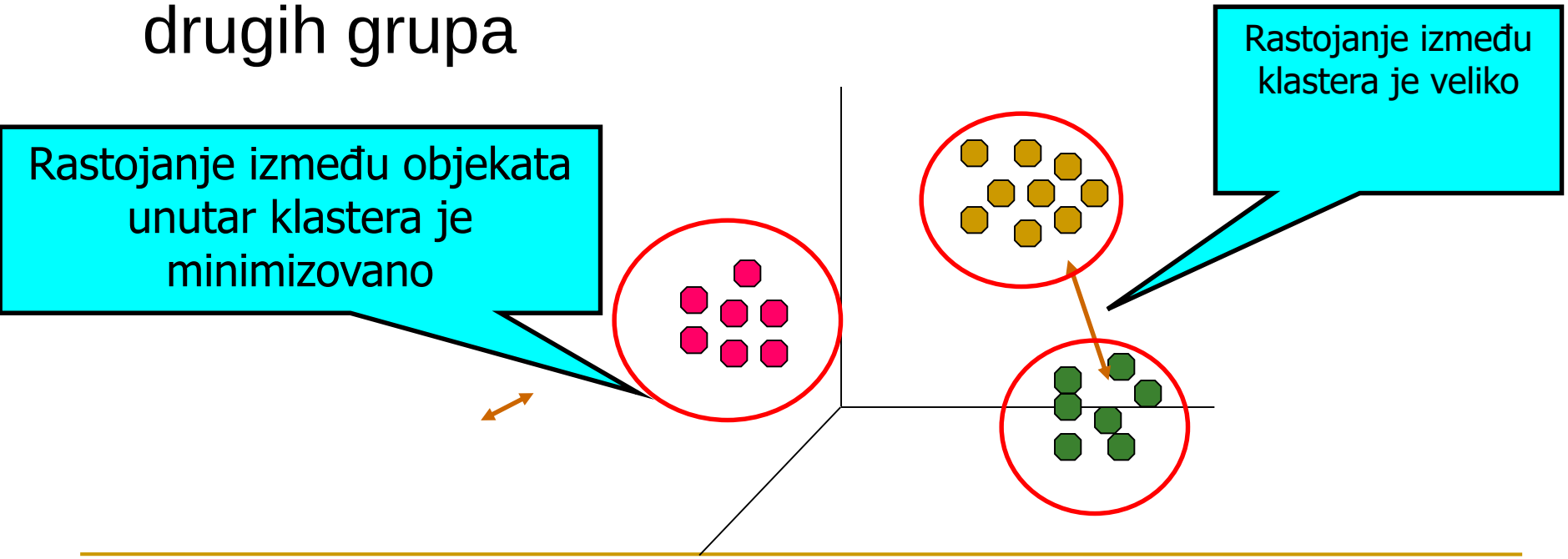
Klasterizacija

Klaster analiza

- Identifikovanje grupa (klastera) objekata tako da su objekti unutar grupe međusobno slični i istovremeno različiti u odnosu na objekte iz drugih grupa



Primjene klaster analize

- Clustering for understanding
  - Grupisanje povezanih veb dokumenata
  - Grupisanje gena sa sličnim funkcionalnostima
  - Grupisanje proizvoda sa sličnim promjenama cijene
- Clustering for utility
  - Smanjivanje obima velikih skupova podataka

Šta nije klaster analiza?

- Klasifikacija
  - Nadgledano učenje, za svaki objekat poznata je klasa kojoj pripada
- “Jednostavne” podjele (segmentacija)
  - Particionisanje na osnovu početnog slova u imenu
- SQL upit sa GROUP BY

Šta je klaster?

Koliko ima klastera u datom skupu objekata?

6 klastera

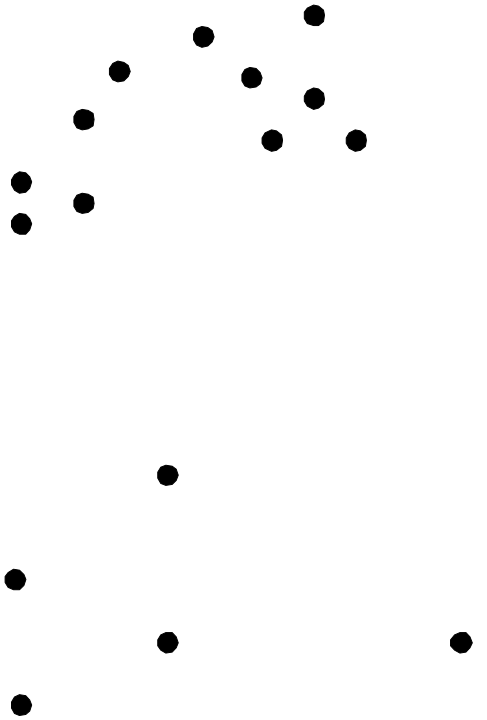
2 klastera

4 klastera

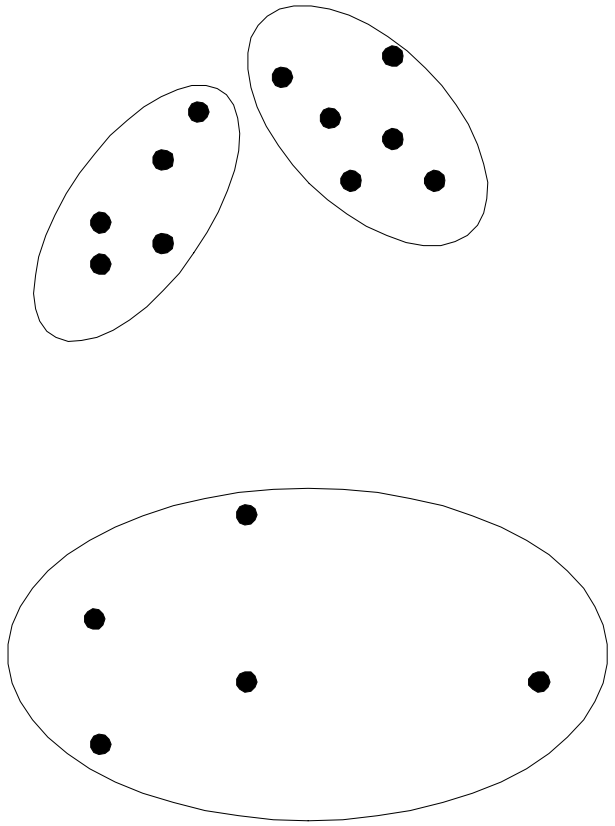
Tehnike klasterizacije

- Tehnike particionisanja
  - Podjela datog skupa objekata na disjunktne klastere tako da svaki objekat pripada samo jednom klasteru
- Hijerarhijske tehnike
  - Podjela datog skupa objekata na ugnježdene klastere koji su hijerarhijski uređeni u obliku stabla

Tehnike particionisanja

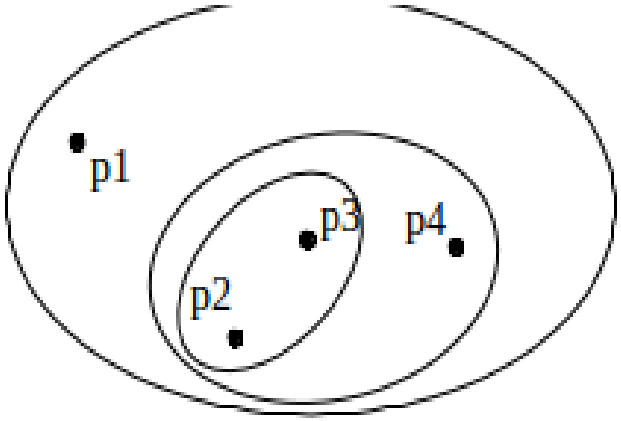


Skup objekata

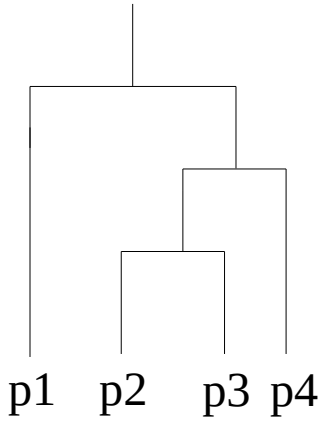


Generisani klasteri

Hijerarhijske tehnike



Skup podataka



Dendogram reprezentacija

Dodatni kriterijumi za podjelu tehnika klasterizacije

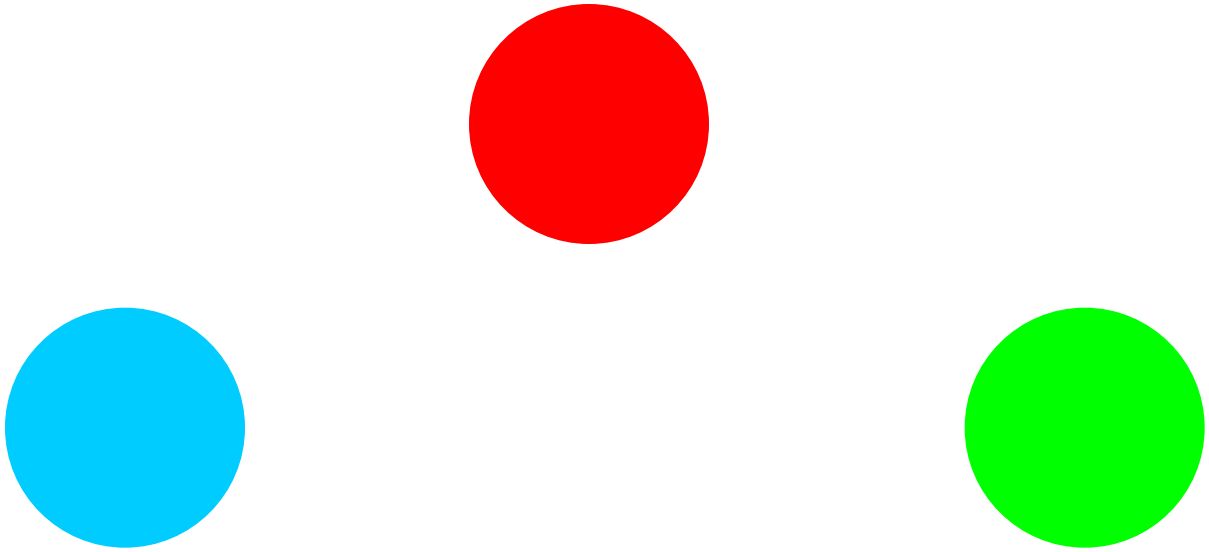
- Exclusive versus non-exclusive
  - U non-exclusive klasterima jedan objekat može da pripada većem broju klastera
- Fuzzy versus non-fuzzy
  - U fuzzy klasterima objekat pripada svakom klasteru sa nekom težinom $0 \leq w \leq 1$
- Partial versus complete
  - Partial klasterizacija obuhvata samo dio polaznog skupa

Tipovi klastera

- Well-separated clusters
- Center-based clusters
- Contiguous clusters
- Density-based clusters
- Property or Conceptual
- Described by an Objective Function

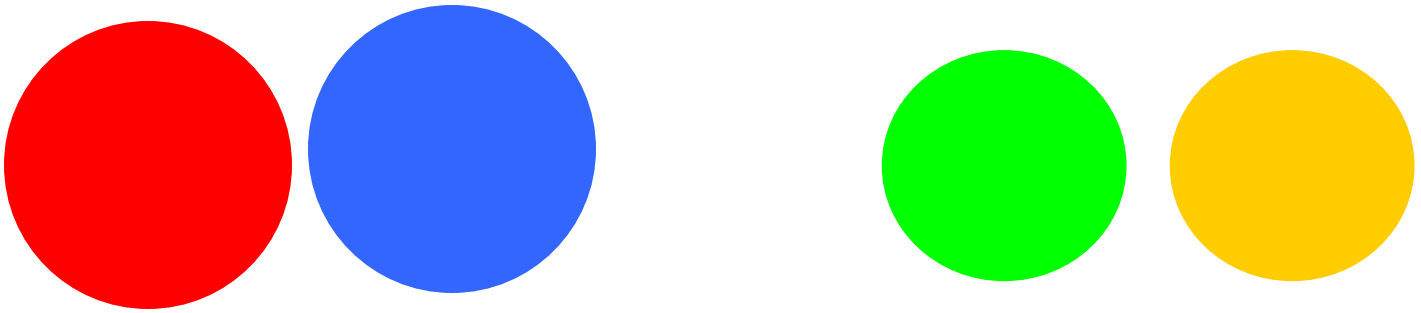
Tipovi klastera: well-separated

- Klasteri u kojima je ma koji objekat bliži svim objektima iz istog klastera nego ma kom objektu izvan istog klastera



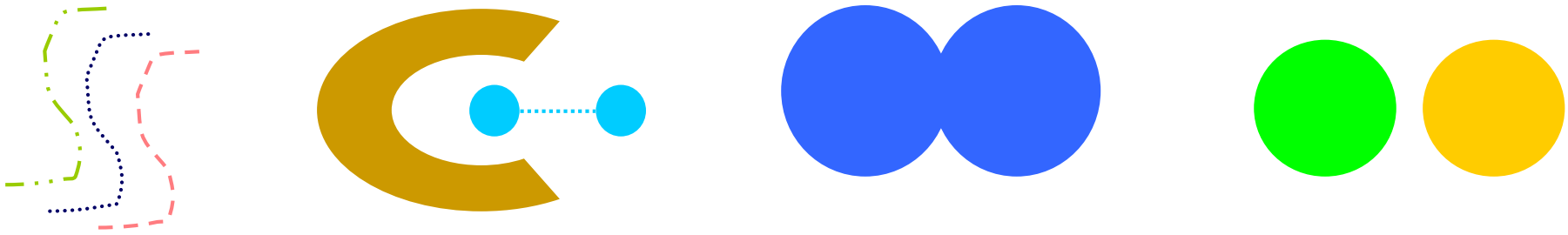
Tipovi klastera: center-based clusters

- Klasteri u kojima je ma koji objekat bliži centru klastera kojem pripada nego centru ma kog drugog klastera



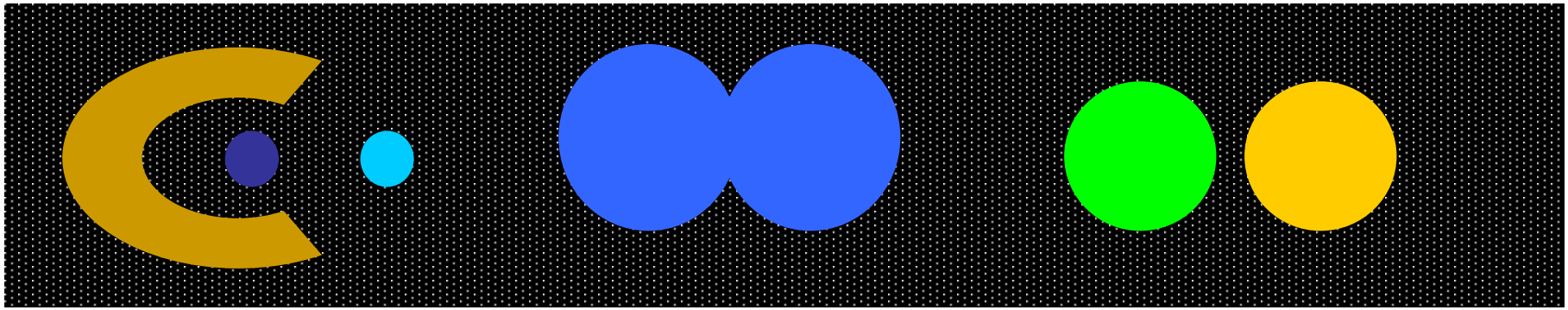
Tipovi klastera: contiguous clusters

- Klasteri u kojima je ma koji objekat bliži bar jednom objektu iz tog klastera nego ma kom objektu izvan istog klastera



Tipovi klastera: density-based clusters

- Klasteri su “regioni” sa velikom gustinom, razdvojeni regionima male gustine



Algoritmi klasterizacije

- K-means
- Agglomerative hierarchical clustering
- Density-based clustering

K-means

- Tehnika particionisanja
- Svaki klaster predstavljen je centroidom
- Objekat se pridružuje najbližem centroidu
- Broj klastera je ulazni podatak

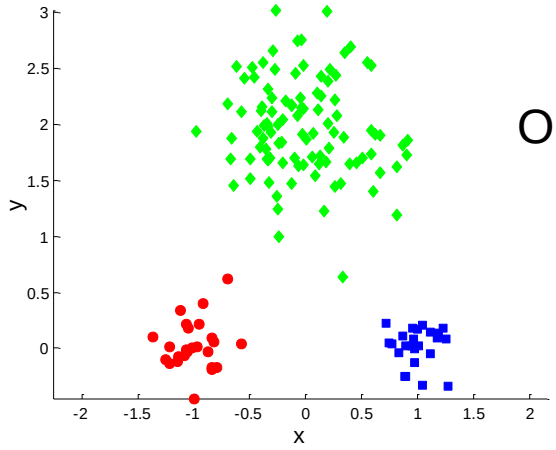
Algorithm 1 Basic K-means Algorithm.

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

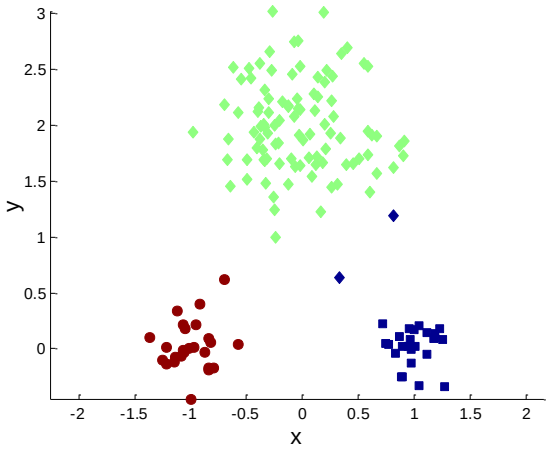
K-means (2)

- Početni izbor centroida je slučajan
- Centroidi se računaju kao aritmetička sredina objekata u klasteru
- Mjere rastojanja: Euklidsko, kosinusno, itd.
- Kriterijum zaustavljanja
  - Mali broj objekata je promijenio klaster
- Složenost O(n*k*i*d)
  - n je broj objekata, k je broj klastera, i je broj iteracija, d je broj atributa

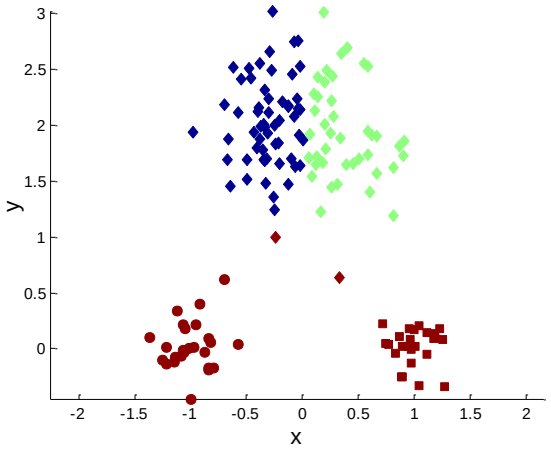
Različita klasterisanja sa K-means



Original Points

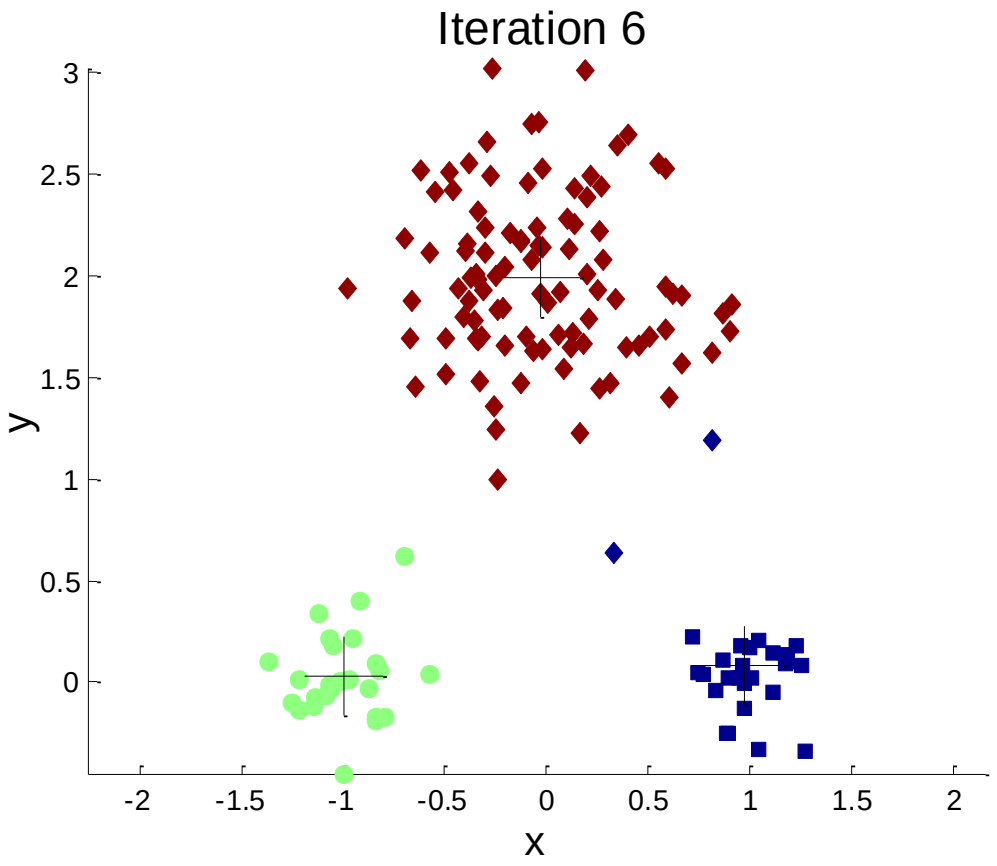


Optimal Clustering

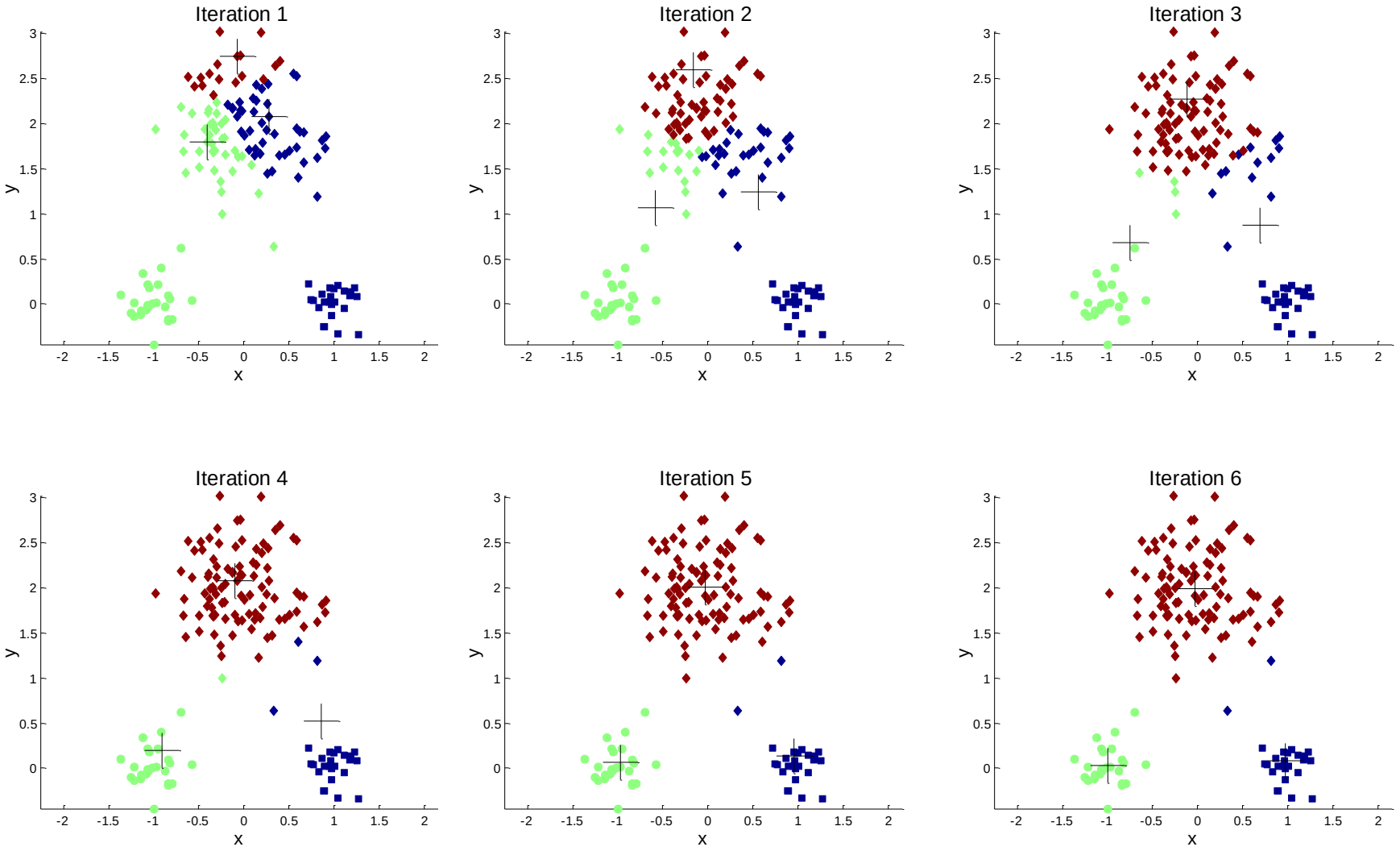


Sub-optimal Clustering

Značaj izbora centroida



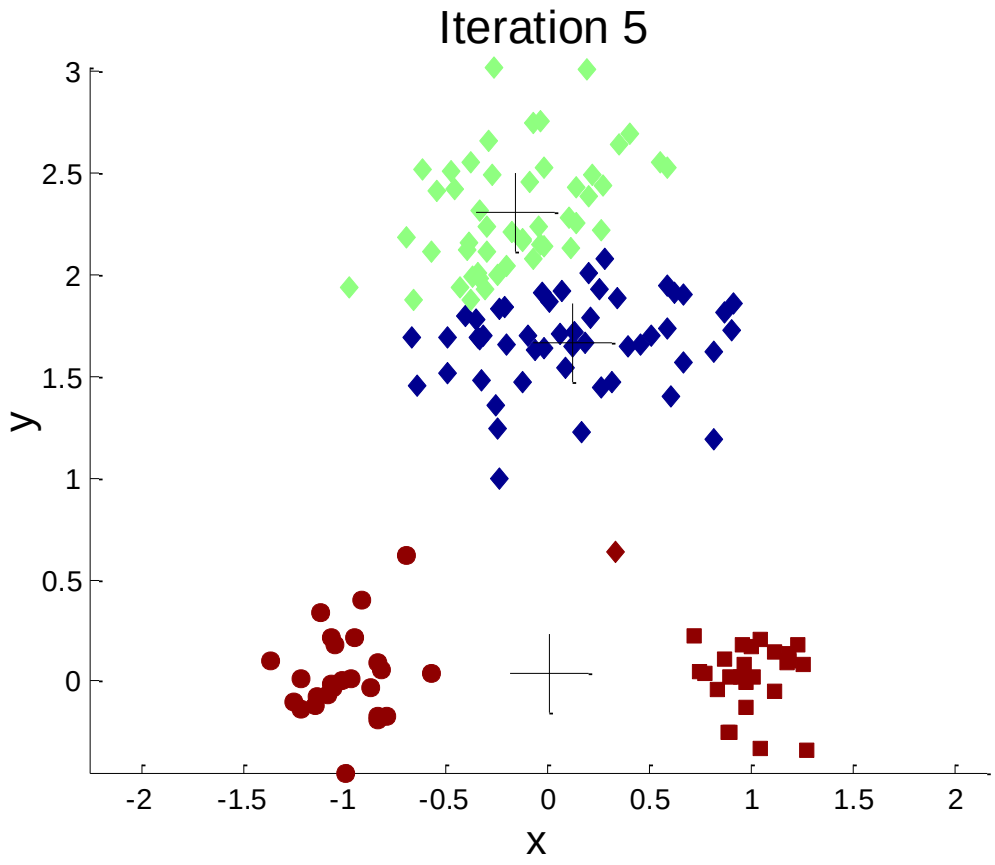
Značaj izbora centroida (2)



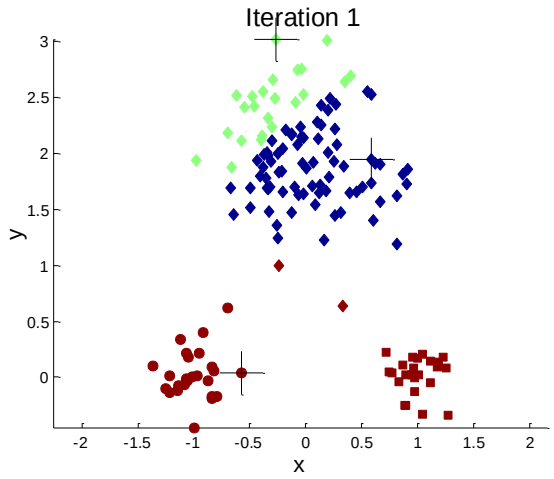
Procjena generisanih klastera

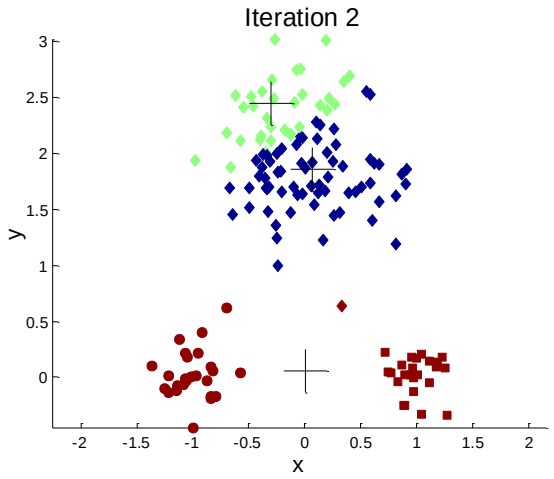
- Uobičajeno se koristi SSE
  - Za svaki objekta greška je rastojanje do najbližeg centroida
  - Za data dva skupa klastera biramo onaj sa manjom SSE
  - Način da se smanji SSE je da se poveća ukupan broj klastera

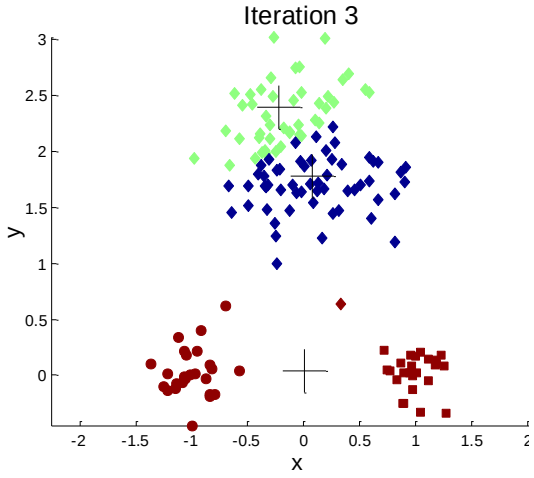
Značaj izbora centroida (3)

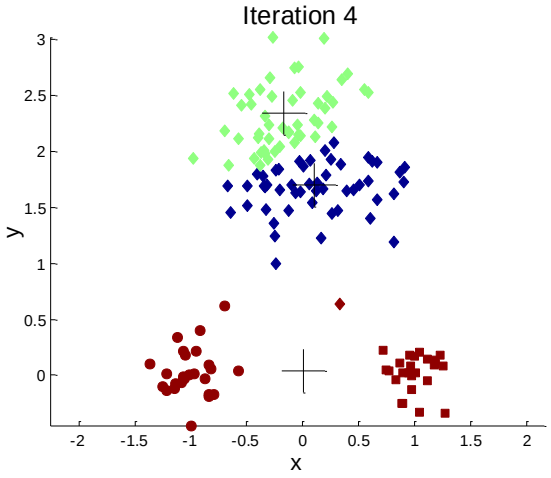


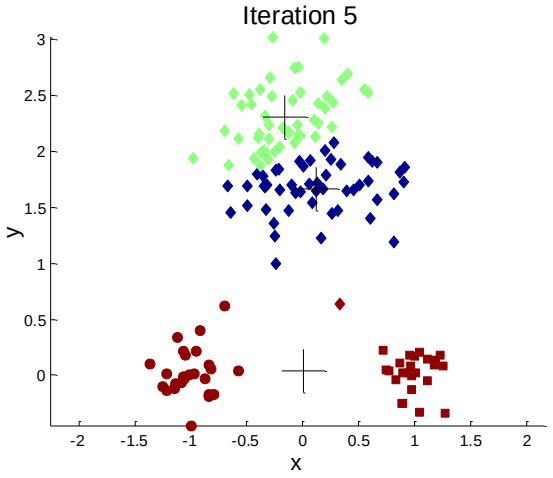
Značaj izbora centroida (4)











Biranje centroida

- Ako u skupu objekata postoji K klastera, vjerovatnoća da iz svakog klastera izaberemo po jedan centroid je mala
  - Ako svaki klaster sadrži po n objekata

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

  - Ako je k = 10, onda je P = 0.00036
  - Nekad se i pogrešno odabrani centroidi ispravno preraspodijele

Prazni klasteri

- Osnovni K-means algoritam može da generiše prazne klastere
- Neka rješenja
  - Izabira se objekat koji najviše utiče na SSE
  - Izabira se objekat iz klastera sa najvećom SSE
  - Ako postoji više praznih klastera, prethodni koraci se ponavljaju više puta

Promjena centroida inkrementalno

- U osnovnom K-means algoritmu centroidi se preračunavaju na kraju iteracije glavne petlje
- Alternativa je da se preračunavanje izvrši poslije svakog pridruživanja objekta najbližem centroidu
  - Mijenja se 0 ili 2 centroida
  - Računarski složenije
  - Redosljed kojim se objektima pristupa utiče na rezultat

Preprocesiranje i postprocesiranje

- Preprocesiranje
  - Normalizacija
  - Eliminacija izuzetaka
- Postprocesiranje
  - Brisanje malih klastera koji moguće predstavljaju izuzetke
  - Podjela klastera sa velikom SSE
  - Spajanje bliskih klastera

Bisecting K-means

- Varijanta K-means algoritma koja može da generiše klastere kao i hijerarhijske tehnike

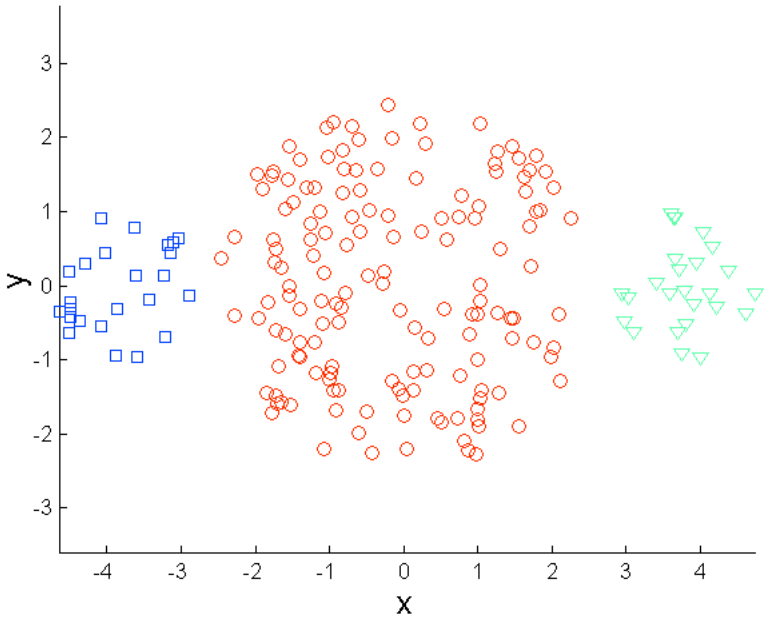
Algorithm 3 Bisecting K-means Algorithm.

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

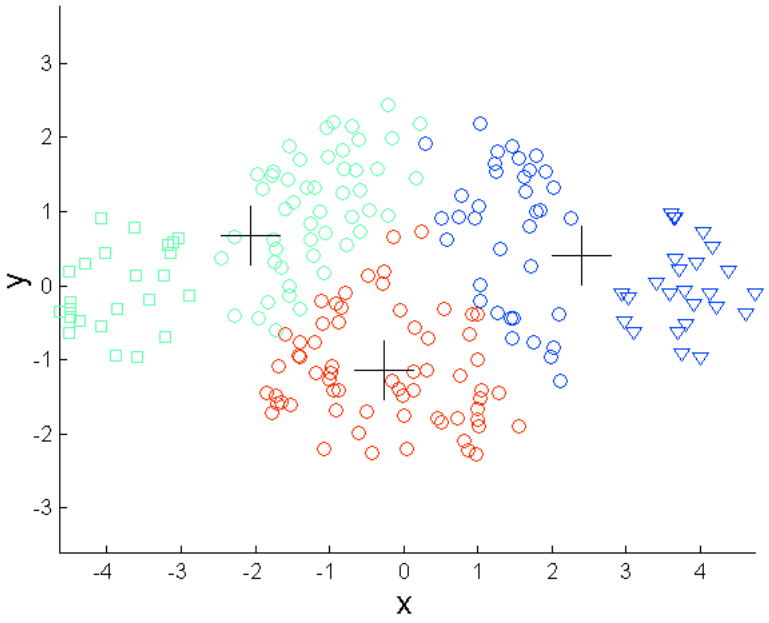
Ograničenja K-means algoritma

- Kada su klasteri
  - Različitih veličina
  - Različitih gustina
  - Ne-sfernog oblika
- Kada originalni podaci sadrže izuzetke

Ograničenja K-means algoritma: klasteri različitih veličina

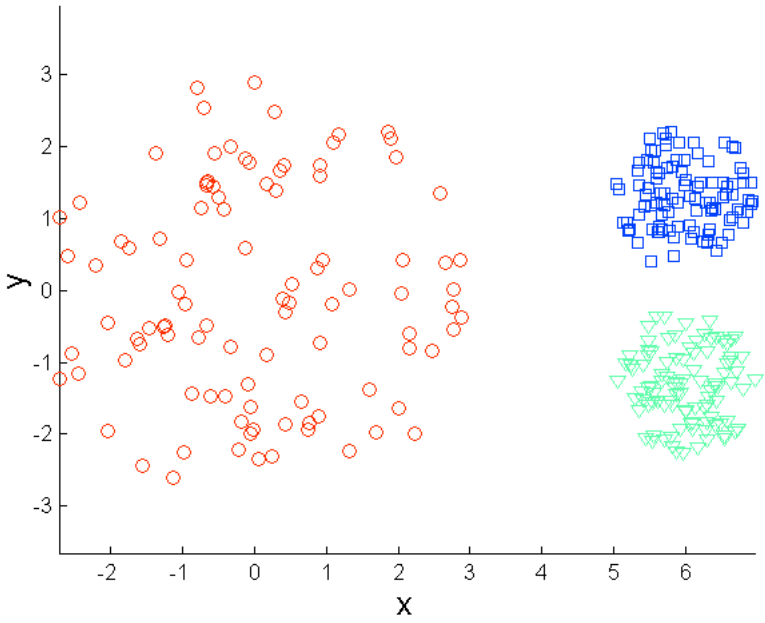


Originalne tačke

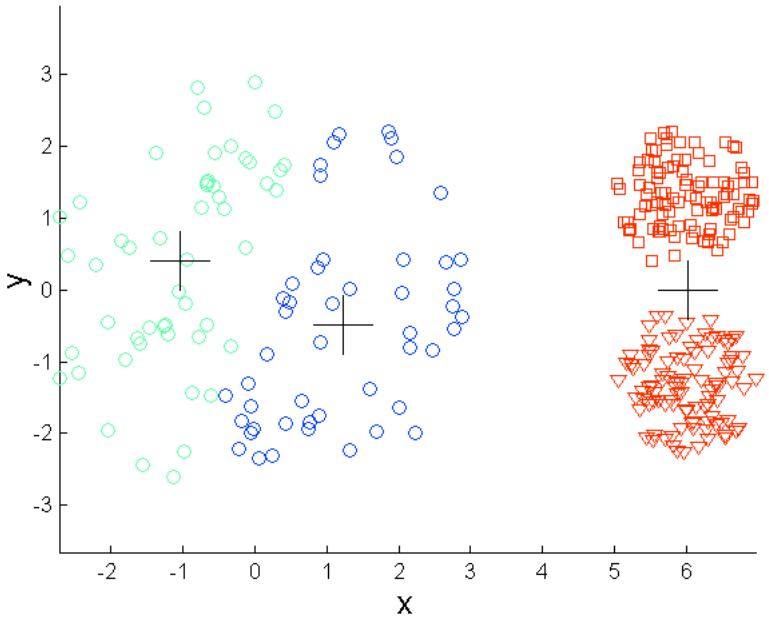


K-means (3 Klastera)

Ograničenja K-means algoritma: klasteri različitih gustina

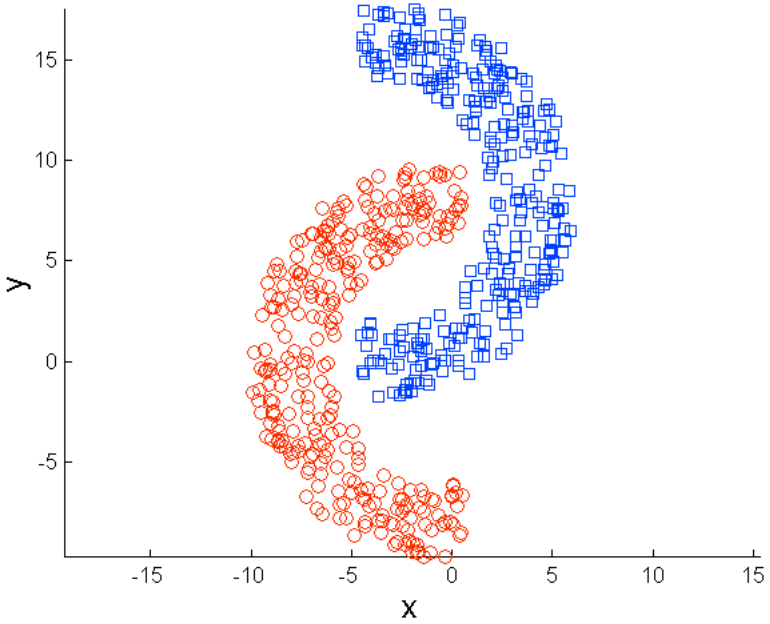


Originalne tačke

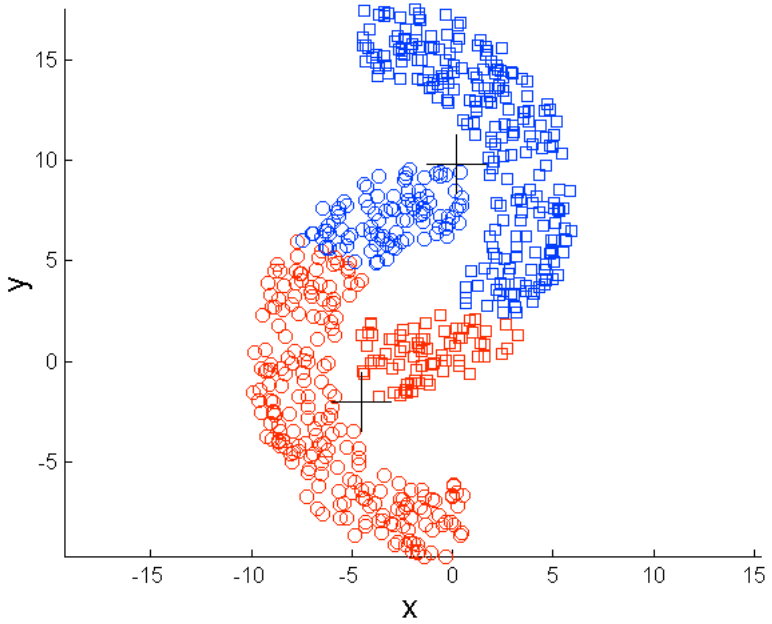


K-means (3 Klastera)

Ograničenja K-means algoritma: klasteri ne-sfernog oblika

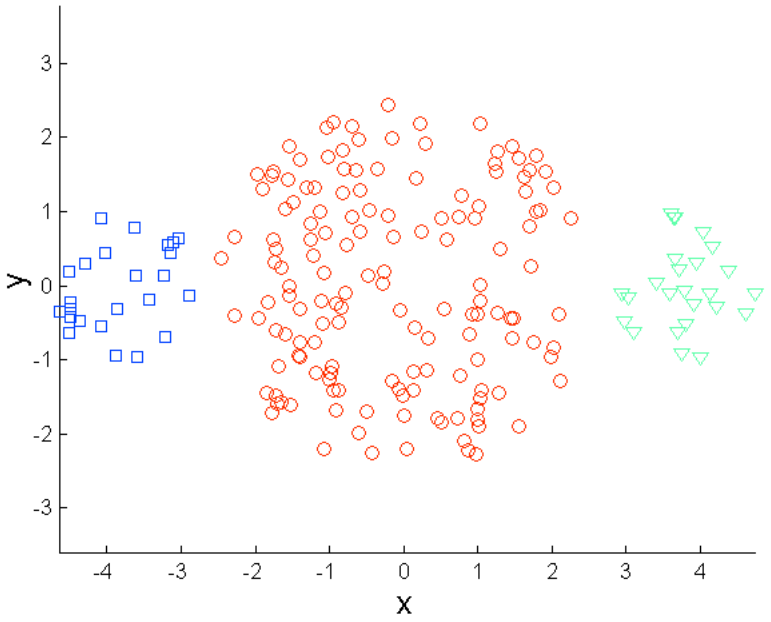


Originalne tačke

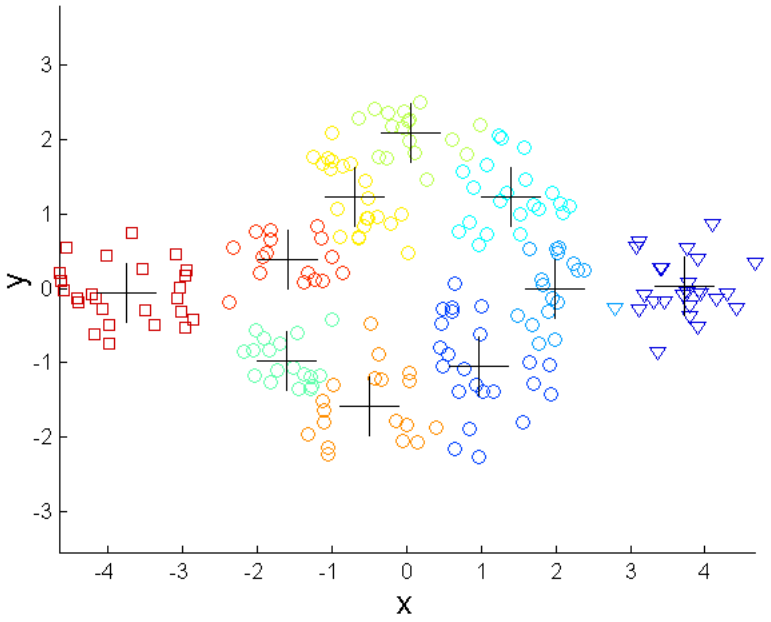


K-means (2 Klastera)

Rješenja



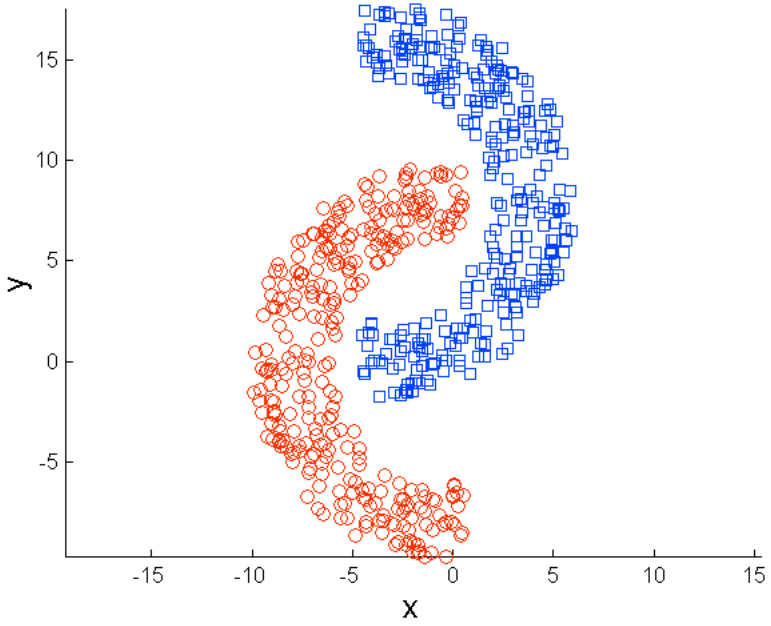
Originalne tačke



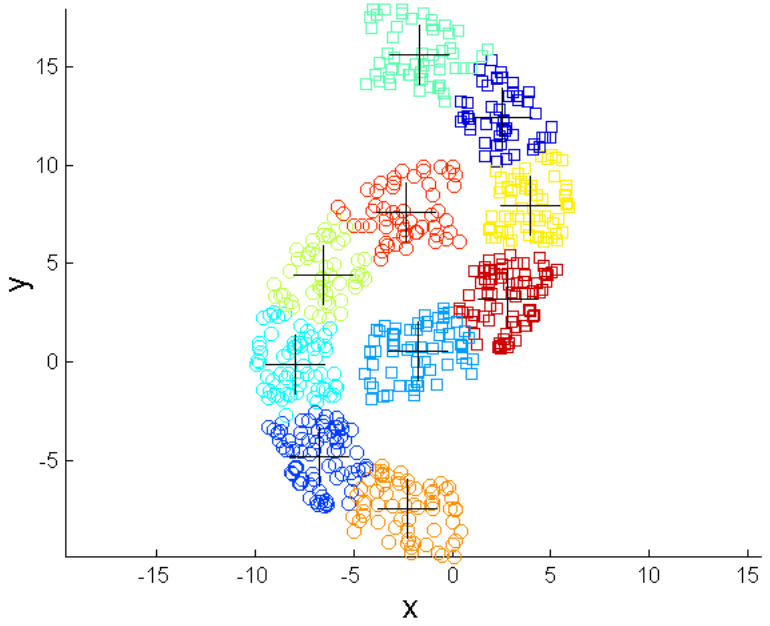
K-means Klasteri

Algoritam se pokrene sa velikim brojem klastera.

Rješenja (2)



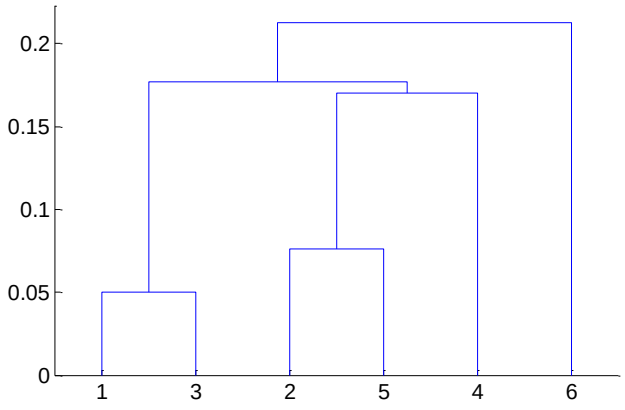
Originalne tačke

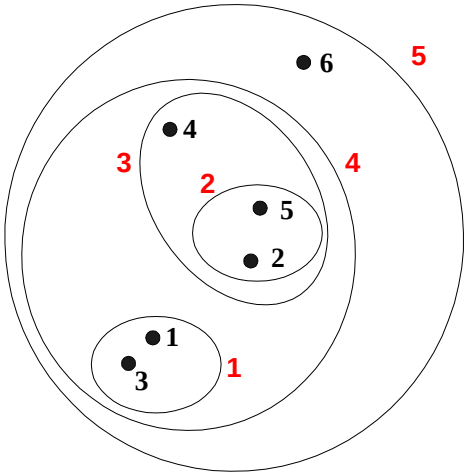


K-means Klasteri

Hijerarhijske tehnike klasterizacije

- Generiše se skup ugnježdenih klastera koji formiraju stablo
- Vizuelizacija dendogramom





Svojstva hijerarhijskih tehnika

- Nije potrebno unaprijed zadati broj klastera
  - Željeni broj klastera dobija se sječenjem dendograma na odgovarajući nivo
- Dendogrami često odgovaraju “prirodnim” hijerarhijama

Vrste hijerarhijskih tehnika

- Dvije osnovne vrste
  - Agglomerative
    - Inicijalno je svaki objekat posebni klaster
    - U svakom koraku spaja se par najbližih klastera sve dok se ne dobije jedan klaster (korijen dendograma)
  - Divisive
    - Inicijalno svi objekti pripadaju jednom klasteru
    - U svakom koraku dijele se klasteri sve dok se ne dobiju singlton klasteri

Agglomerative algoritmi

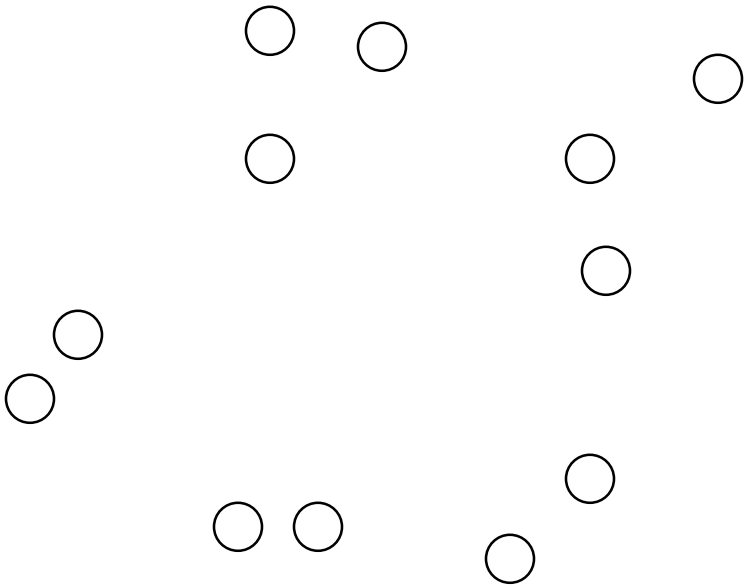
- Češće korišćeni algoritmi
- Osnovni algoritam
  1. Računanje proximity matrice
  2. Svaki objekat je klaster
  3. **Repeat**
  4. spajaju se dva najbliža klastera
  5. Preračunavanje proximity matrice
  6. **Until** postoji samo jedan klaster

Agglomerative algoritmi (2)

- Osnovna operacija je računanje proximity funkcije za dva klastera
  - Različite proximity funkcije određuju različite algoritme

Agglomerative algoritmi: Korak 1

- Svaki objekat je klaster za koje se računa proximity matrica



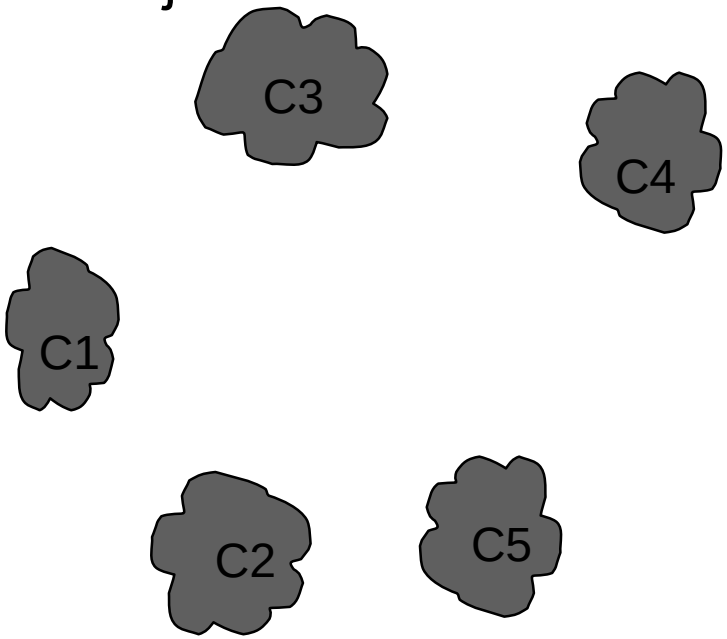
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

Agglomerative algoritmi: Petlja

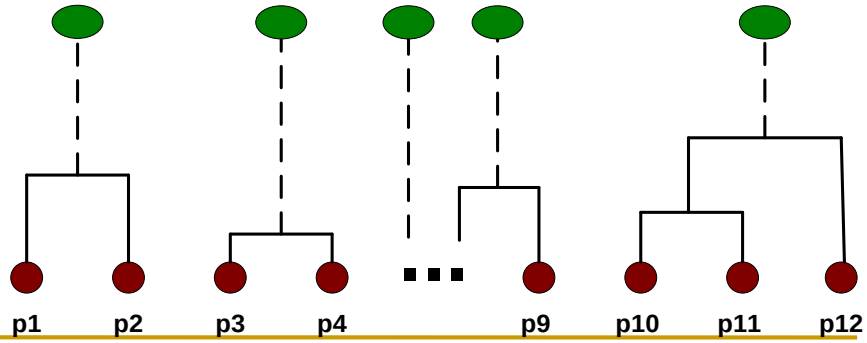
- Spajanjem najbližih klastera

formiraju se veći



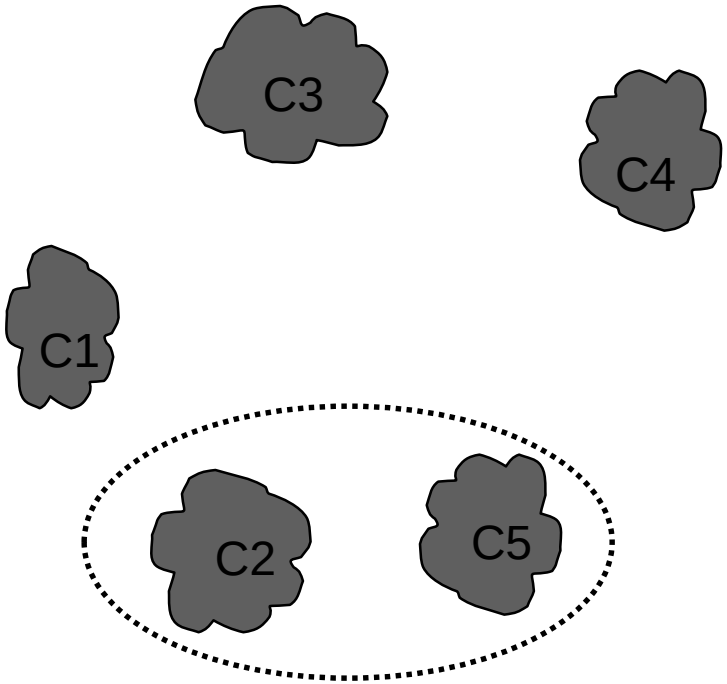
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

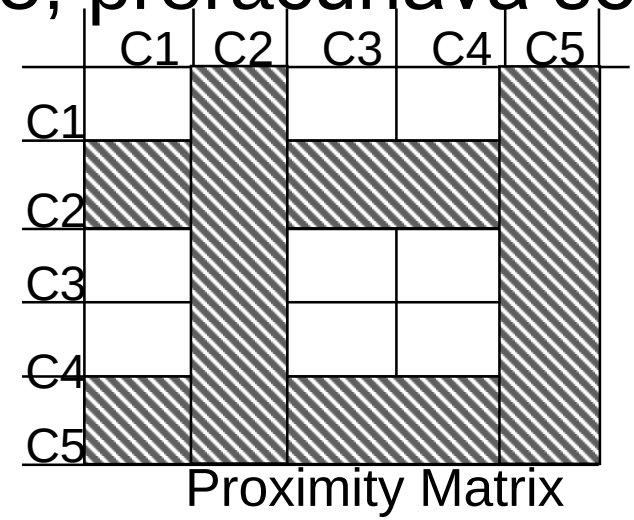
Proximity Matrica



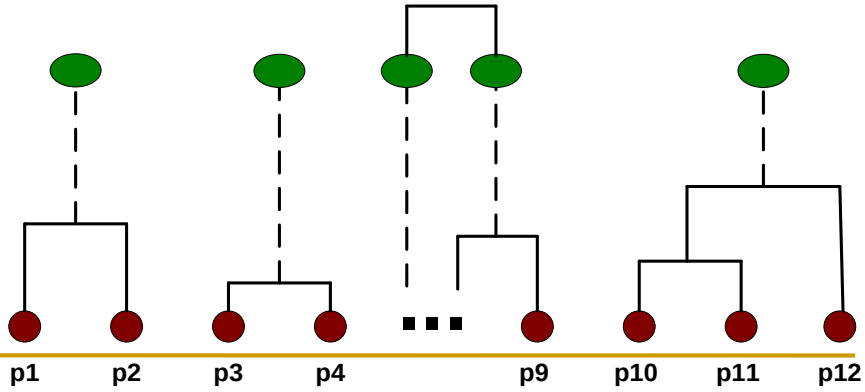
Agglomerative algoritmi: Petlja (2)

- Spajaju se klasteri C2 i C5; preračunava se proximity matrica

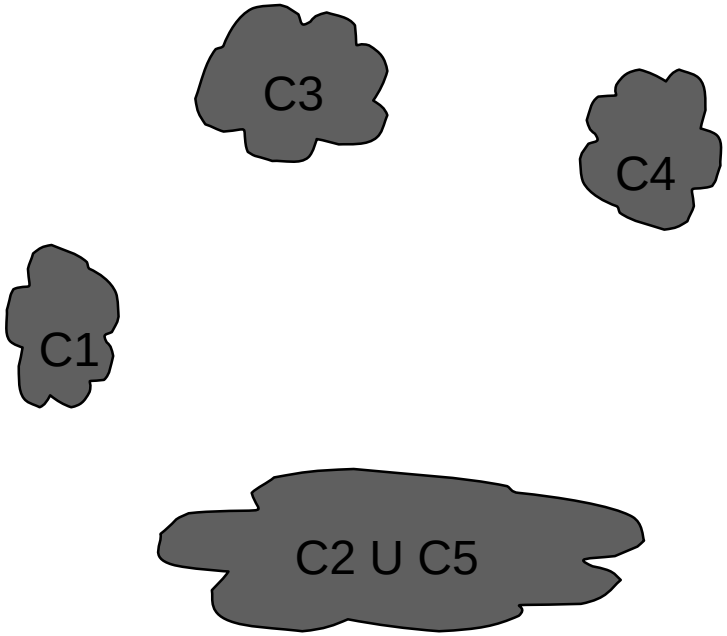




Proximity Matrix



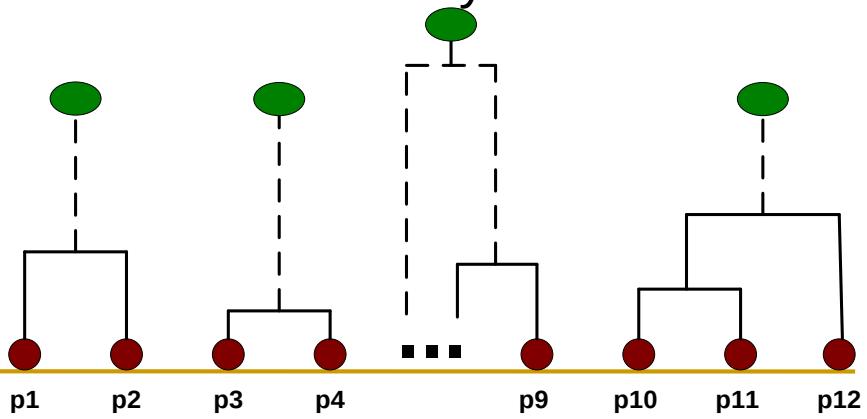
Agglomerative algoritmi: spajanje klastera



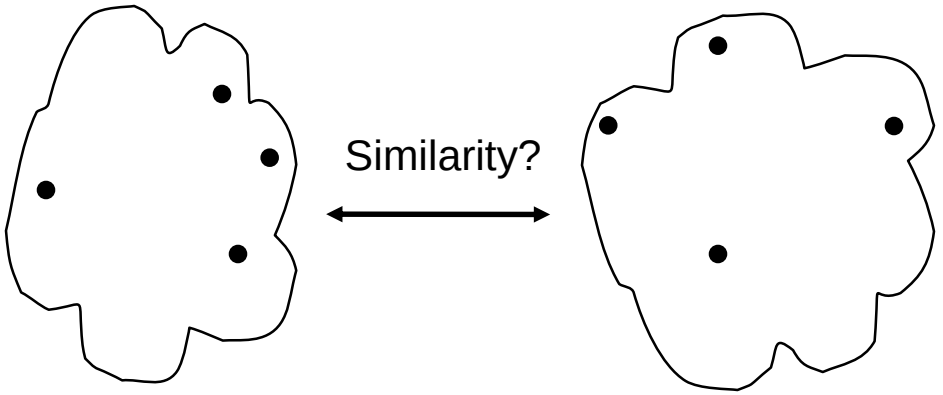
C2 U C5

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

Proximity Matrix



Sličnost između dva klastera

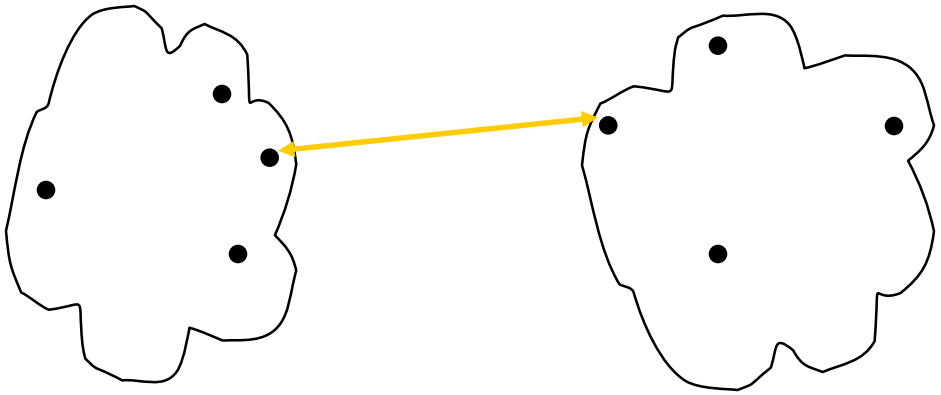


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Prosječno rastojanje svih parova objekata iz dva klastera
- Rastojanje između centroida

Sličnost između dva klastera (2)

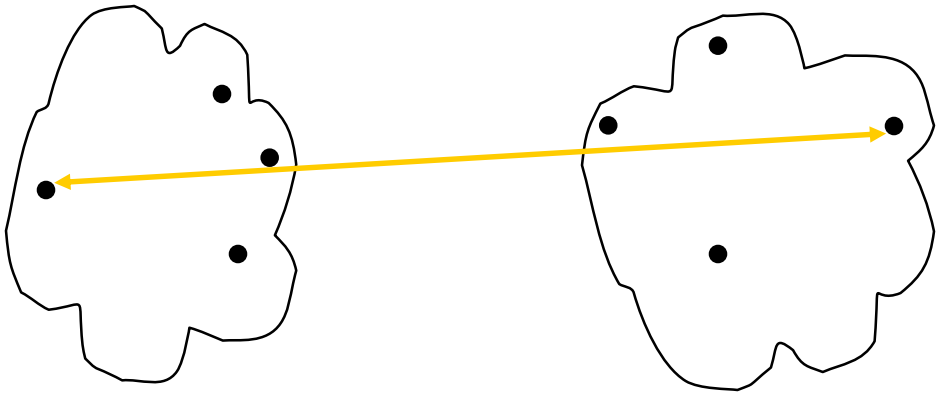


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Prosječno rastojanje svih parova objekata iz dva klastera
- Rastojanje između centroida

Sličnost između dva klastera (3)

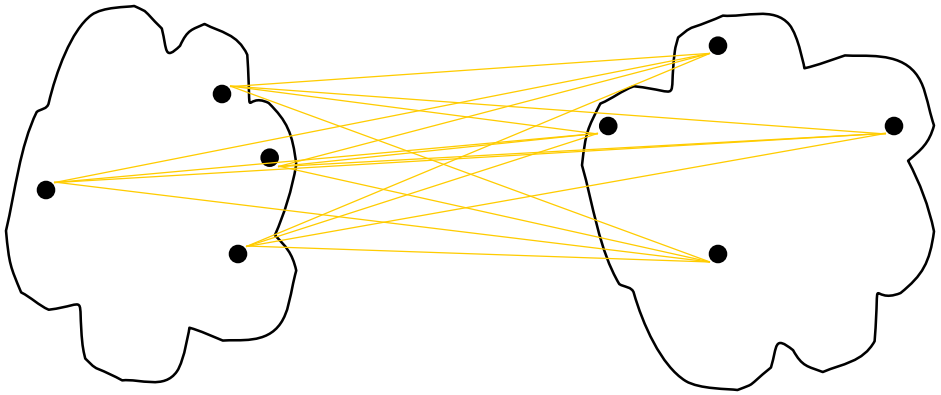


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Prosječno rastojanje svih parova objekata iz dva klastera
- Rastojanje između centroida

Sličnost između dva klastera (4)

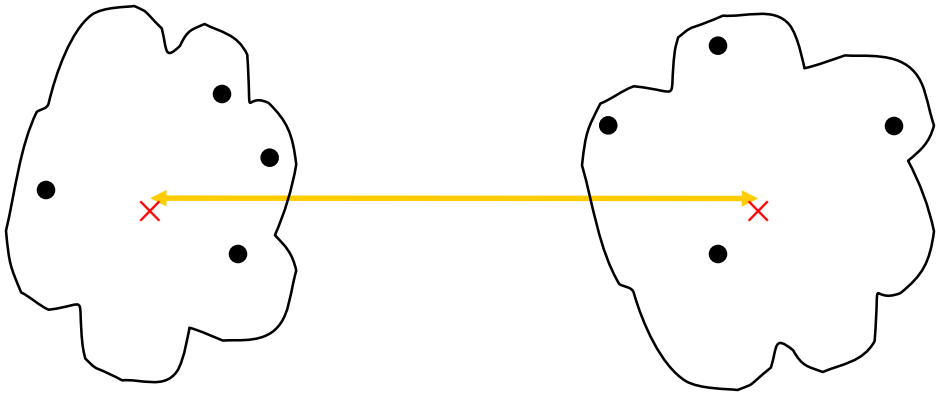


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Prosječno rastojanje svih parova objekata iz dva klastera
- Rastojanje između centroida

Sličnost između dva klastera (5)

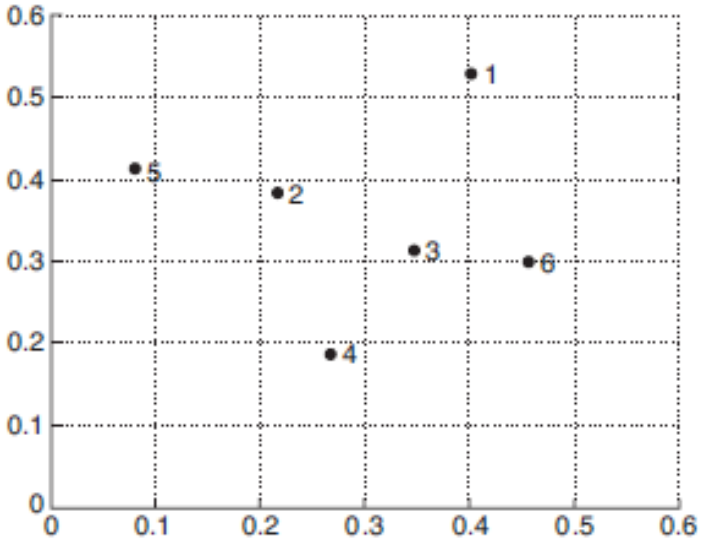


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Prosječno rastojanje svih parova objekata iz dva klastera
- Rastojanje između centroida

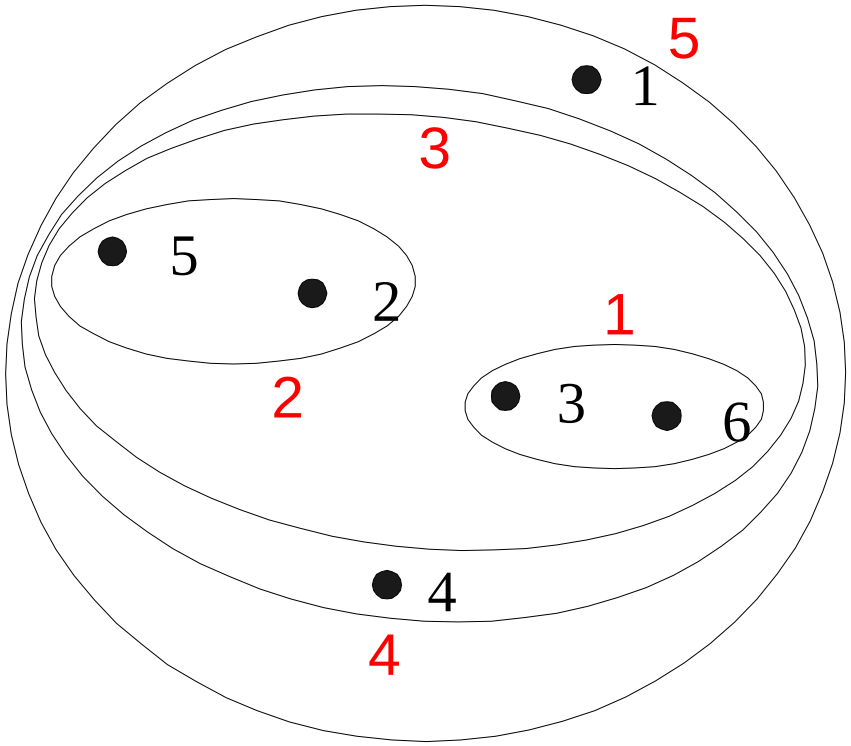
Primjer skupa podataka

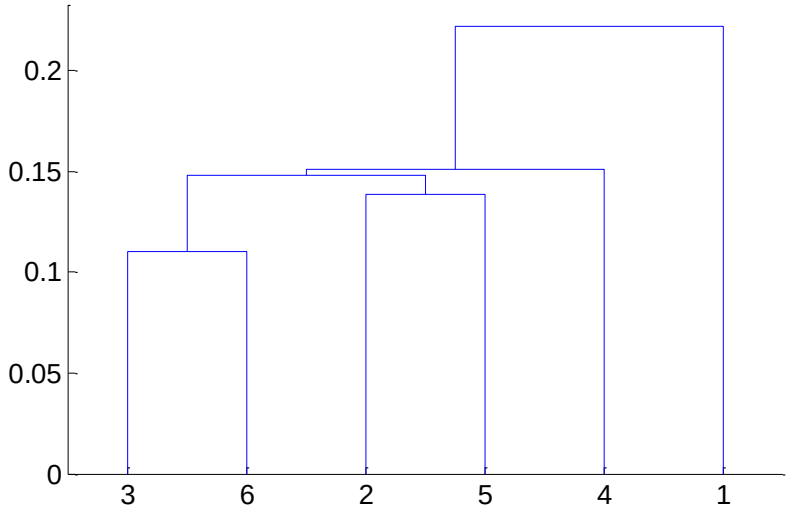


Point	x Coordinate	y Coordinate
p1	0.40	0.53
p2	0.22	0.38
p3	0.35	0.32
p4	0.26	0.19
p5	0.08	0.41
p6	0.45	0.30

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Agglomerative sa MIN





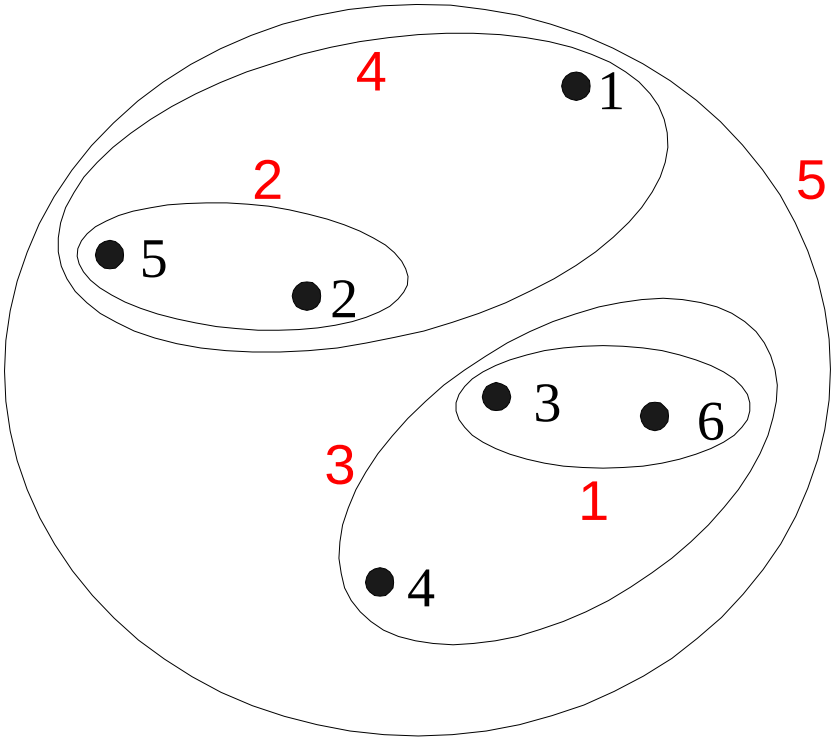
Klasteri

Dendrogram

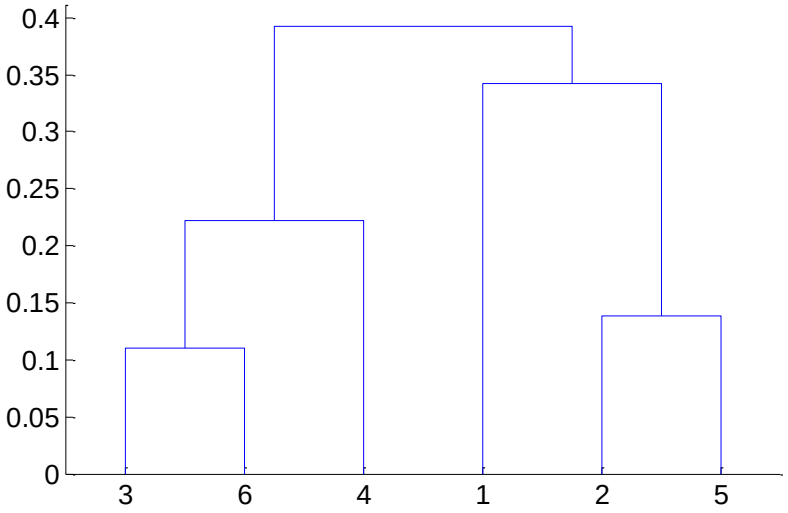
Agglomerative sa MIN (2)

$$
\begin{aligned}
dist(\{3,6\},\{2,5\}) &= \min(dist(3,2), dist(6,2), dist(3,5), dist(6,5)) \\
&= \min(0.15, 0.25, 0.28, 0.39) \\
&= 0.15.
\end{aligned}
$$

Agglomerative sa MAX



Klasteri



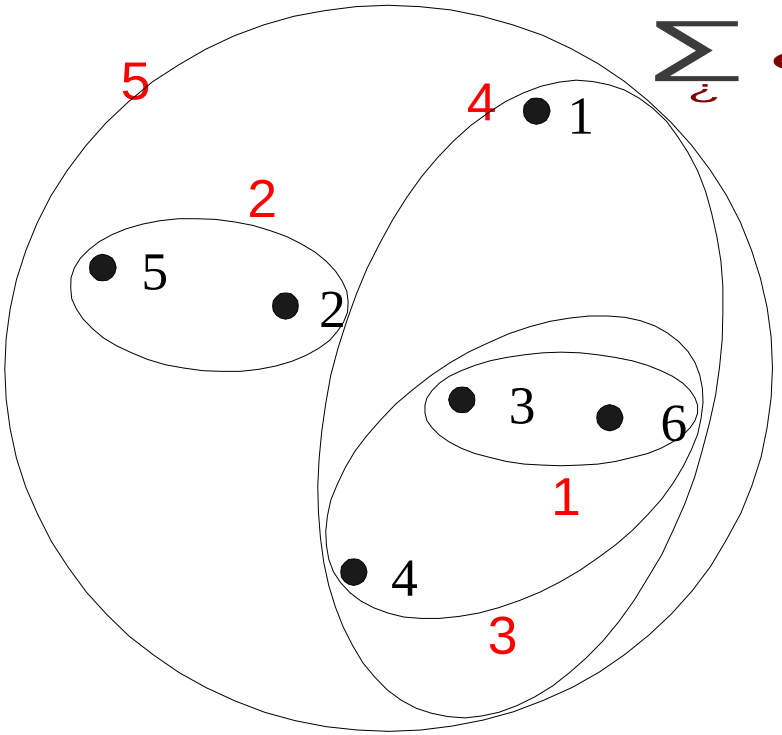
Dendrogram

Agglomerative sa MAX (2)

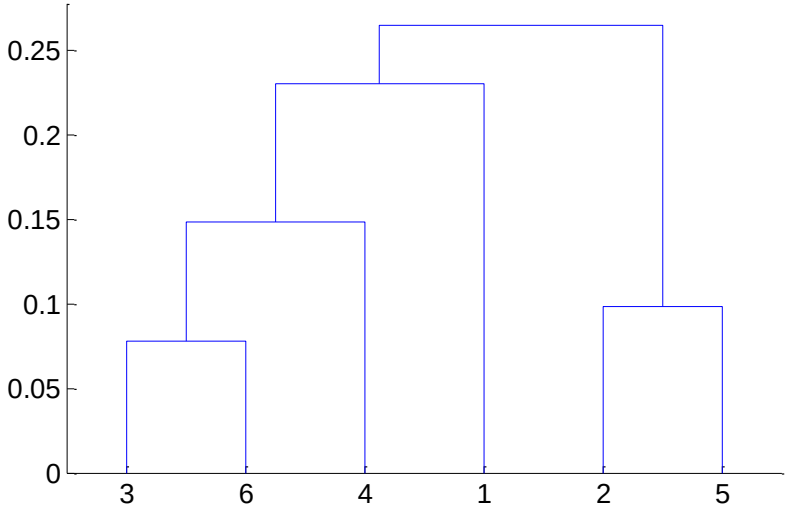
$$
\begin{aligned}
dist(\{3,6\},\{4\}) &= \max(dist(3,4), dist(6,4)) \\
&= \max(0.15, 0.22) \\
&= 0.22. \\
dist(\{3,6\},\{2,5\}) &= \max(dist(3,2), dist(6,2), dist(3,5), dist(6,5)) \\
&= \max(0.15, 0.25, 0.28, 0.39) \\
&= 0.39. \\
dist(\{3,6\},\{1\}) &= \max(dist(3,1), dist(6,1)) \\
&= \max(0.22, 0.23) \\
&= 0.23.
\end{aligned}
$$

Agglomerative sa Prosječnim rastojanjem

$$\sum_{p_j \in Cluster_j} \sum_{p_i \in Cluster_i} \frac{¿}{|Cluster_i| * |Cluster_j|}$$



Klasteri



Dendrogram

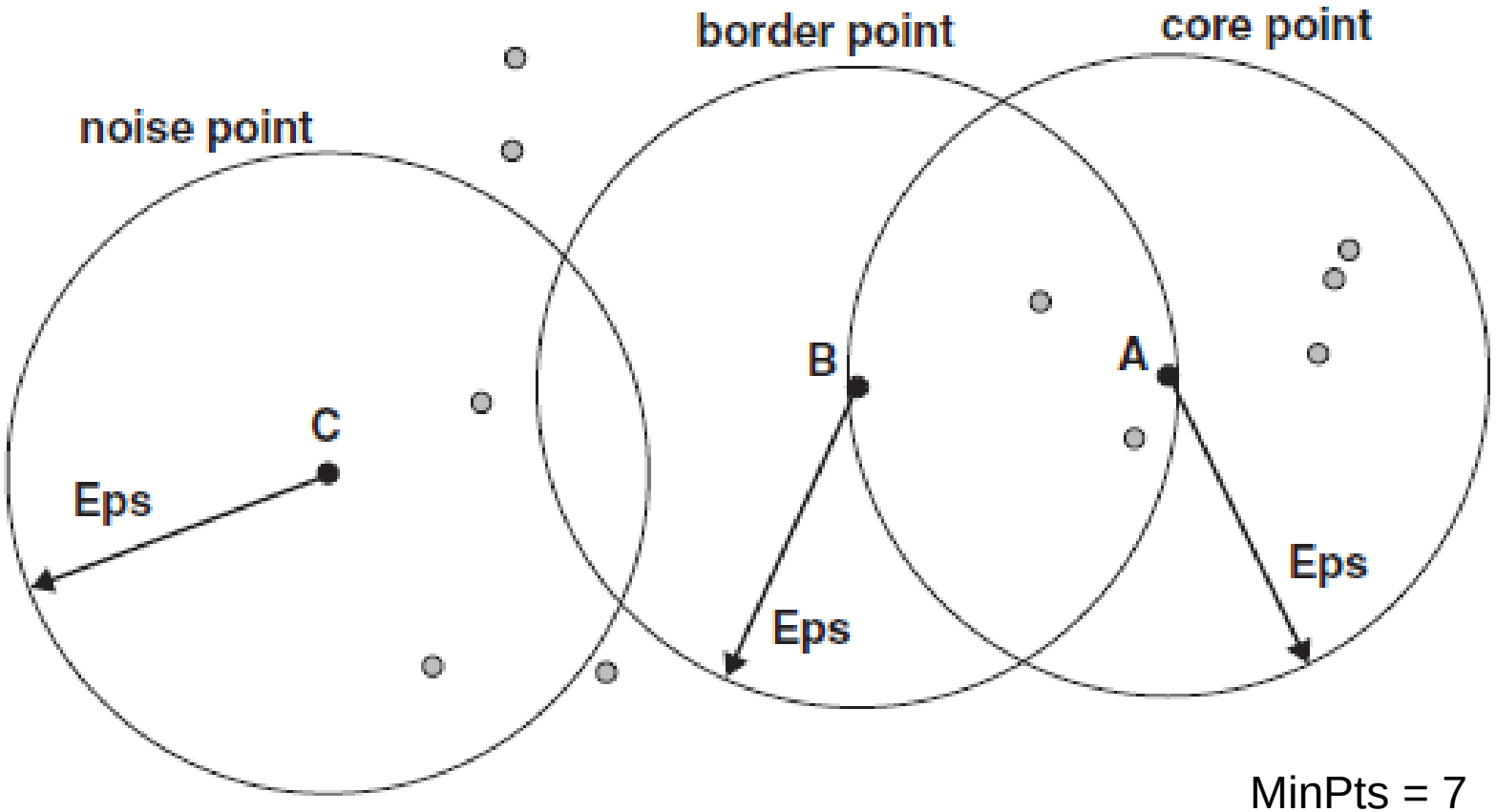
Agglomerative sa Prosječnim rastojanjem (2)

$$
\begin{aligned}
dist(\{3,6,4\},\{1\}) &= (0.22+0.37+0.23)/(3*1) \\
&= 0.28 \\
dist(\{2,5\},\{1\}) &= (0.2357+0.3421)/(2*1) \\
&= 0.2889 \\
dist(\{3,6,4\},\{2,5\}) &= (0.15+0.28+0.25+0.39+0.20+0.29)/(6*2) \\
&= 0.26
\end{aligned}
$$

DBSCAN

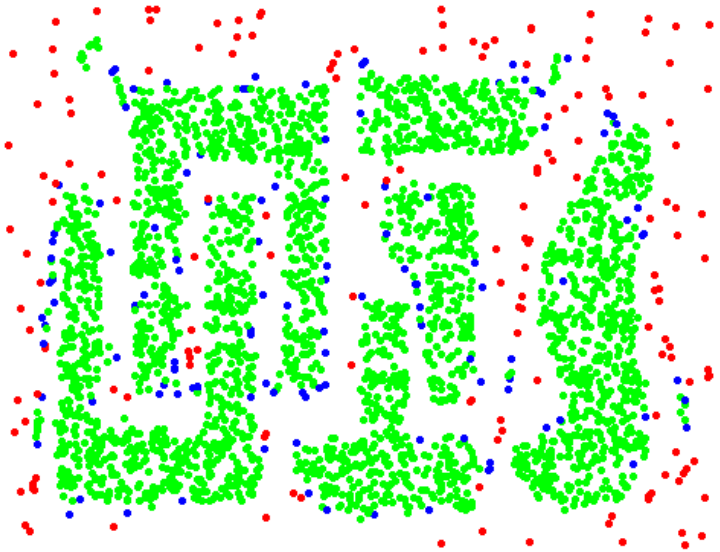
- Density-based algoritam
- Gustina je broj objekata unutar kruga poluprečnika Eps
- Objekti se klasifikuju na
  - Core point: ako u Eps okolini sadrži više od MinPts objekata
  - Border point: u Eps okolini nekog objekta koji je core point
  - Noise: ostali objekti

Core, Border i Noise point



Core, Border i Noise point (2)

Originalni skup tačaka



core, border, noise

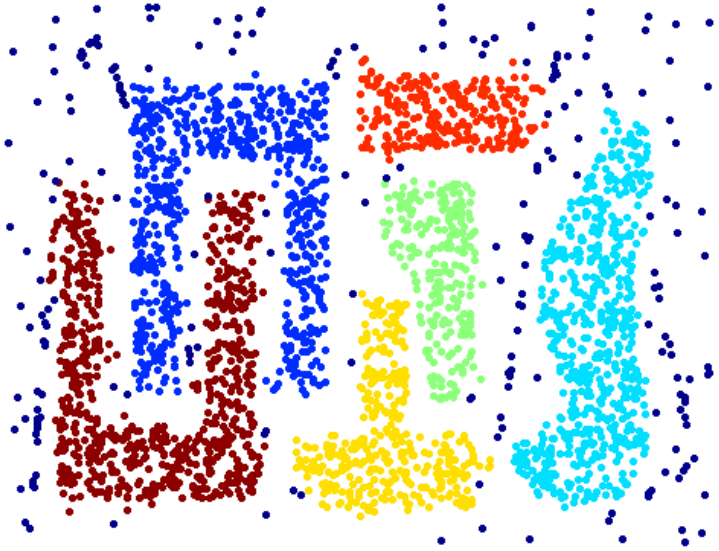
DBSCAN algoritam

Algorithm 8.4 DBSCAN algorithm.

1: Label all points as core, border, or noise points.
2: Eliminate noise points.
3: Put an edge between all core points that are within *Eps* of each other.
4: Make each group of connected core points into a separate cluster.
5: Assign each border point to one of the clusters of its associated core points.

DBSCAN algoritam (2)

Originalni skup tačaka



Klasteri

DBSCAN: određivanje MinPts i Eps

- k-dist je rastojanje nekog objekta do njegovog k-tog najbližeg susjeda
  - Za objekte koji pripadaju nekom klasteru k-dist nije veliko ako je k manje od veličine klastera
  - Za objekte koji su šum k-dist je relativno veliko

Sortirati k-dist za sve objekte i napraviti grafikon