Odkrivanje skupin

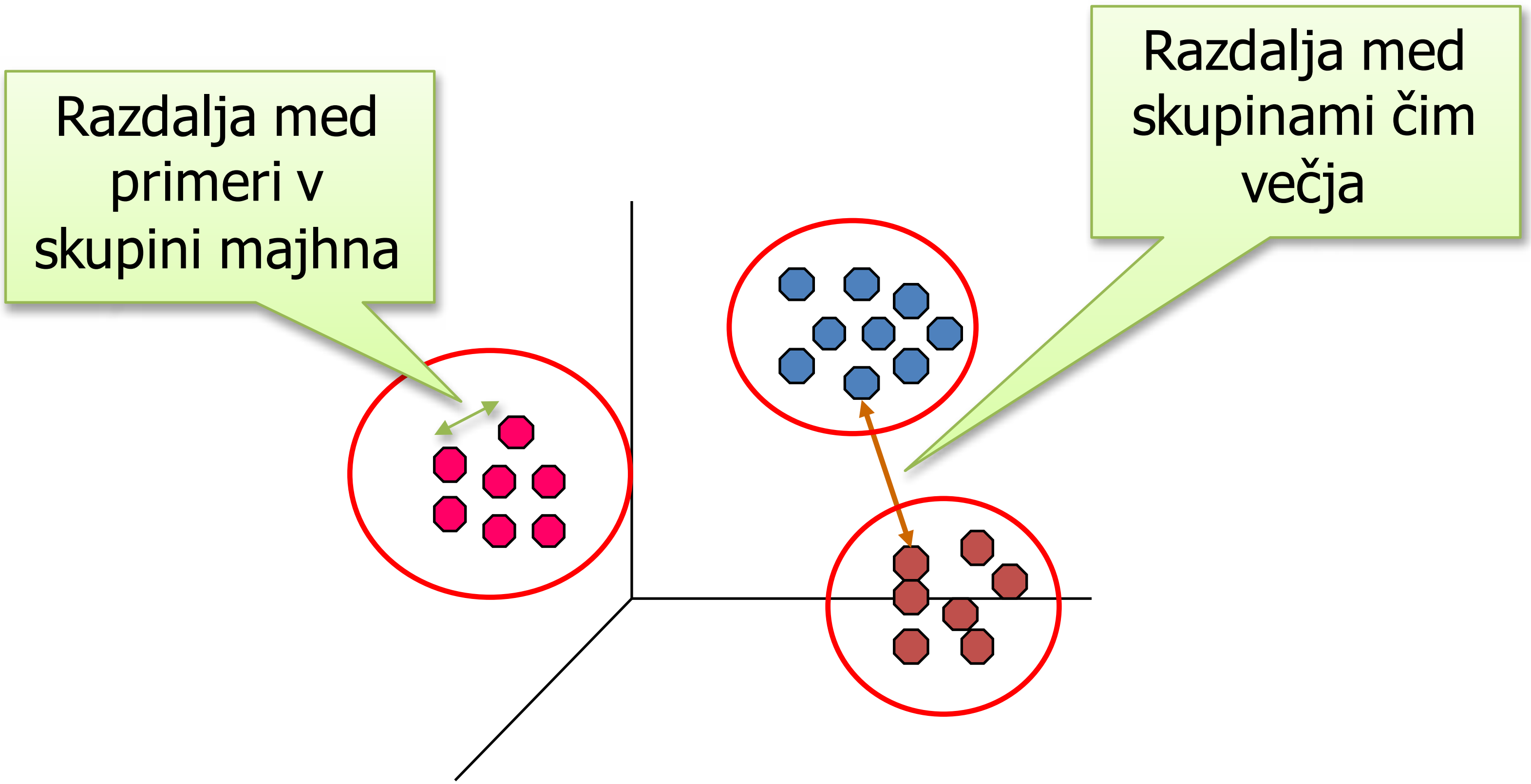
Tomaž Curk

metoda voditeljev (k-means)
hierarhično razvrščanje (hierarchical clustering)

Tan, Steinbach & Kumar: Introduction to Data Mining
http://www-users.cs.umn.edu/~kumar/dmbook/index.php

Odkrivanje skupin

Poišči skupine primerov tako, da bo vsaka skupina vsebovala le podobne objekte.



Koliko skupin?

Koliko skupin?

Koliko skupin?

Koliko skupin?

Metode za razvrščanje v skupine

- Delitev primerov (Partitional clustering)
  - Primere delimo v neprekrivajoče skupine (clusters)
  - Primer: metoda voditeljev (k-means clustering)
- Hierarhično razvrščanje (Hierarchical clustering)
  - Odkrijemo hierarhijo skupin, kar prikažemo z hierarhičnim drevesom (dendrogramom)

Metoda k-voditeljev (k-means)

- Vsaka skupina ima svojega voditelja (centroid).
- Vsak primer pripišemo najbližjemu voditelju.
- Podati moramo K - število voditeljev (skupin).
- Preprost algoritem.

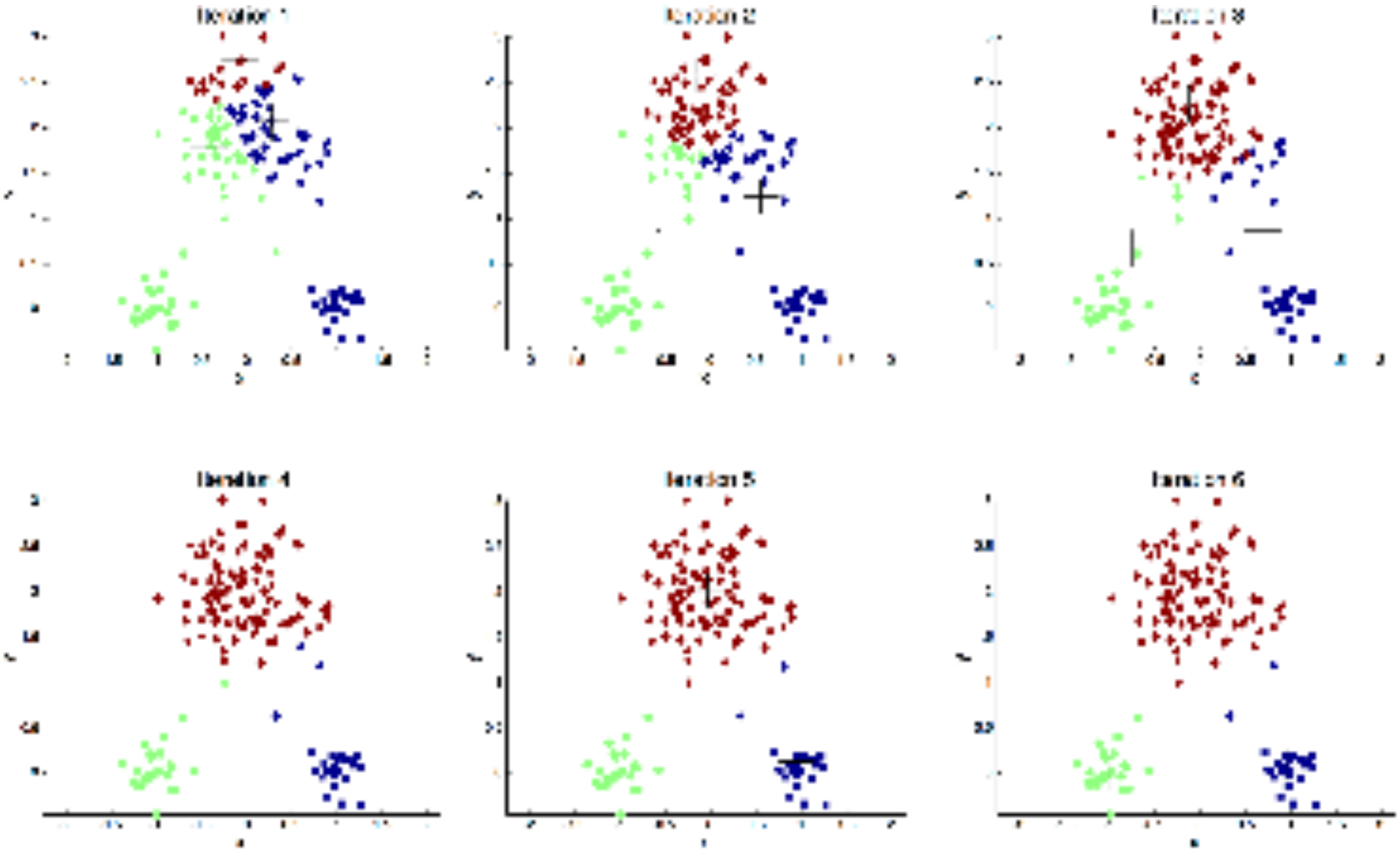
Algorithm 1 Basic K-means Algorithm.

Select K points as the initial centroids.
repeat
    Form K clusters by assigning all points to the closest centroid.
    Recompute the centroid of each cluster.
until The centroids don't change

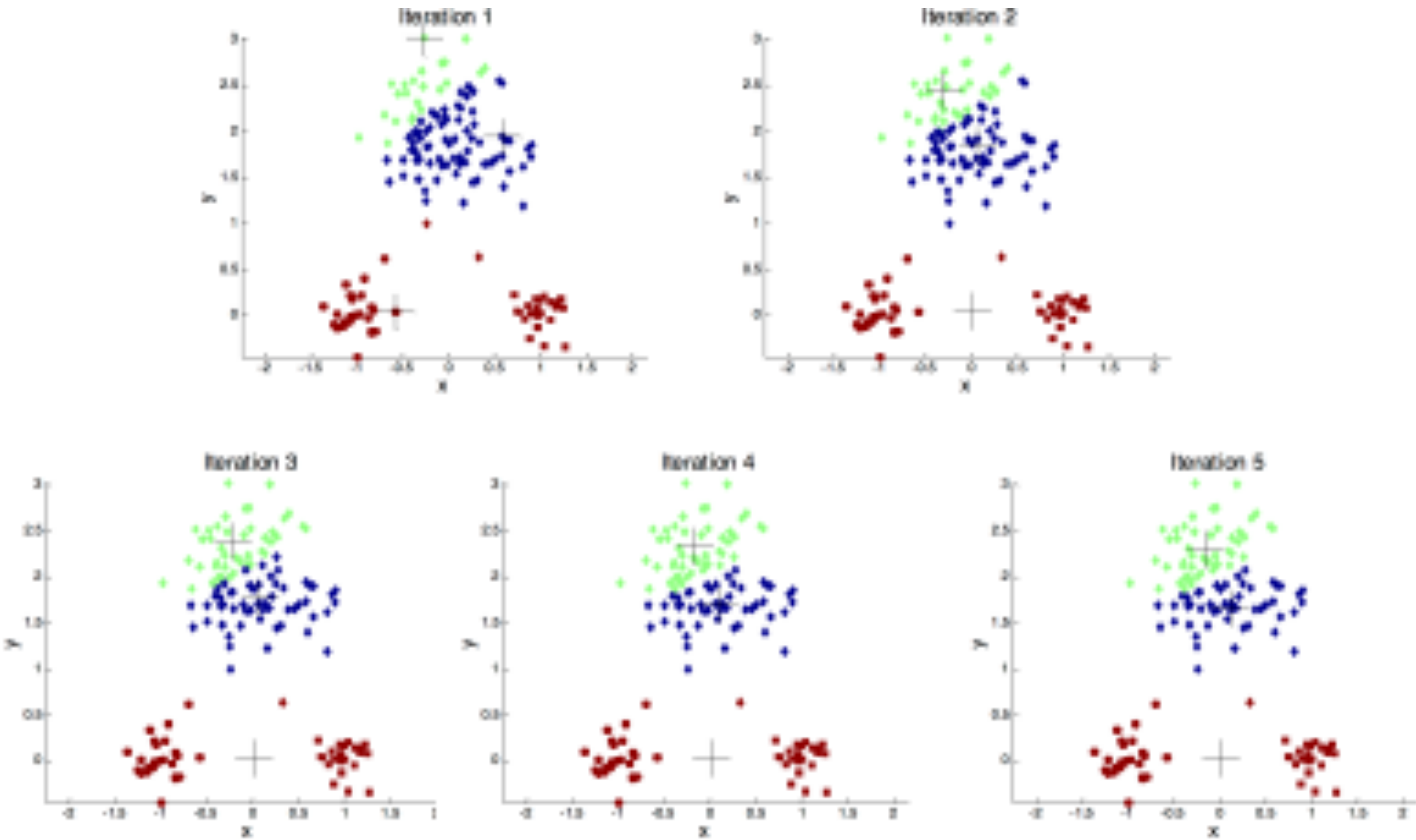
Metoda k-voditeljev (k-means)

- Začetni voditelji izbrani naključno.
  – metoda je občutljiva na to izbiro
- Bližino merimo z Evklidsko razdaljo, kosinusno razdaljo, korelacijo, itd.
- Algoritem konvergira za zgornje mere.
  – Konvergira hitro, le v nekaj korakih.
  – Ustavitveni pogoj je navadno: “Until relatively few points change clusters”
- Potrebno predprocesiranje podatkov (normalizacija).

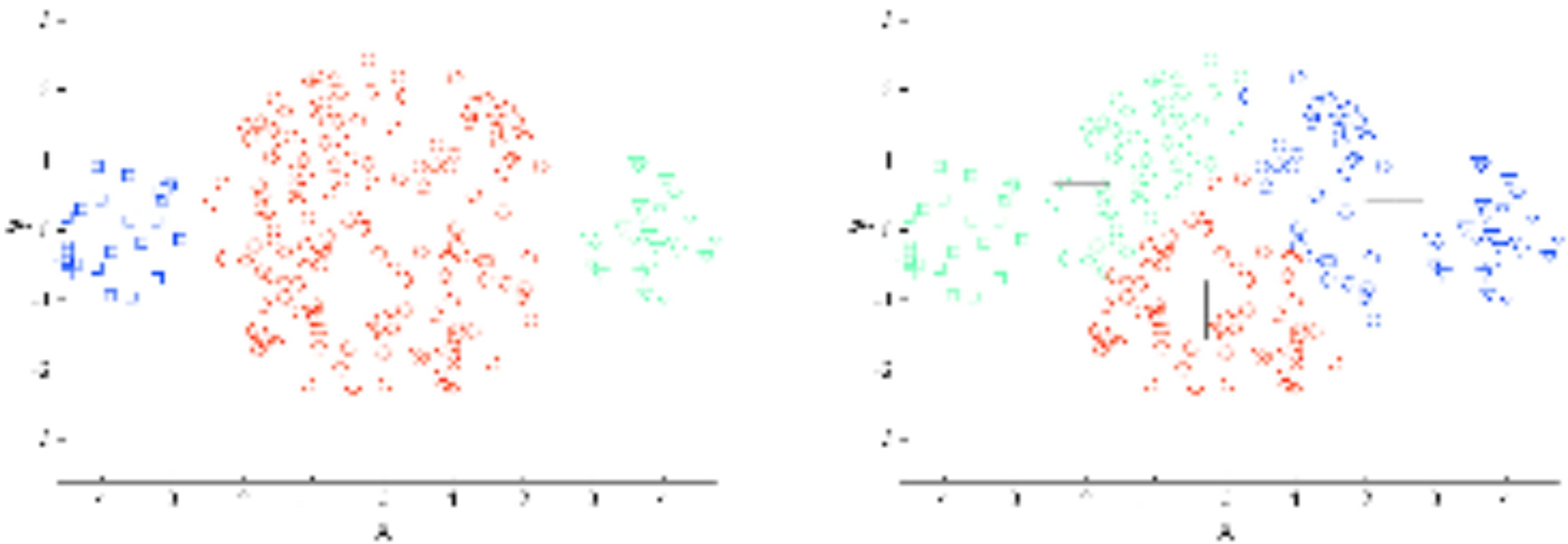
Metoda k-voditeljev (k-means)



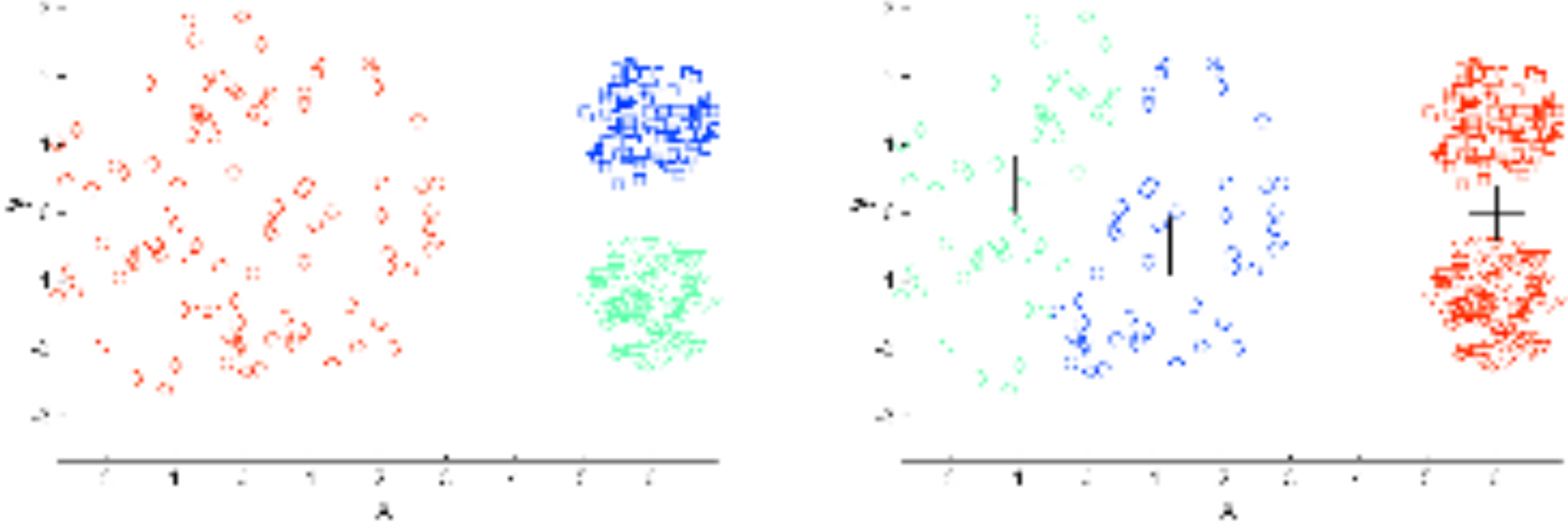
Izbira začetnih voditeljev (centroidov)



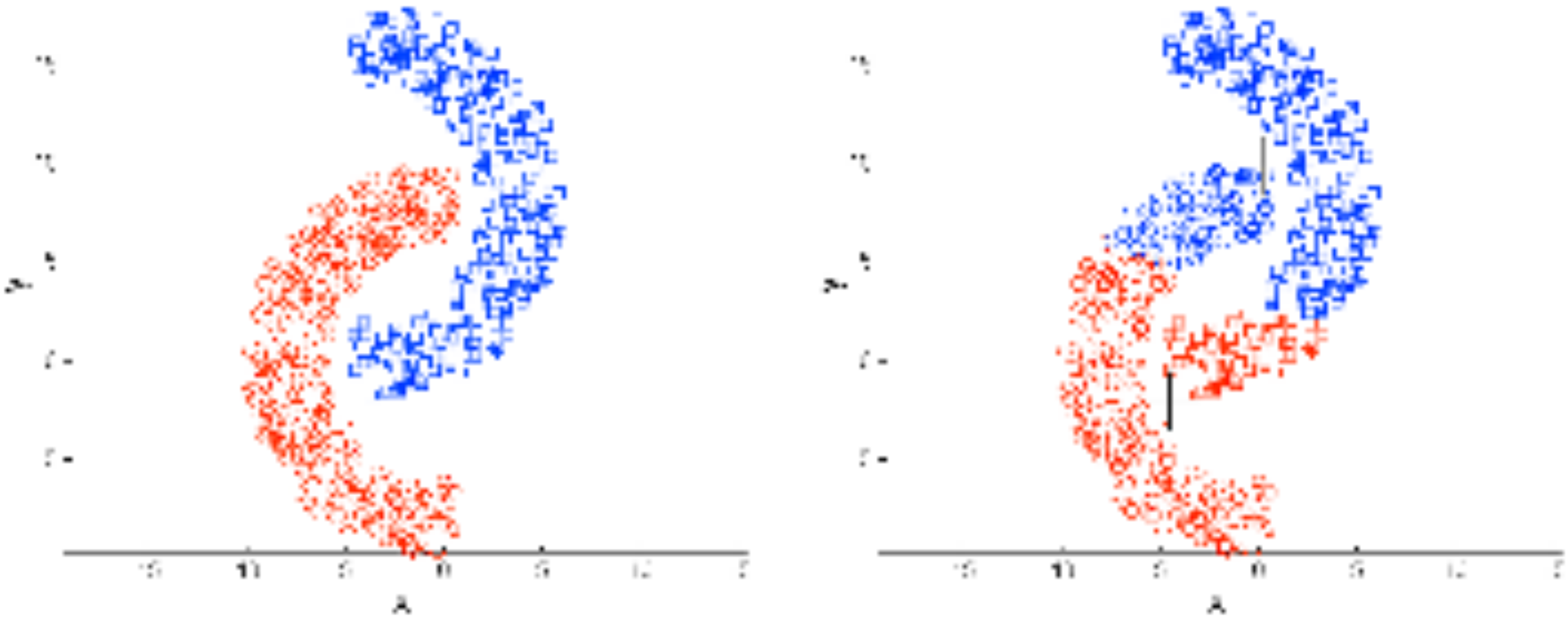
Omejitve: različne velikosti skupin



Omejitve: različne gostote skupin

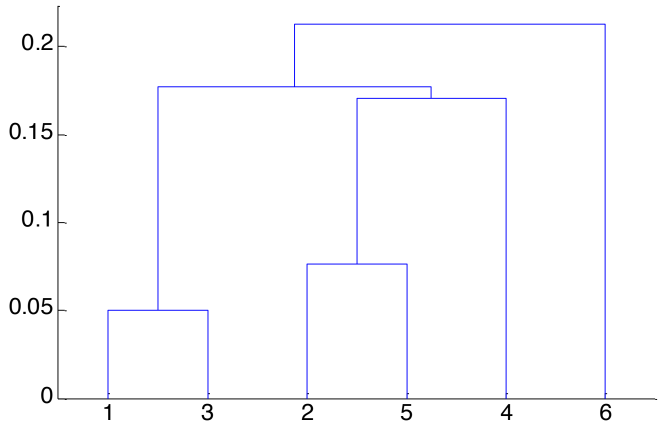


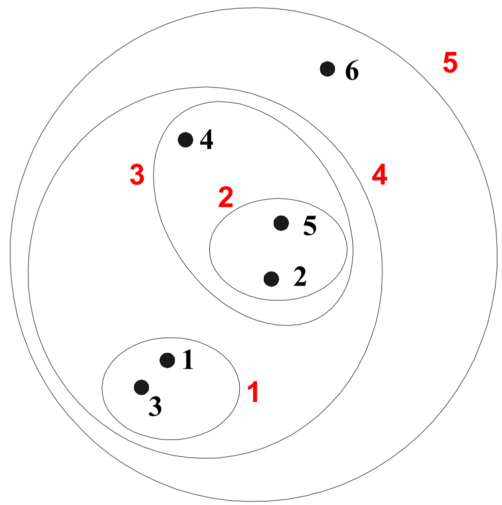
Omejitve: različne oblike



Hierarhično razvrščanje

- Rezultat je hierarhija skupin primerov.
- Prikažemo z dendrogramom
  - Iz drevesa je razvidno, kako so se združevale skupine.





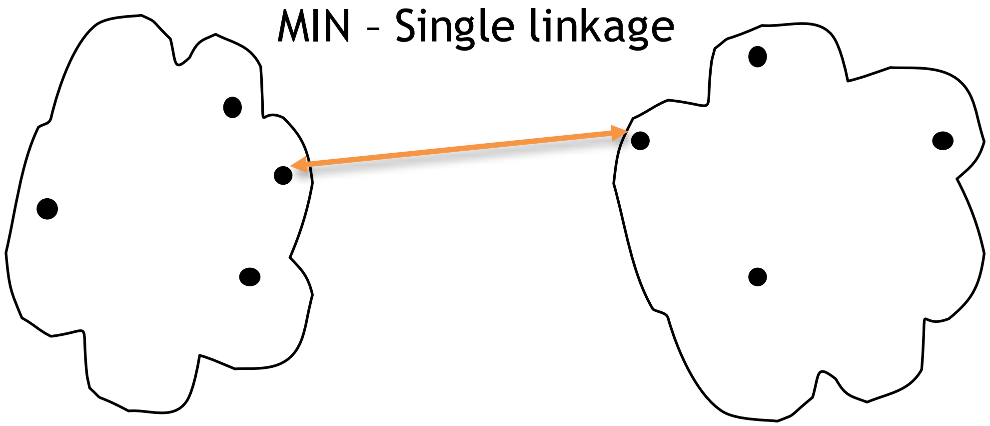
Prednosti

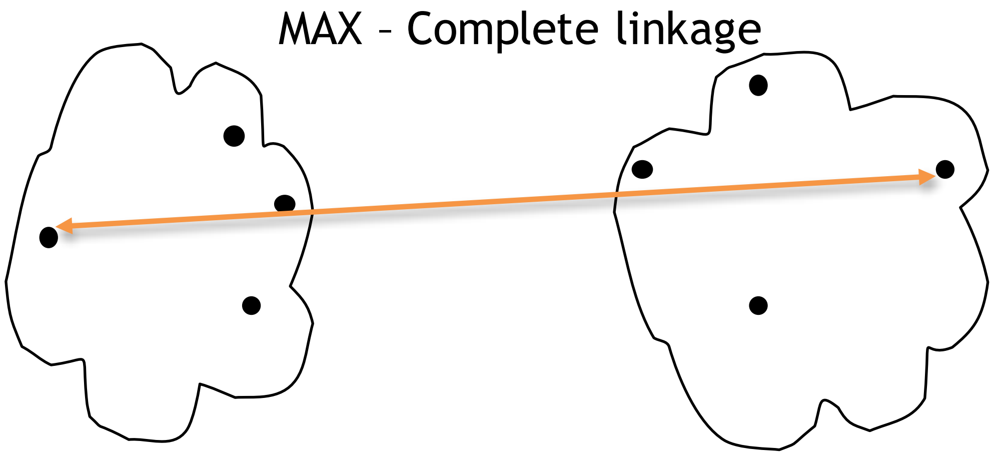
- Števila skupin ni potrebno podati.
  - Z “rezanjem” drevesa lahko pridobimo razbitje na poljubno število skupin.
- Skupine navadno sovpadajo s taksonomijami (razredom)
  - N.pr., rekonstrukcija filogenetskih dreves (zoo)

Razvrščanje v skupine na podlagi združevanja primerov (Agglomerative Clustering)

1. Izračunaj matriko podobnosti
2. Vsak primer svoja skupina
3. **Repeat**
   - Združi dve najbolj podobni skupini
   - Posodobi matriko podobnosti
4. **Until** ena skupina

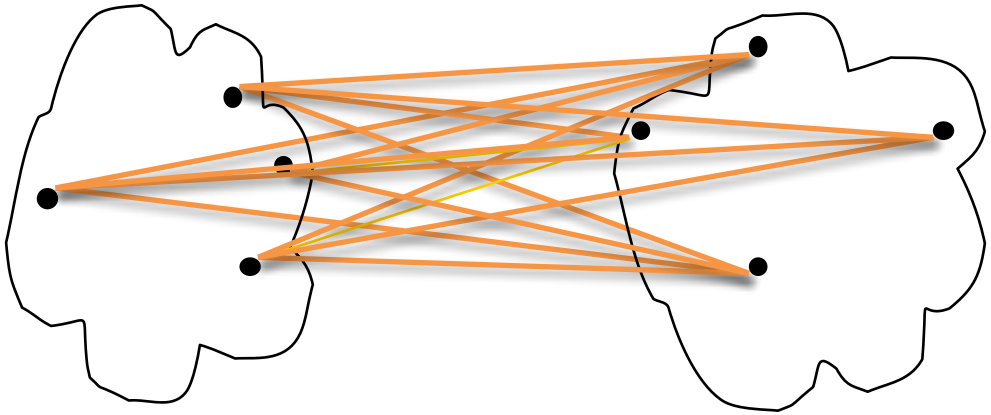
Podobnost med skupinami (Inter-Cluster Similarity)





Podobnost med skupinami (Inter-Cluster Similarity)

Group average – average linkage



Primer hierarhičnega razvrščanja

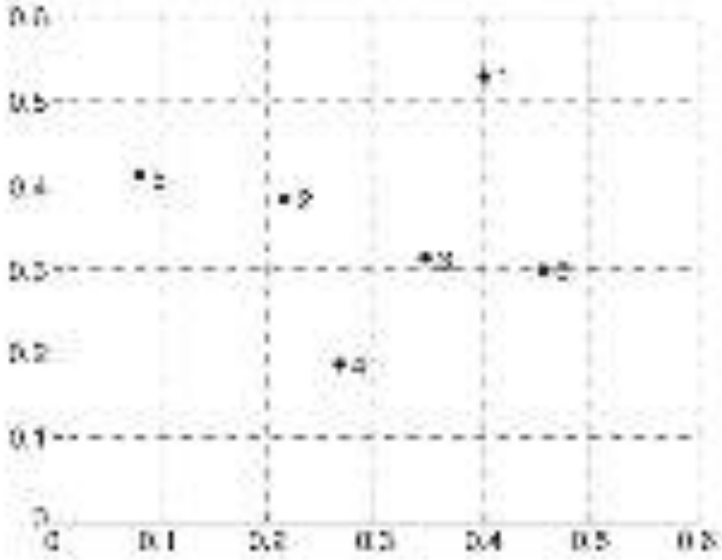


Figure 8.15. Set of 6 two-dimensional points.

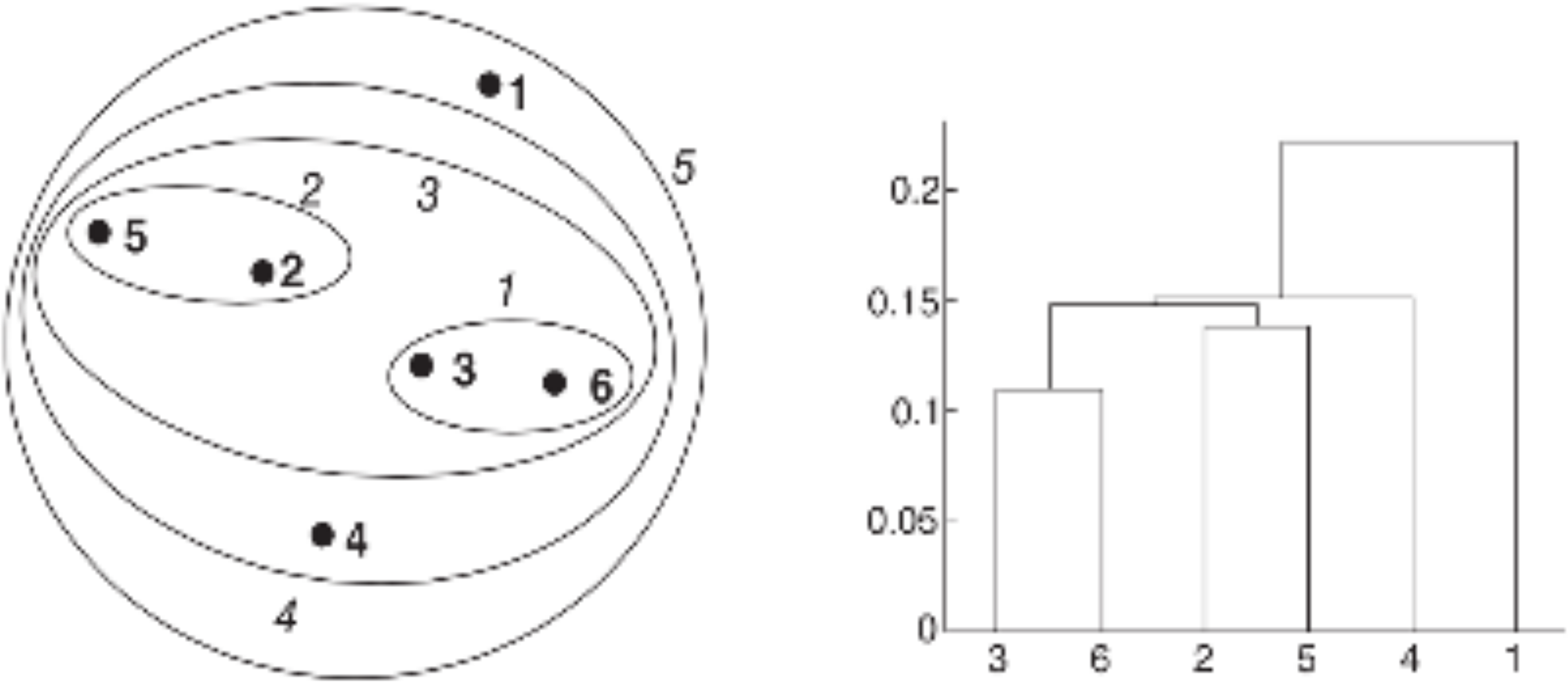
Point	x Coordinate	y Coordinate
p1	0.40	0.53
p2	0.22	0.38
p3	0.35	0.32
p4	0.26	0.19
p5	0.08	0.41
p6	0.45	0.30

Table 8.3. xy coordinates of 6 points.

	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

Table 8.4. Euclidean distance matrix for 6 points.

Single linkage

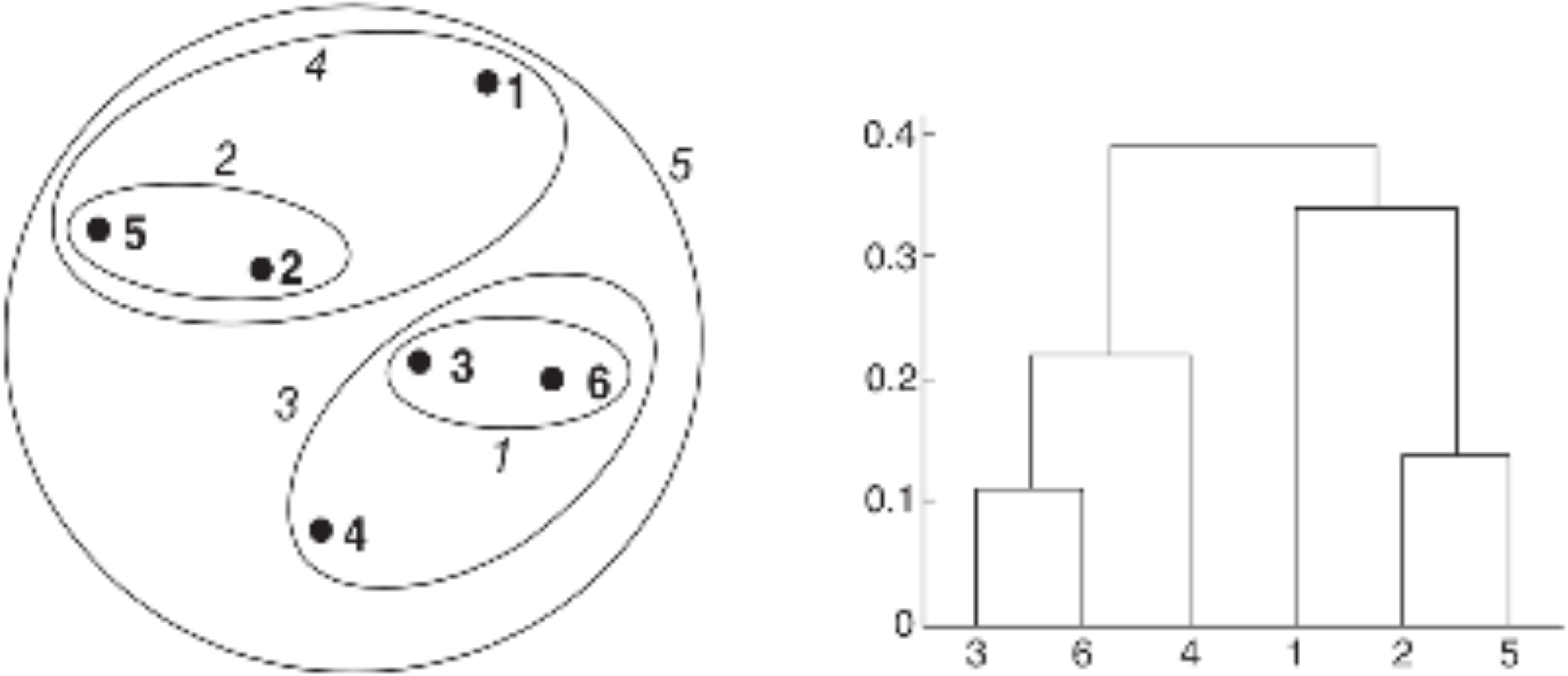


(a) Single link clustering.

(b) Single link dendrogram.

Figure 8.16. Single link clustering of the six points shown in Figure 8.15.

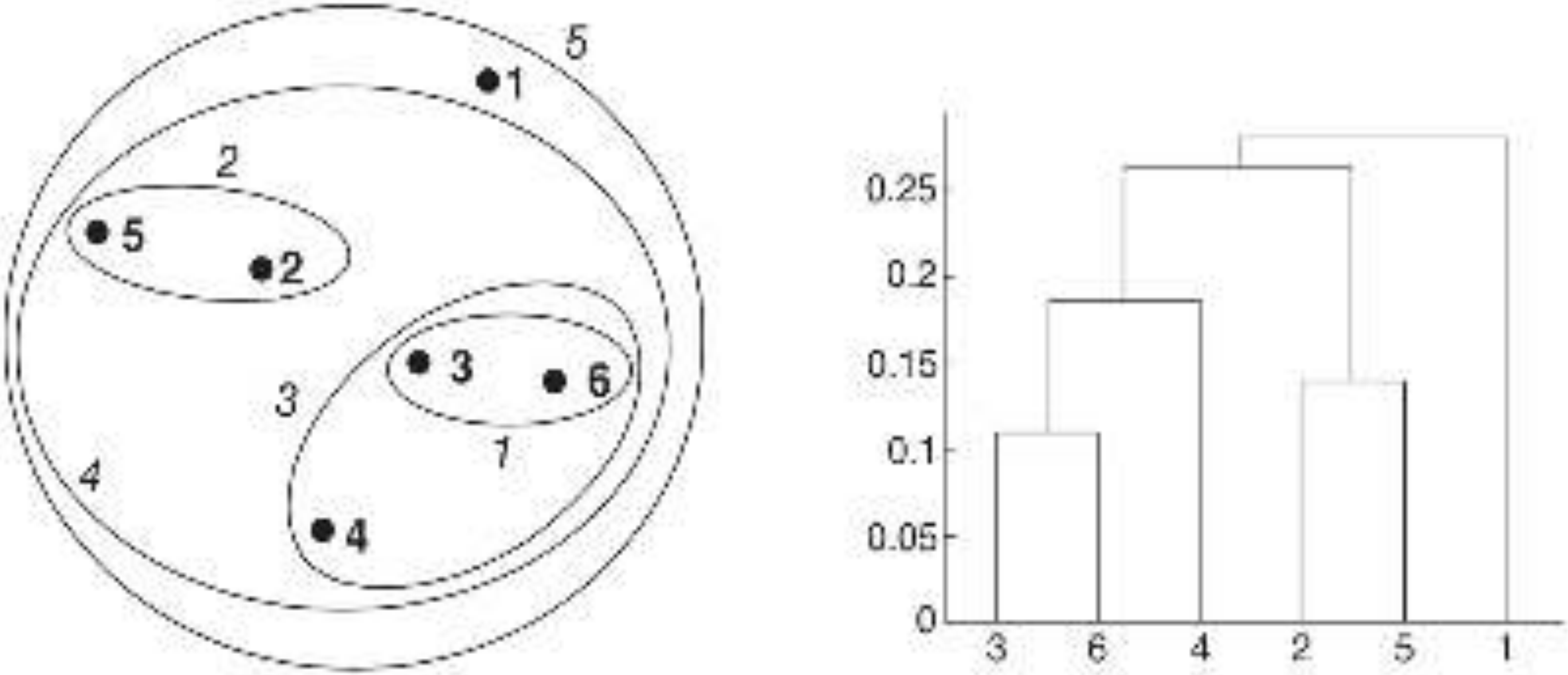
Complete linkage



(a) Complete link clustering. (b) Complete link dendrogram.

Figure 8.17. Complete link clustering of the six points shown in Figure 8.15.

Average linkage

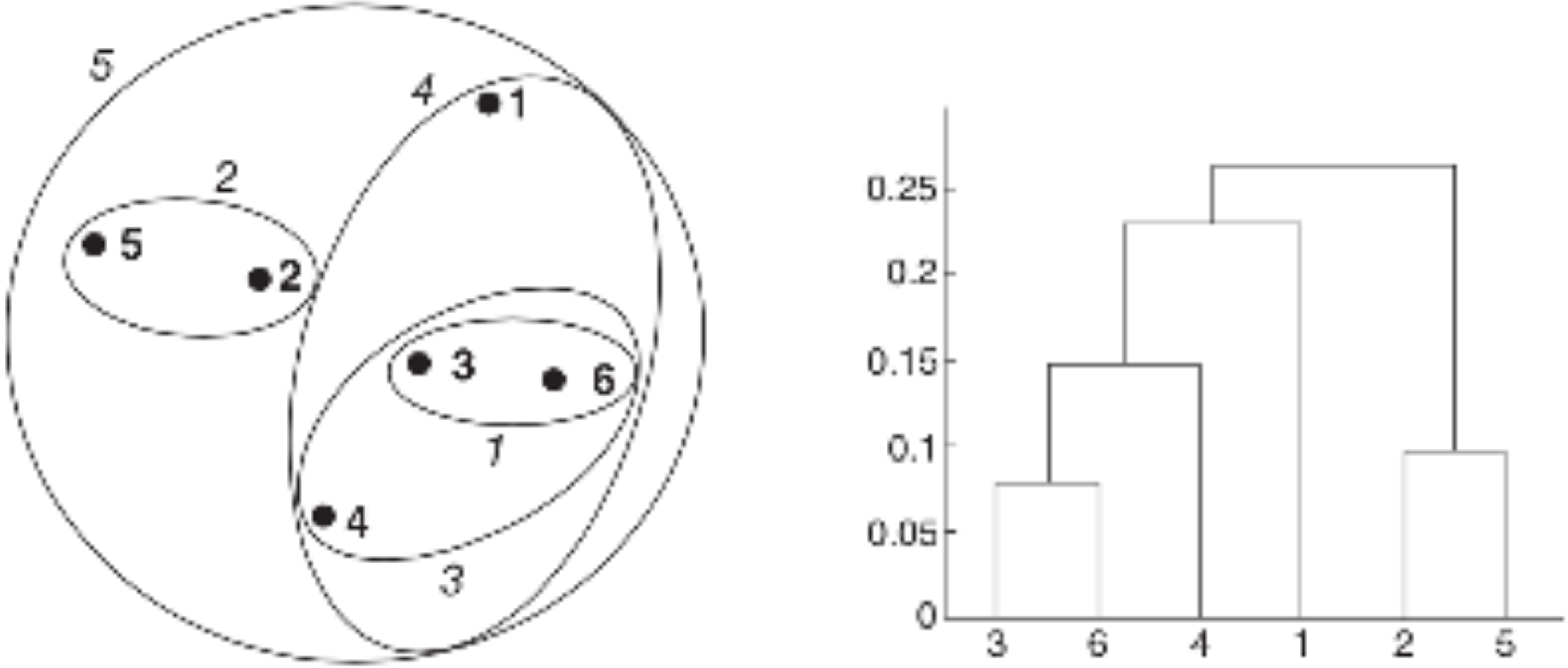


(a) Group average clustering.

(b) Group average dendrogram.

Figure 8.18. Group average clustering of the six points shown in Figure 8.15.

Wardovo združevanje



(a) Ward's clustering. (b) Ward's dendrogram.

Figure 8.19. Ward's clustering of the six points shown in Figure 8.15.

Wardovo združevanje

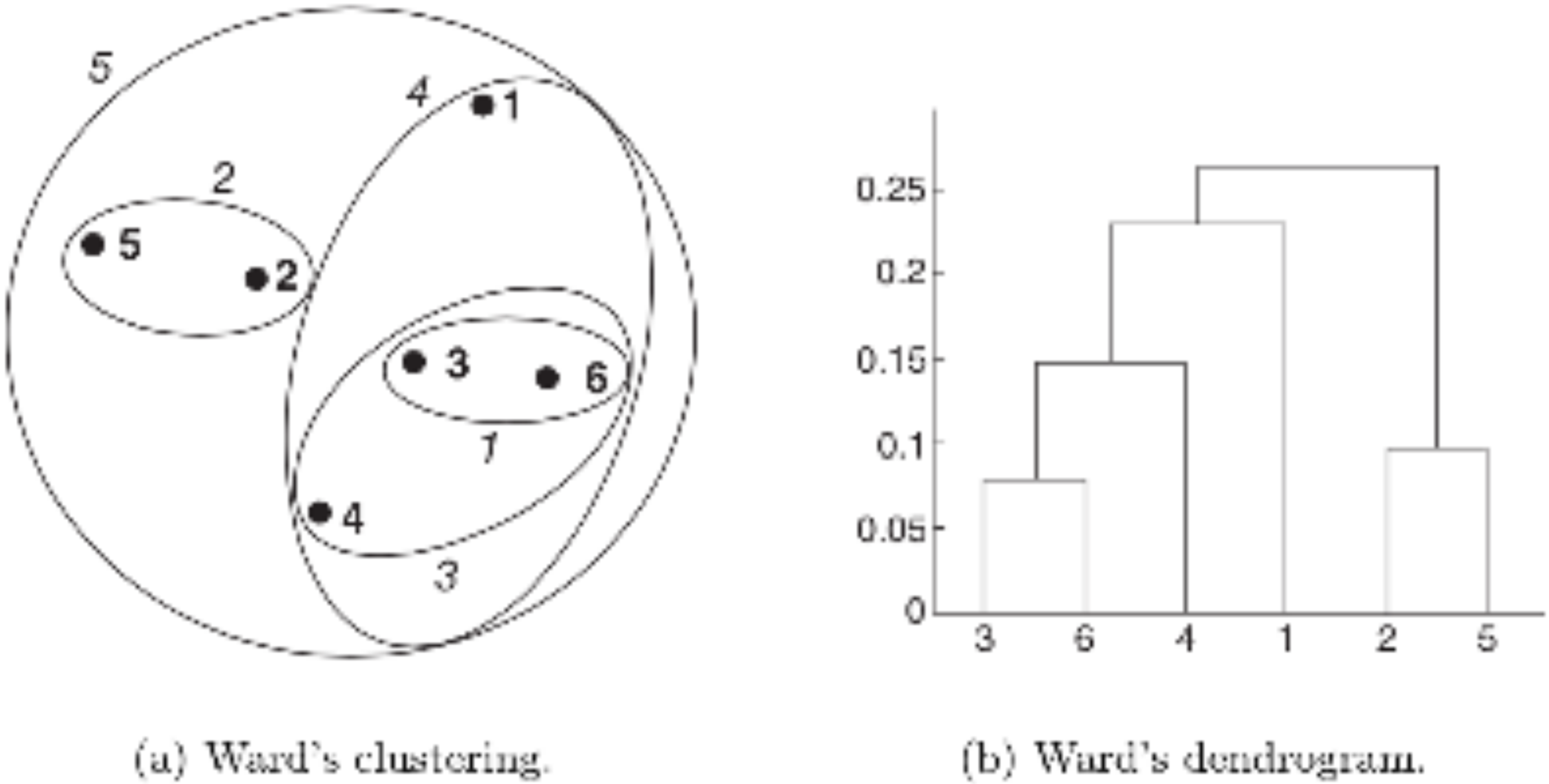
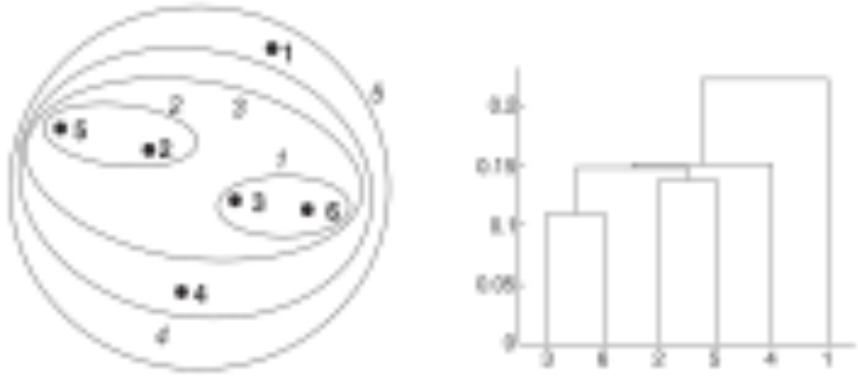


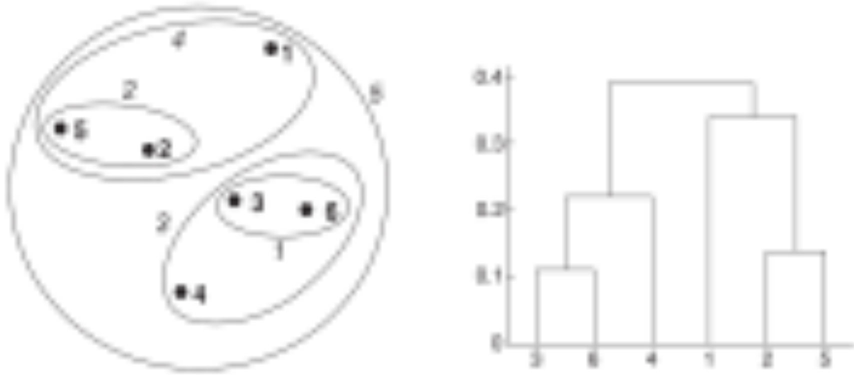
Figure 8.19. Ward's clustering of the six points shown in Figure 8.15.

Različne skupine



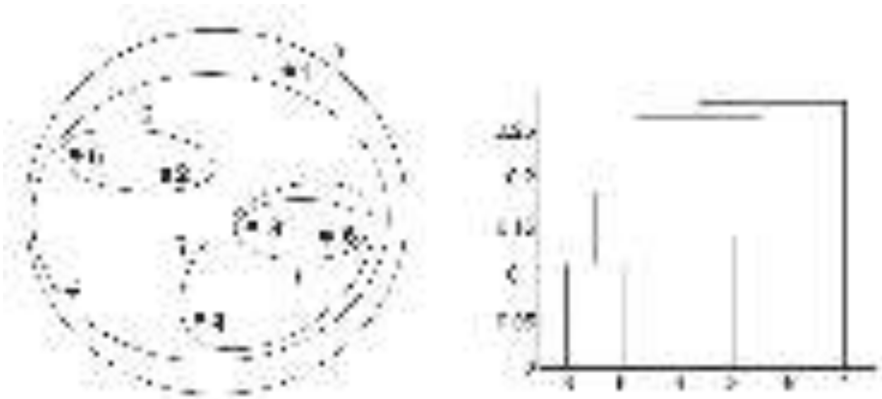
(a) Single link clustering. (b) Single link dendrogram.

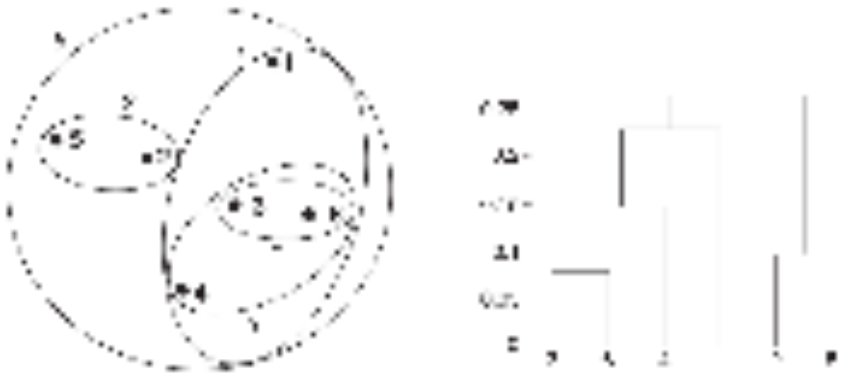
Figure 8.16. Single link clustering of the six points shown in Figure 8.15.



(a) Complete link clustering. (b) Complete link dendrogram.

Figure 8.17. Complete link clustering of the six points shown in Figure 8.15.





Kvaliteta razbitja (Silhouette score)

http://orange.biolab.si/docs/latest/reference/rst/Orange.clustering.kmeans.html

The following code computes the silhouette score for k=2..7 and plots a silhuette plot for k=3 (kmeans-silhouette.py):

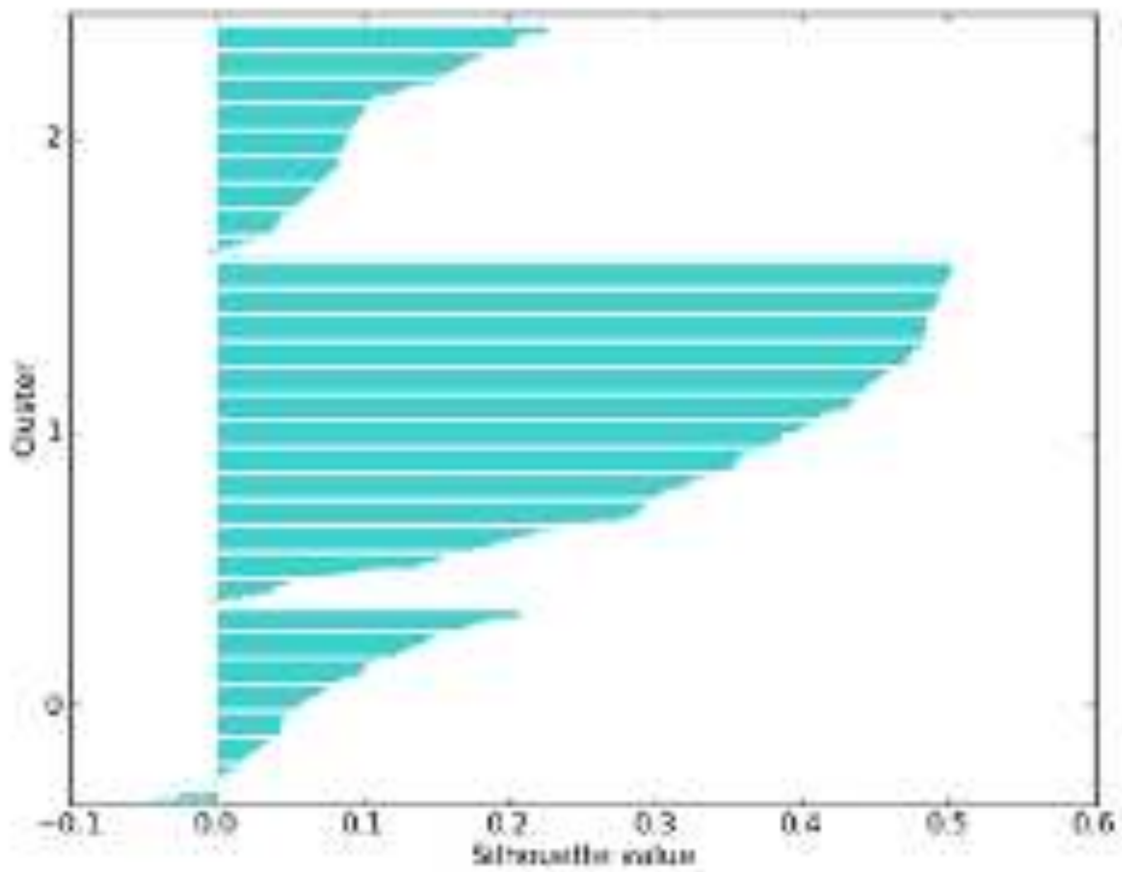
```
import Orange

voting = Orange.data.Table("voting")
# table = Orange.data.Table("iris")

for k in range(2, 8):
    km = Orange.clustering.kmeans.Clustering(voting, k, initialization=Orange.clustering.kmeans.init_c
    score = Orange.clustering.kmeans.score_silhouette(km)
    print k, score

km = Orange.clustering.kmeans.Clustering(voting, 3, initialization=Orange.clustering.kmeans.init_diver
Orange.clustering.kmeans.plot_silhouette(km, "kmeans-silhouette.png")
```

The analysis suggests that k=2 is preferred as it yields the maximal silhouette coefficient:

```
2 0.629447553352
3 0.584210855354
4 0.487259377354
5 0.398628975381
6 0.353228492088
7 0.366357876944
```


[http://orange.biolab.si/docs/latest/reference/rst/Orange.clustering.kmeans/]

Silhouette



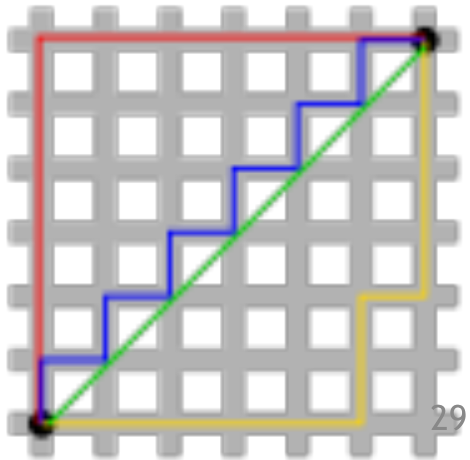
Silhouette plot for k=3.

[http://orange.biolab.si/docs/latest/reference/rst/Orange.clustering.kmeans/]

Primer

	sepal length	sepal width	petal length	petal width
1	4.9	2.4	3.3	1.0
2	6.7	3.0	5.0	1.7
3	5.5	2.4	3.7	1.0
4	6.2	2.9	4.3	1.3
5	6.8	3.0	5.5	2.1
6	6.9	3.1	5.4	2.1
7	6.2	3.4	5.4	2.3
8	5.9	3.0	5.1	1.8

- Single or complete linkage
- Manhattansko razdaljo
  $p_1 = (x_1, y_1)$, $p_2 = (x_2, y_2)$, then
  $d_M = |x_1 - x_2| + |y_1 - y_2|$.
- Nariši dendrogram



Iris virginica

petal leaves

Iris versicolor

sepal leaves

http://statlab.uni-heidelberg.de/data/iris/

Skupine so seveda znane …

	sepal length	sepal width	petal length	petal width	iris
1	4.9	2.4	3.3	1.0	Iris-versicolor
2	6.7	3.0	5.0	1.7	Iris-versicolor
3	5.5	2.4	3.7	1.0	Iris-versicolor
4	6.2	2.9	4.3	1.3	Iris-versicolor
5	6.8	3.0	5.5	2.1	Iris-virginica
6	6.9	3.1	5.4	2.1	Iris-virginica
7	6.2	3.4	5.4	2.3	Iris-virginica
8	5.9	3.0	5.1	1.8	Iris-virginica

- Kakšno je ujemanje med odkritimi skupinami in dejanskim razredom?