Clustering Methods

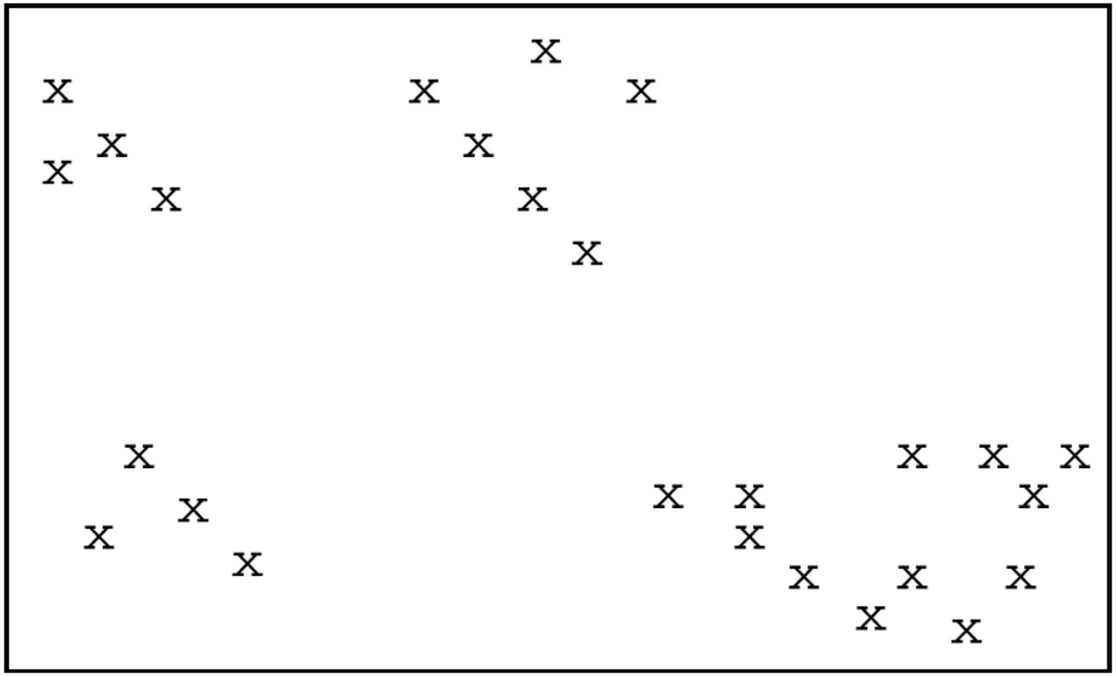
Prof. Asim Tewari
IIT Bombay

What are Clustering Methods?

- Clustering refers to a very broad set of techniques for finding subgroups, or clusters, in a data set.
- Clustering is concerned with grouping together objects that are similar to each other and dissimilar to the objects belonging to other clusters.
- Clustering is unsupervised learning that assigns labels to objects in unlabeled data.

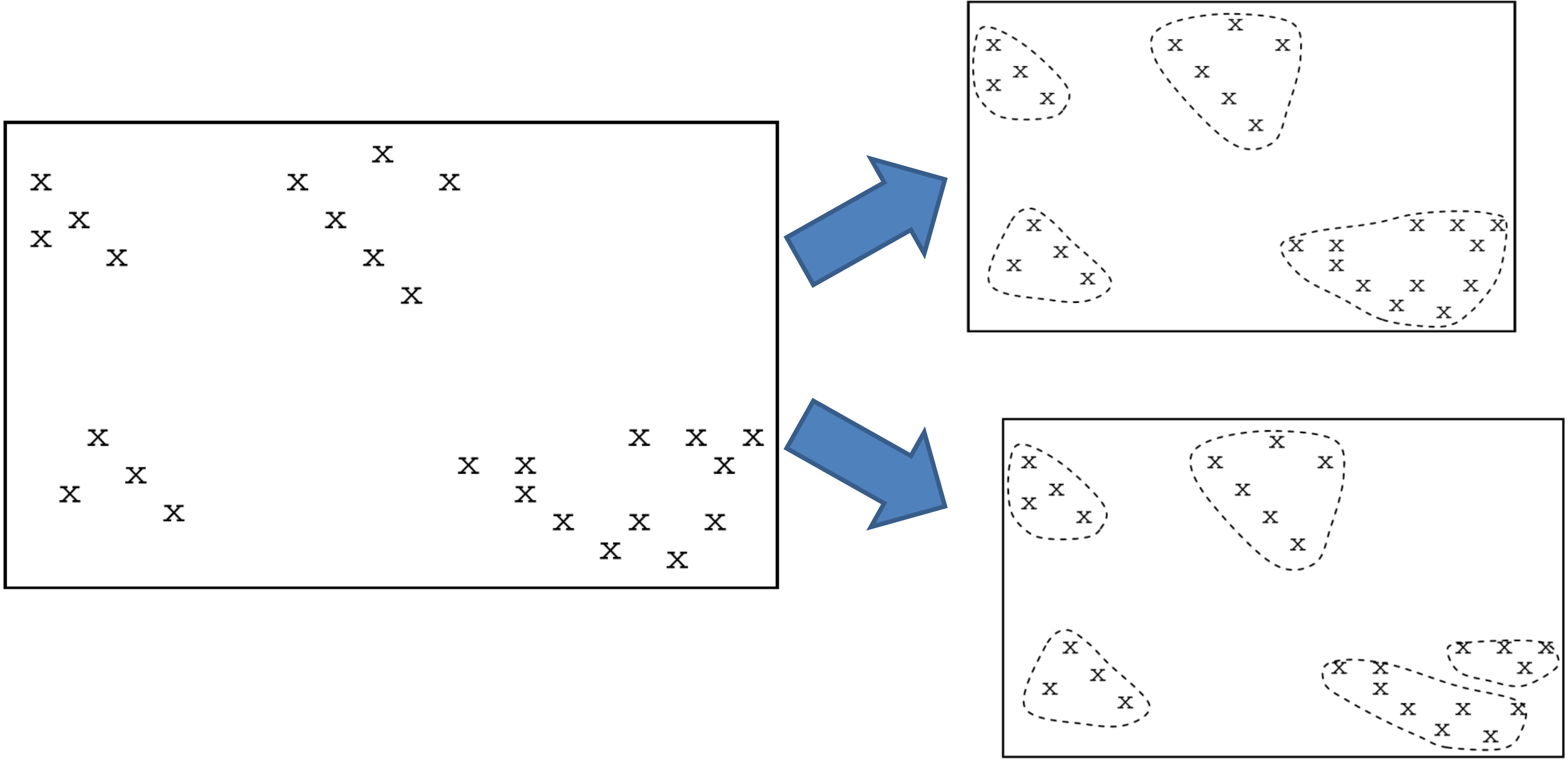
Example of clustering

- How many clusters do you see?



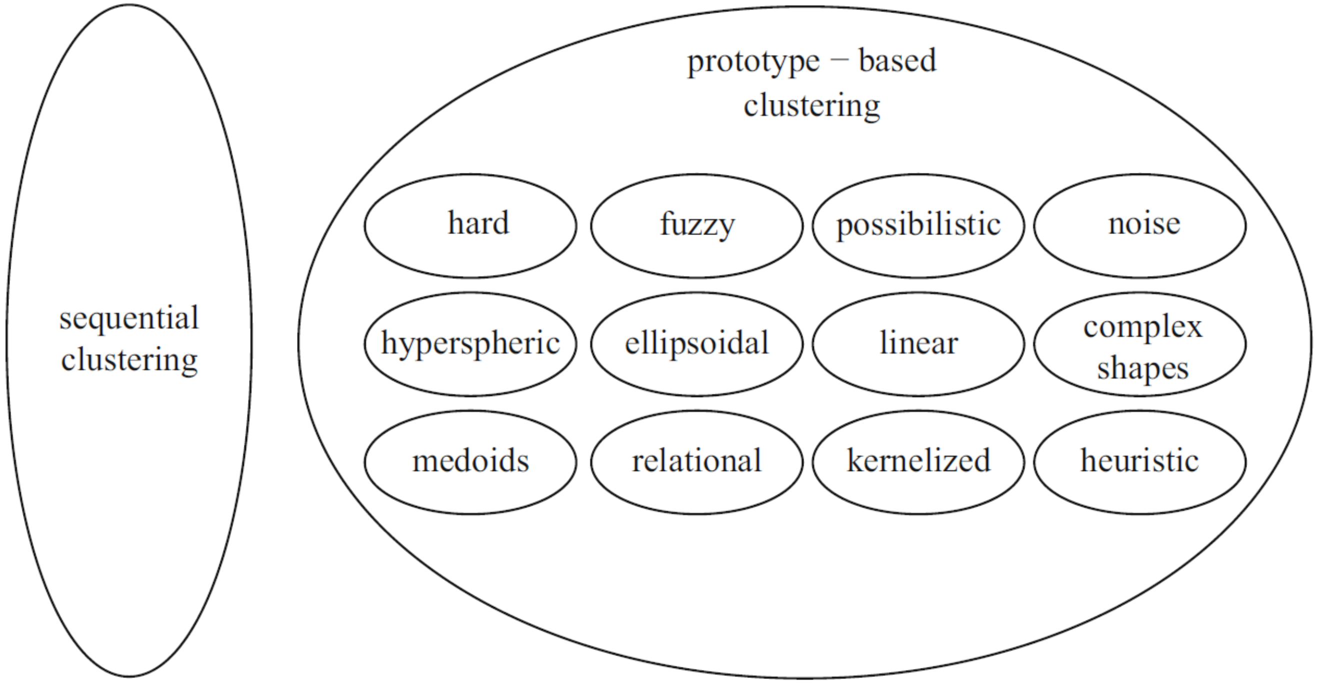
Example of clustering

- How many clusters do you see?



Types of Clustering Classification-1

- Sequential Clustering
- Prototype based clustering



Types of Clustering Classification-2

- K-means clustering
  - In this we seek to partition the observations into a pre-specified number of clusters
- Hierarchical clustering
  - In this we do not know in advance how many clusters we want; in fact, we end up with a tree-like visual representation of the observations, called a *dendrogram*

K-Means Clustering

Partitioning a data set into *K* distinct, non-overlapping clusters

- First specify the desired number of clusters *K*
- Then the *K*-means algorithm will assign each observation to exactly one of the *K* clusters
- The idea behind K-means clustering is that a good clustering is one for which the within-cluster variation is as small as possible

K-Means Clustering method

Let $C1, \ldots, CK$ denote sets containing the indices of the observations in each cluster. These sets satisfy two properties:

1. $C1 \cup C2 \cup \ldots \cup CK = \{1, \ldots, n\}$. In other words, each observation belongs to at least one of the K clusters.
2. $Ck1 \cap Ck2 = \emptyset$ for all $k1 \neq k2$. In other words, the clusters are nonoverlapping: no observation belongs to more than one cluster.

K-Means Clustering method..

- Define a $W(C_k)$ as within-cluster variation measure.
- To obtain the k-means cluster solve the following problem

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} W(C_k) \right\}$$

This formula says that we want to partition the observations into K clusters such that the total within-cluster variation, summed over all K clusters, is as small as possible

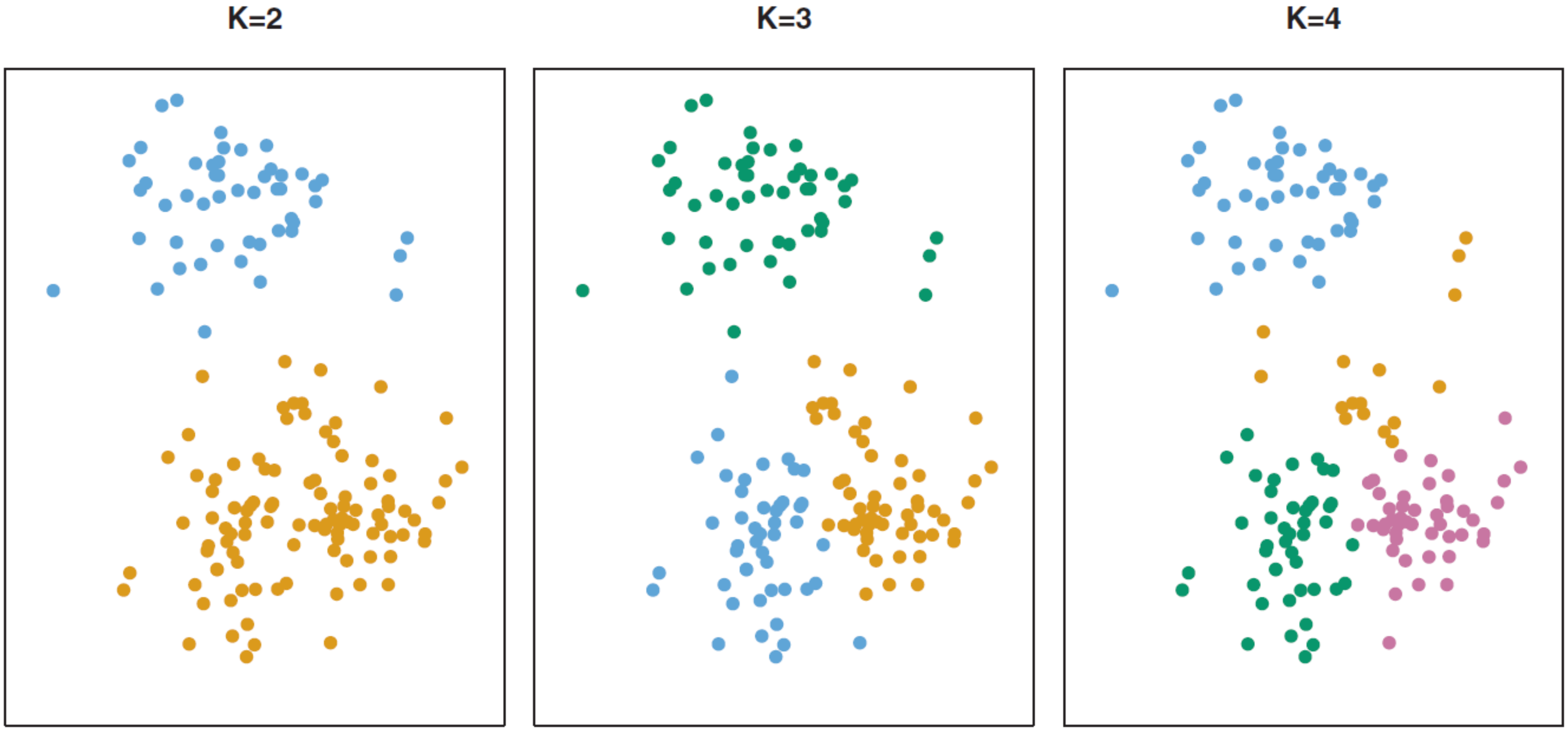
K-Means Clustering method..

- There are many possible ways to define the *$W(C_k)$*, within-cluster variation measure
- Most common is squared Euclidean distance

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

Where $|C_k|$ denotes the number of observations in the kth cluster

K-Means Clustering



K-Means Clustering Algorithm

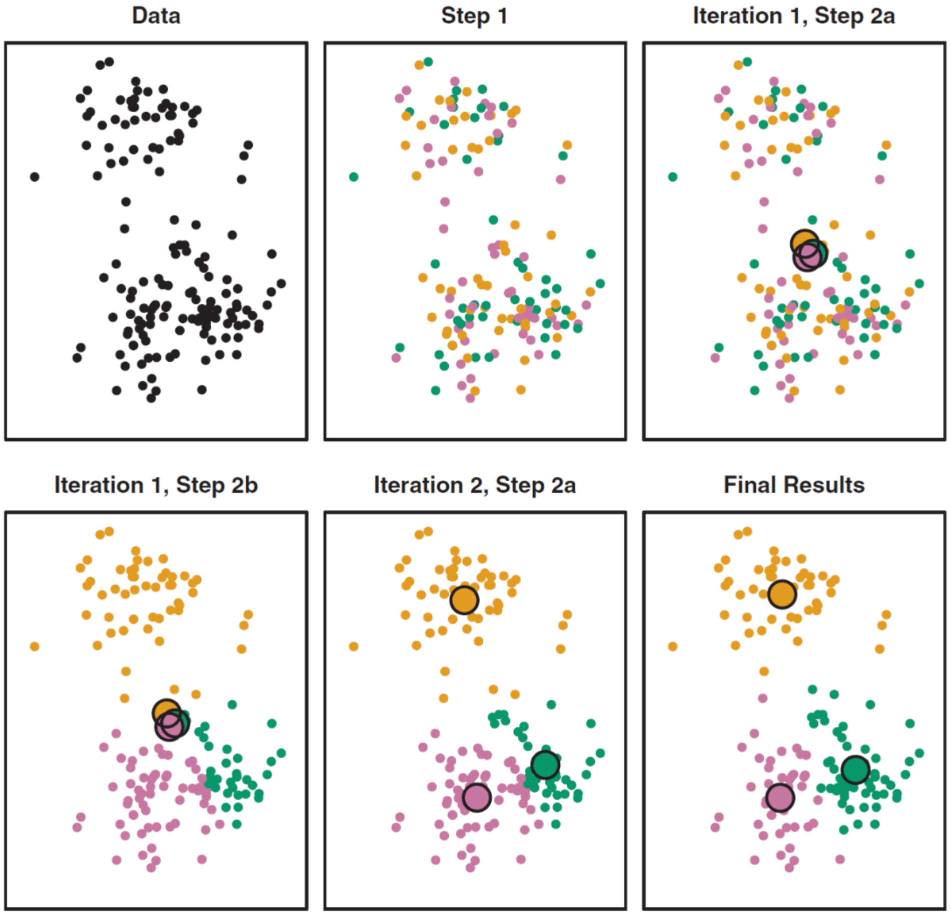
1. Randomly assign a number, from 1 to *K*, to each of the observations. These serve as initial cluster assignments for the observations.

2. Iterate until the cluster assignments stop changing:
   a) For each of the *K* clusters, compute the cluster *centroid*. The *k*th cluster centroid is the vector of the *p* feature means for the observations in the *k*th cluster.
   b) Assign each observation to the cluster whose centroid is closest (where *closest* is defined using Euclidean distance).

K-Means Clustering Algorithm..

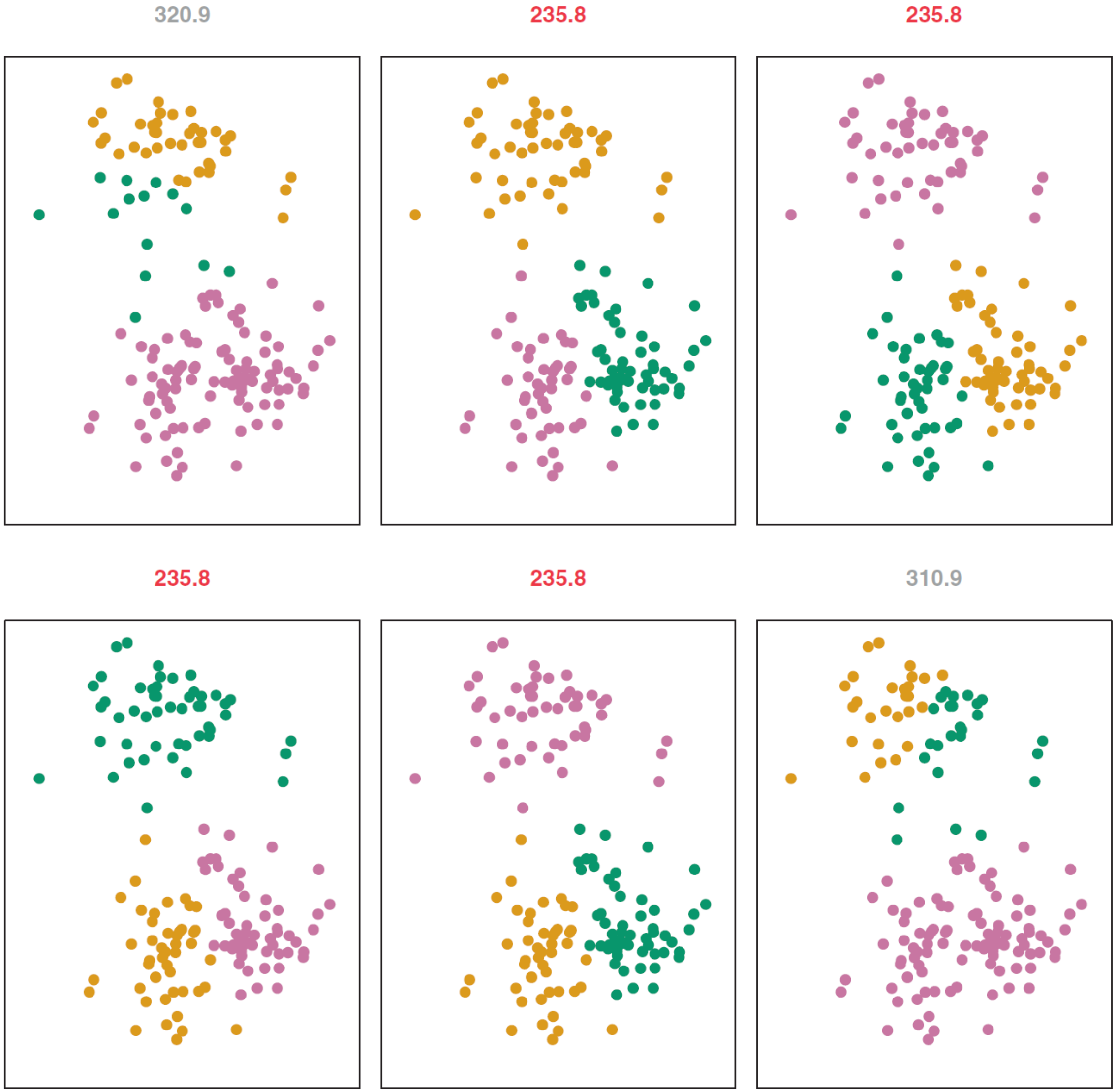
- This algorithm is guaranteed to decrease the value of the objective at each step
- This means that as the algorithm is run, the clustering obtained will continually improve until the result no longer changes (the objective function will never increase)
- When the result no longer changes, a **local optimum** has been reached
- The resulting clustering will depend on the initial (random) cluster assignment

K-Means Clustering Algorithm..



Top left: the observations are shown.
Top center: in Step 1 of the algorithm, each observation is randomly assigned to a cluster.
Top right: in Step 2(a), the cluster centroids are computed. These are shown as large colored disks. Initially the centroids are almost completely overlapping because the initial cluster assignments were chosen at random.
Bottom left: in Step 2(b), each observation is assigned to the nearest centroid.
Bottom center: Step 2(a) is once again performed, leading to new cluster centroids.
Bottom right: the results obtained after ten iterations.

K-Means Clustering Algorithm..



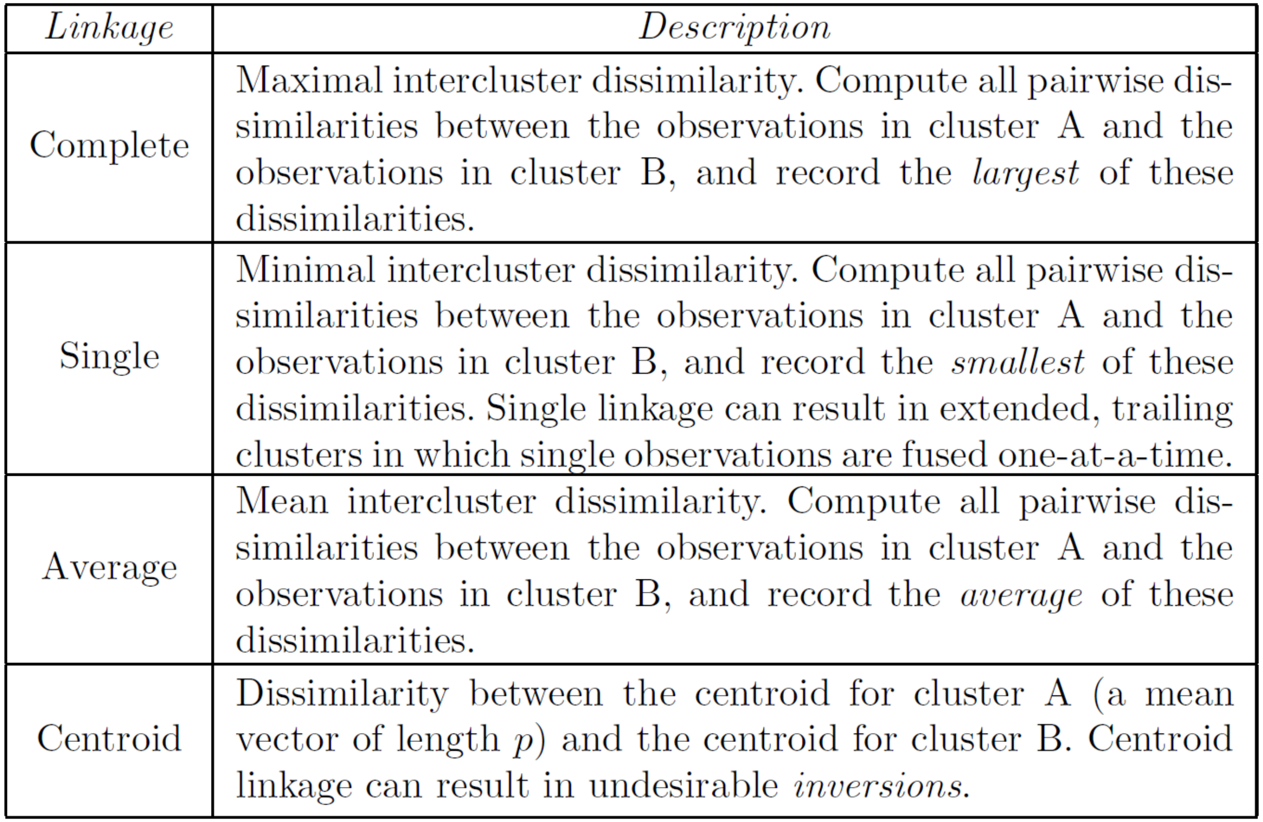
- *K-means clustering with a different initial random assignment. Above each plot is the value of the objective function.*
- *Three different local optima were obtained*

Hierarchical Clustering

- In agglomerative Hierarchical Clustering we start with each object in a cluster of its own and then repeatedly merge the closest pair of clusters until we end up with just one cluster containing everything
- This results in a tree-based representation of the observations, called a *dendrogram*

Hierarchical Clustering

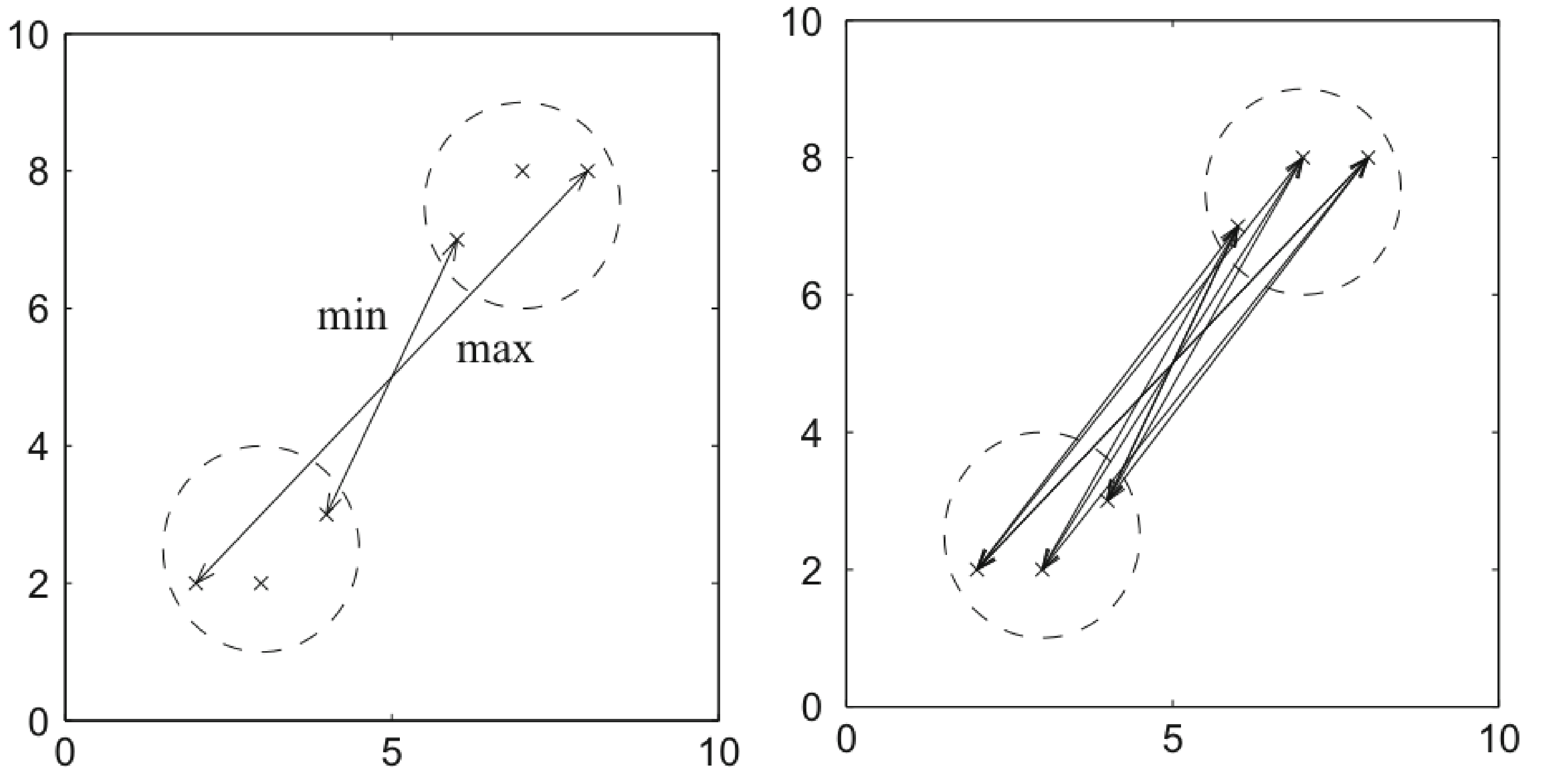
- *Dissimilarity measure between a pair of groups is defined by developing the notion of Linkage*
- *The four most common types of linkage are complete, average, single and centroid*



Linkage	*Description*
Complete	Maximal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *largest* of these dissimilarities.
Single	Minimal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *smallest* of these dissimilarities. Single linkage can result in extended, trailing clusters in which single observations are fused one-at-a-time.
Average	Mean intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *average* of these dissimilarities.
Centroid	Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B. Centroid linkage can result in undesirable *inversions*.

Hierarchical Clustering

Dissimilarity measure between a pair of groups.



Distances between clusters (left: single and complete linkage, right: average linkage)

Hierarchical Clustering

Dissimilarity measure between a pair of groups.

- minimum distance or *single linkage*

$$d(C_r, C_s) = \min_{x \in C_r,\, y \in C_s} d(x, y)$$

- maximum distance or *complete linkage*

$$d(C_r, C_s) = \max_{x \in C_r,\, y \in C_s} d(x, y)$$

- average distance or *average linkage*

$$d(C_r, C_s) = \frac{1}{|C_r| \cdot |C_s|} \sum_{x \in C_r,\, y \in C_s} d(x, y)$$

Hierarchical Clustering

Dissimilarity measure between a pair of groups.

- distance of the centers

$$d(C_r, C_s) = \left\| \frac{1}{|C_r|} \sum_{x \in C_r} x - \frac{1}{|C_s|} \sum_{x \in C_s} x \right\|$$

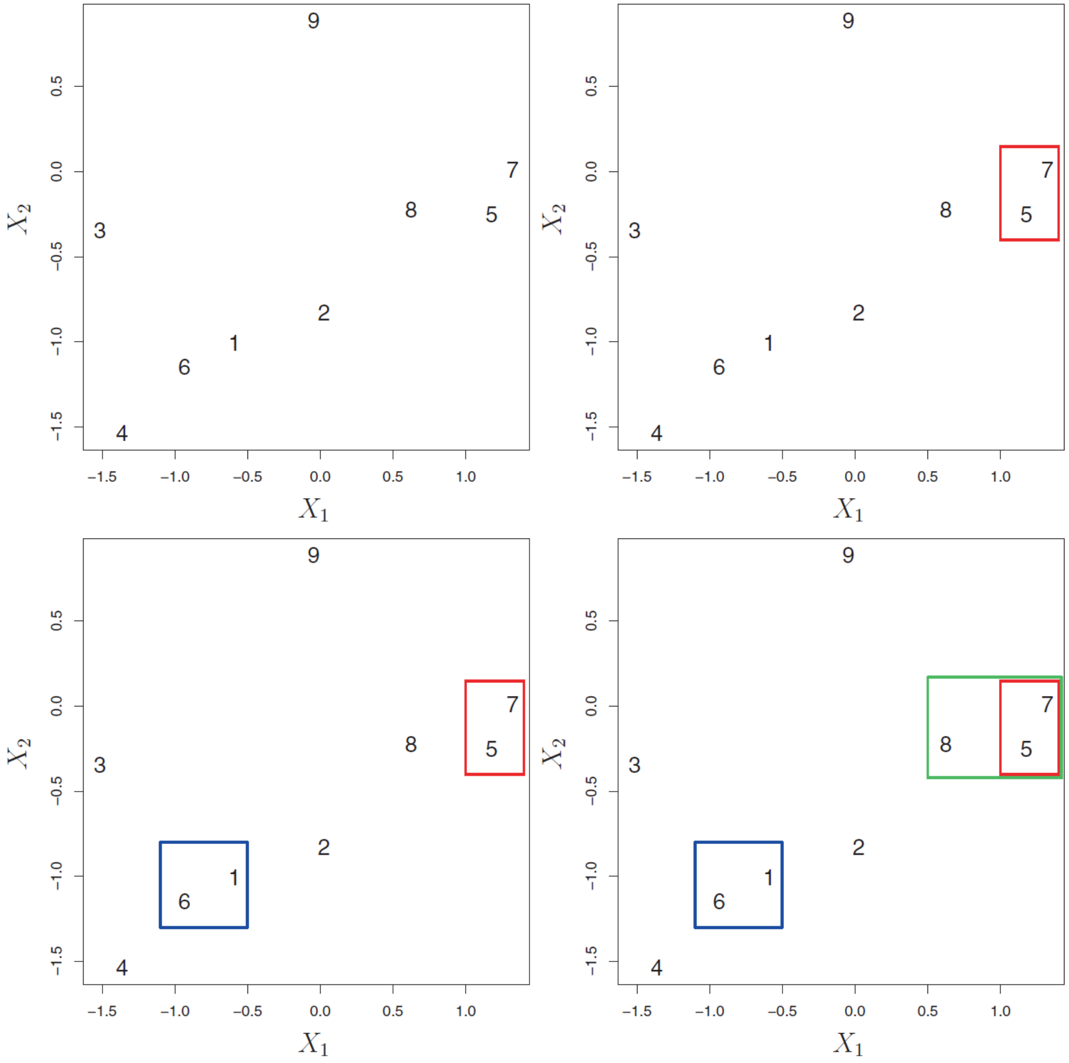
- Ward's measure

$$d(C_r, C_s) = \frac{|C_r| \cdot |C_s|}{|C_r| + |C_s|} \left\| \frac{1}{|C_r|} \sum_{x \in C_r} x - \frac{1}{|C_s|} \sum_{x \in C_s} x \right\|$$

Hierarchical Clustering Algorithm

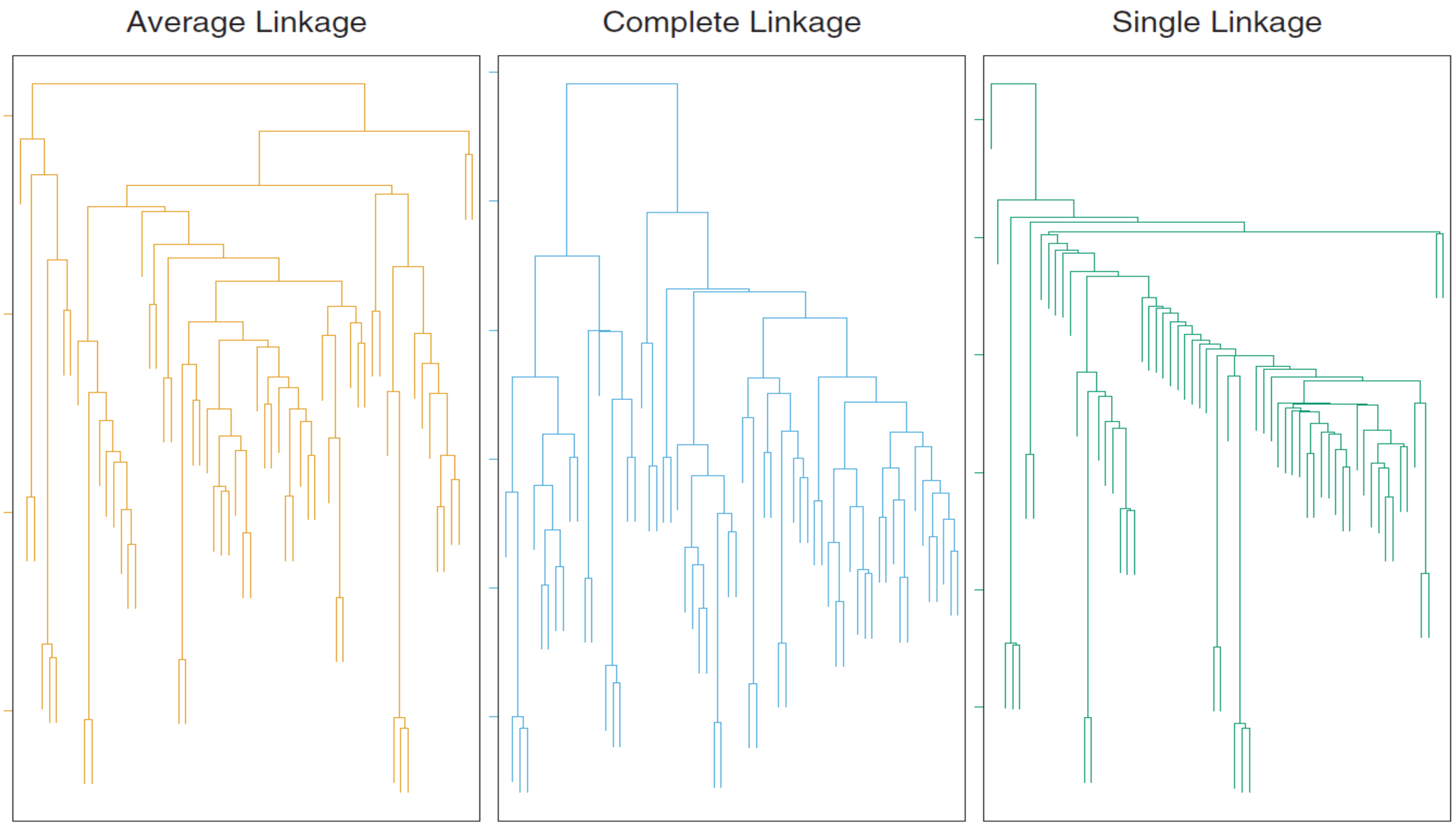
1. Assign each object to its own single-object cluster. Calculate the dissimilarity between each pair of clusters.
2. Choose the closest pair of clusters and merge them into a single cluster (so reducing the total number of clusters by one).
3. Calculate the dissimilarity between the new cluster and each of the old clusters.
4. Repeat steps 2 and 3 until all the objects are in a single cluster.

Hierarchical Clustering Algorithm…



An illustration of the first few steps of the hierarchical clustering algorithm,

Effect of linkage on Hierarchical Clustering



Average Linkage

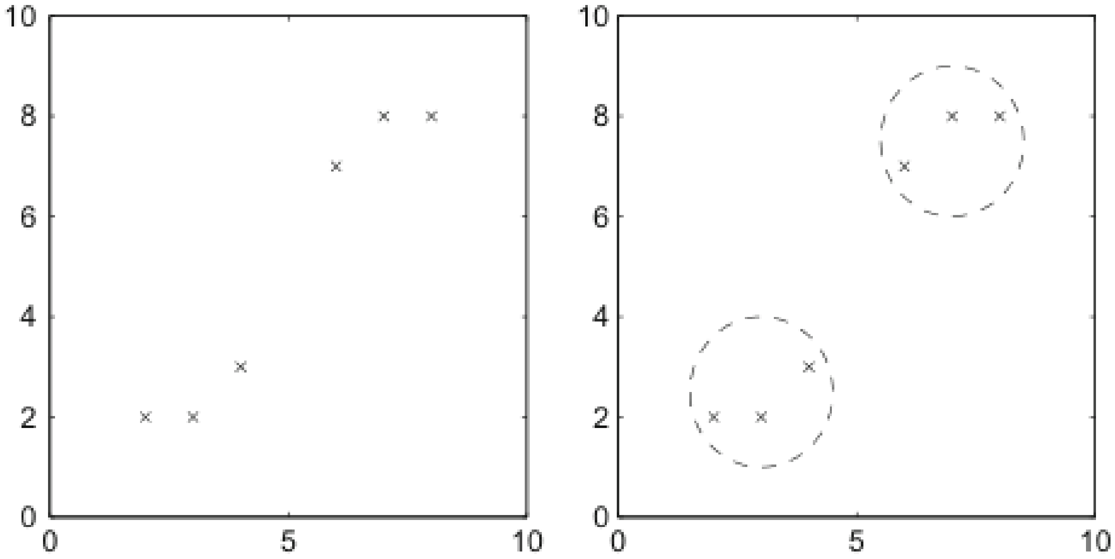
Complete Linkage

Single Linkage

- *Average, complete, and single linkage applied to an example data set.*
- *Average and complete linkage tend to yield more balanced clusters.*

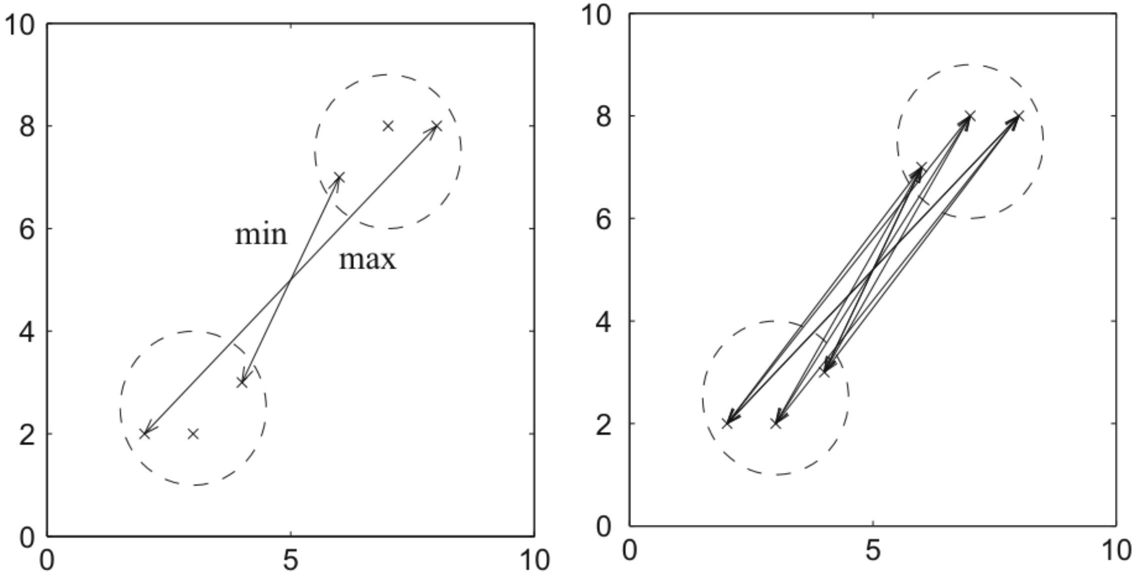
Hierarchical Clustering

X= {(2,2), (3,2), (4,3), (6,7), (7,8), (8,8)}



Hierarchical Clustering

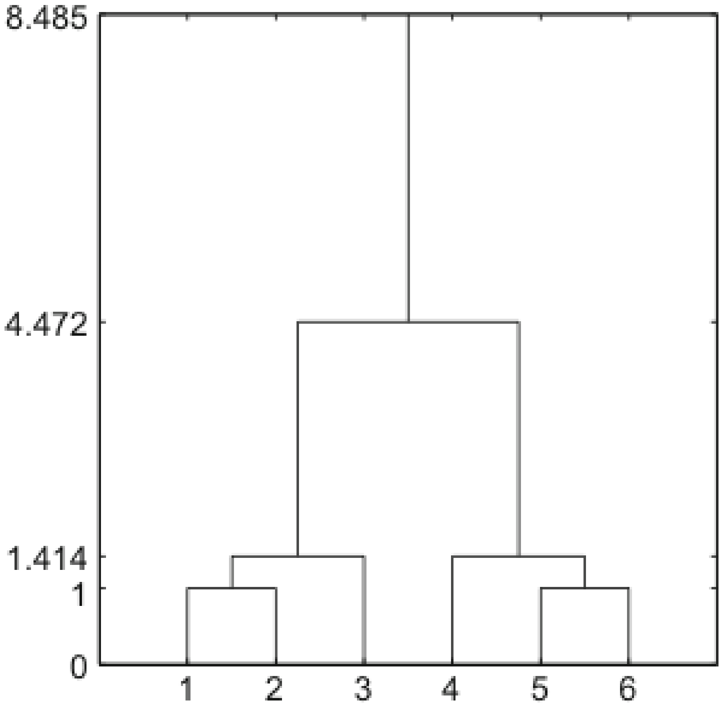
X= {(2,2), (3,2), (4,3), (6,7), (7,8), (8,8)}



Distances between clusters (left: single and complete linkage, right: average linkage)

Hierarchical Clustering Dendrogram

X= {(2,2), (3,2), (4,3), (6,7), (7,8), (8,8)}



The figure shows the single linkage dendrogram for our simple six point data set, that illustrates how the points x1, …, x6 (indices on the horizontal axis) are successively merged. The vertical axis shows the single linkage distances. The single linkage partitions are

$$\Gamma_0 = \{\{x_1\}, \{x_2\}, \{x_3\}, \{x_4\}, \{x_5\}, \{x_6\}\}$$

$$\Gamma_1 = \Gamma_2 = \{\{x_1, x_2\}, \{x_3\}, \{x_4\}, \{x_5, x_6\}\}$$

$$\Gamma_3 = \Gamma_4 = \{\{x_1, x_2, x_3\}, \{x_4, x_5, x_6\}\}$$

$$\Gamma_5 = = \{\{x_1, x_2, x_3, x_4, x_5, x_6\}\} = \{X\}$$

Practical consideration in Clustering

- Should the observations or features first be standardized in some way? For instance, maybe the variables should be centered to have mean zero and scaled to have standard deviation one

- In the case of *K*-means clustering, how many clusters should we look for in the data?

- In the case of hierarchical clustering,
  - What dissimilarity measure should be used?
  - What type of linkage should be used?
  - Where should we cut the dendrogram in order to obtain clusters?

- How to validate the clusters obtained

Prototype-Based Clustering

In the previous section we represented a partition of X by a *partition set* Γ containing disjoint subsets of X. An equivalent representation is a *partition matrix* U with the elements

$$u_{ik} = \begin{cases} 1 & \text{if } x_k \in C_i \\ 0 & \text{if } x_k \notin C_i \end{cases}$$

$i = 1, \ldots, c$, $k = 1, \ldots, n$, so u_{ik} is a *membership value* that indicates if x_k belongs to C_i. For non-empty clusters we require

$$\sum_{k=1} u_{ik} > 0, \quad i = 1, \ldots, c$$

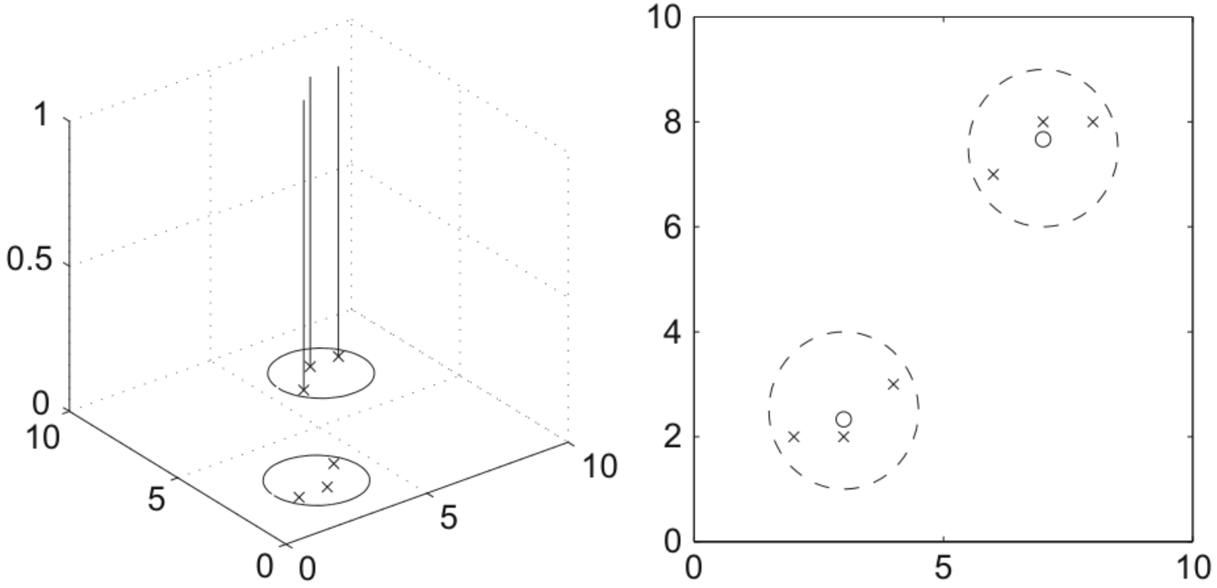
and for pairwise disjoint clusters

$$\sum_{i=1}^{c} u_{ik} = 1, \quad k = 1, \ldots, n$$

Prototype-Based Clustering

$$u_{ik} = \begin{cases} 1 & \text{if } x_k \in C_i \\ 0 & \text{if } x_k \notin C_i \end{cases}$$

$$V = \{v_1, \ldots, v_c\} \subset \mathbb{R}^p$$



Cluster memberships and cluster centers.

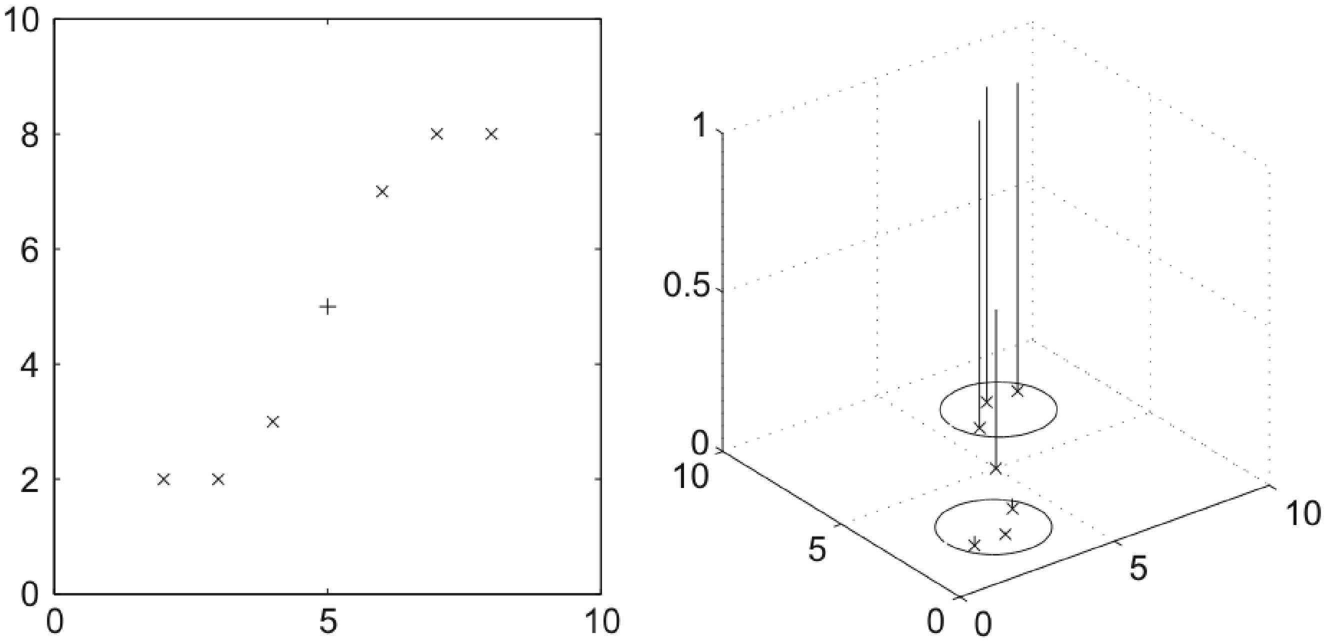
Each cluster may be represented by a (single) center v_i, *i =1,..,c*, so the cluster structure is defined by the set of cluster centers.

Given a data set X, the cluster centers V and the assignment of data points X to clusters can be found by optimizing the c-means (CM) clustering model.

Fuzzy Clustering

CM clustering works well if the clusters are well separated and do not contain inliers or outliers. We add the point $x_7 = (5, 5)$ to the previously considered six point data set and obtain the new seven point data set

$$X = \{(2, 2), (3, 2), (4, 3), (6, 7), (7, 8), (8, 8), (5, 5)\}$$



Data set with a point that belongs to both clusters (left) and fuzzy partition (right)