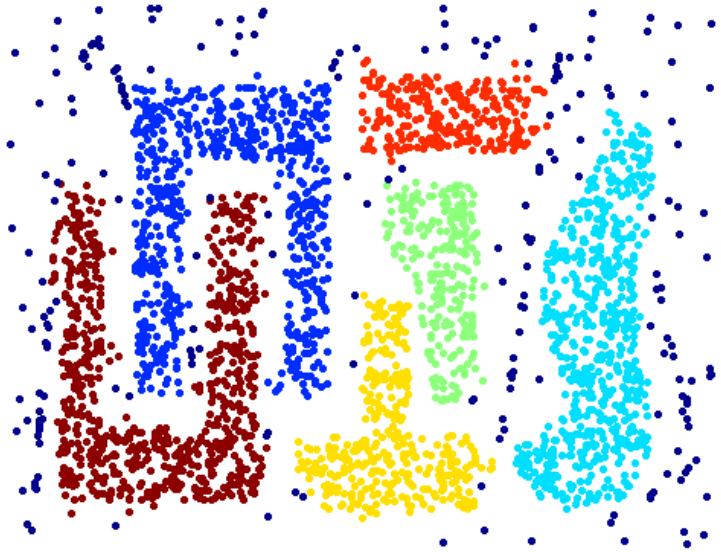


Tip tačke: jezgro, granica i šum

Eps = 10, MinPts = 4

Kada DBSCAN ispravno radi

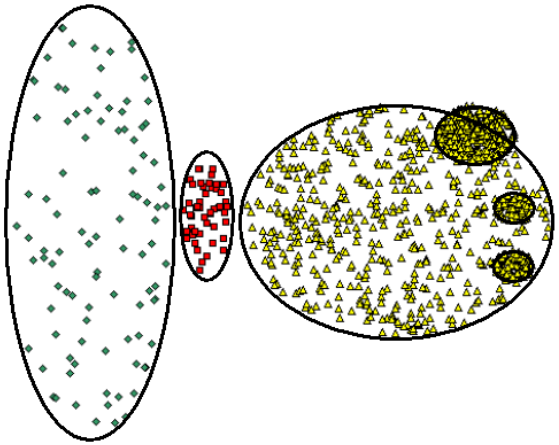
Originalne tačke



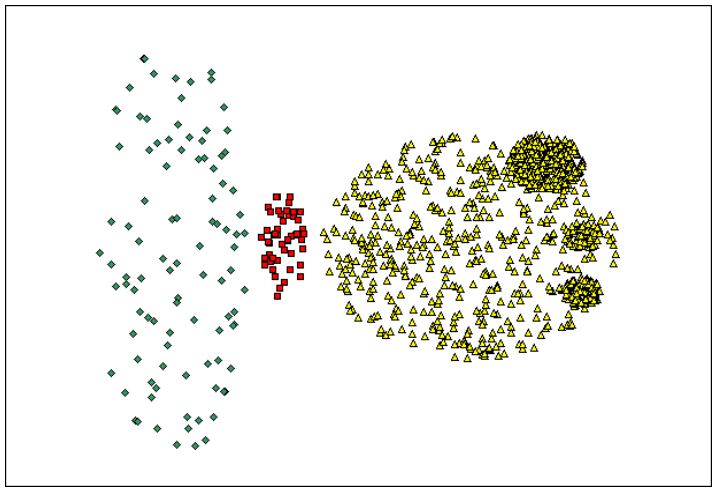
Klasteri

- **Otporan na šum**
- **Može da radi sa klasterima različitih oblika i veličina**

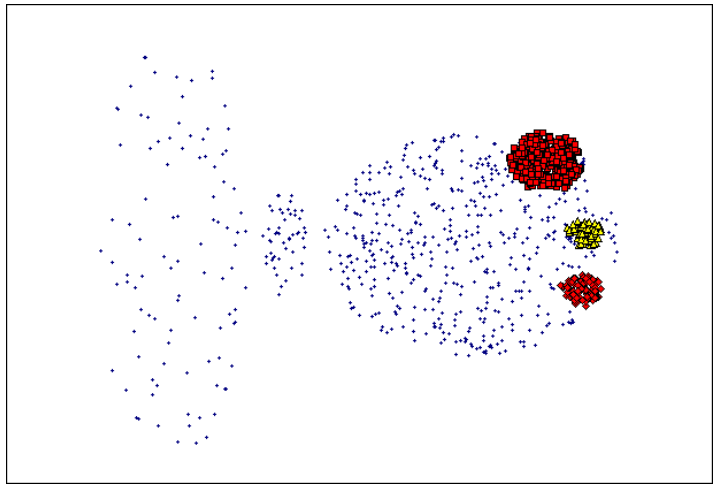
Kada DBSCAN ne radi ispravno



Originalne tačke



(MinPts=4, Eps=9.75).

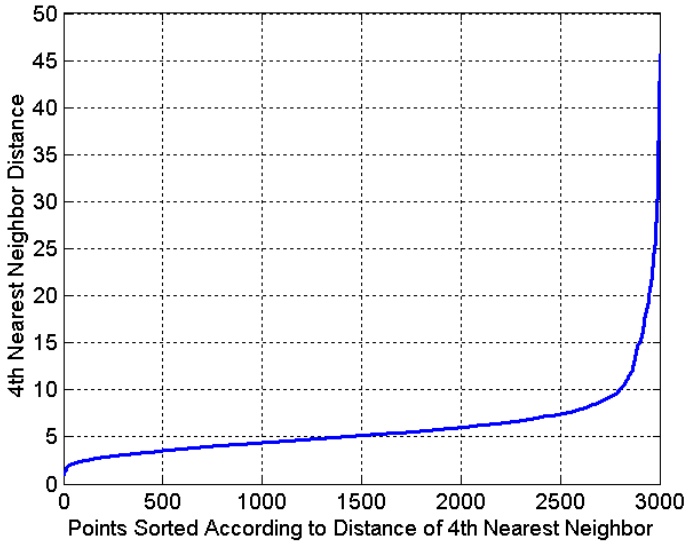


(MinPts=4, Eps=9.92)

- **Različite gustine**
- **Podaci sa velikim brojem dimenzija**

DBSCAN: Određivanje EPS i MinPts

- Ideja je da je za tačku u klasteru njenih k najbližih suseda je skoro na istom rastojanju
- Tačke koje su šum imaju k najbližih suseda na većem rastojanju
- Nacrtati sortirana rastoajanj od svake tačke do njenih ka najbližih suseda

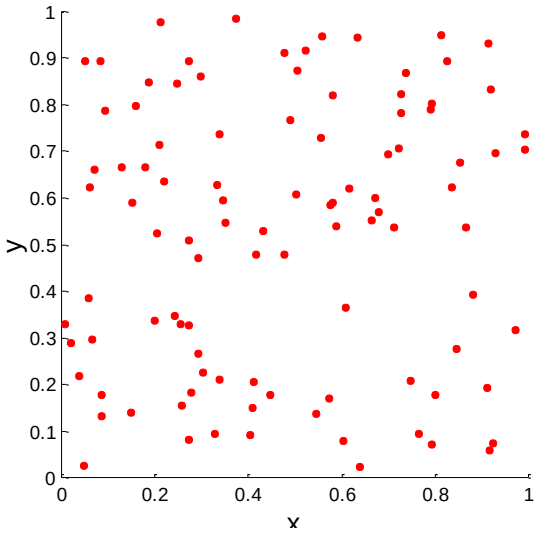


Kvalitet klastera

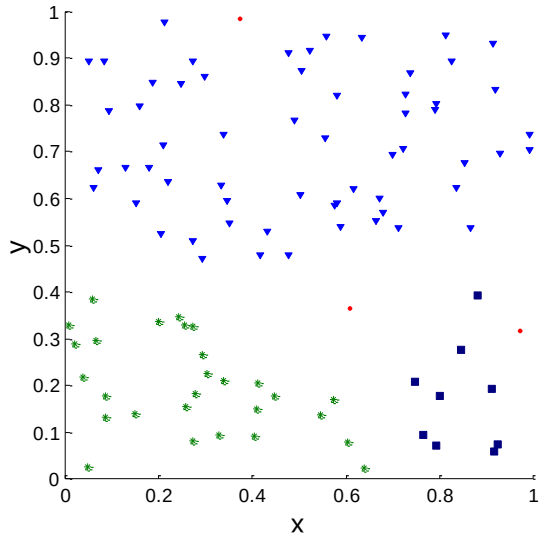
- Za klasifikaciju postoji više mera ua izračunavanje kvaliteta modela
  - Preciznost, punovažnost, odziv (pokrivanje)
- Šta je odgovarajući analogon za klaster analizu?
- Vizuelna ocena
- Zašto onda uopšte želimo da proverimo kvalitet klasterovanja?
  - Radi izbegavanja nađenih obrazaca u šumu
  - Radi poređenja algoritama klasterovanja
  - Radi poređenja dva skupa klastera
  - Radi poređenja dva klastera

Klasteri nađeni u slučajnim podacima

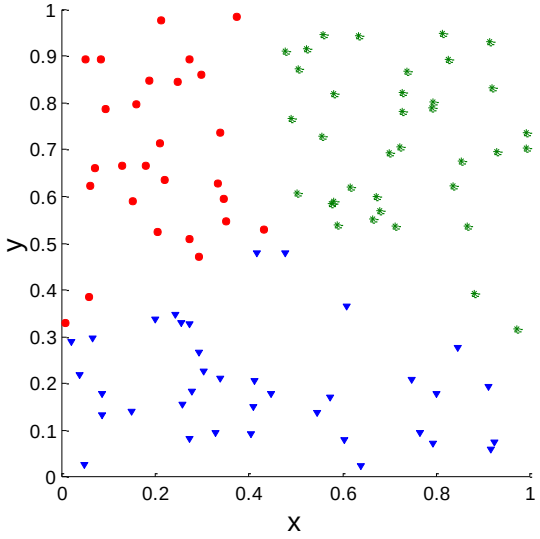
Slučajne tačke



DBSCAN



K-means



Potpuna povezanost