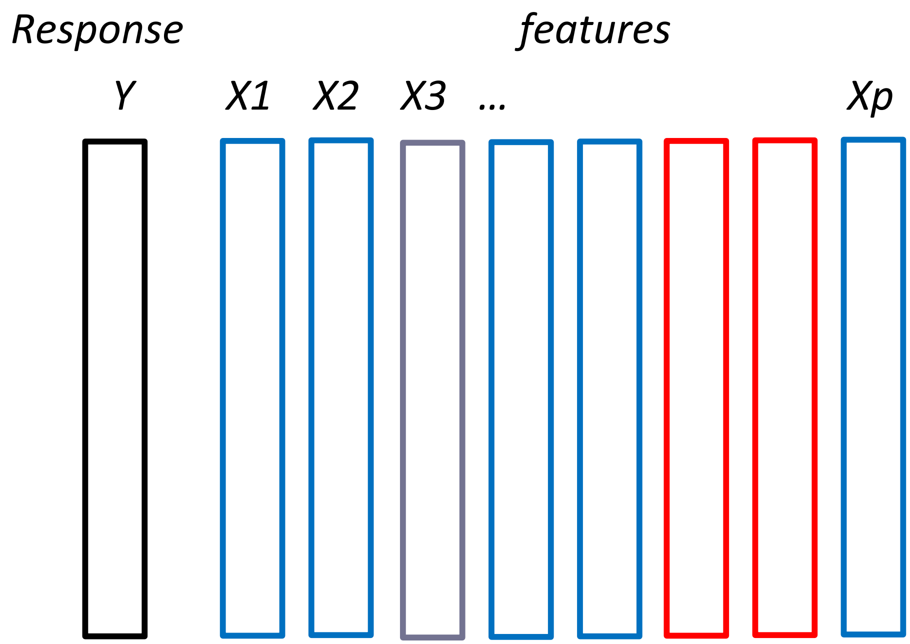
CLUSTERING

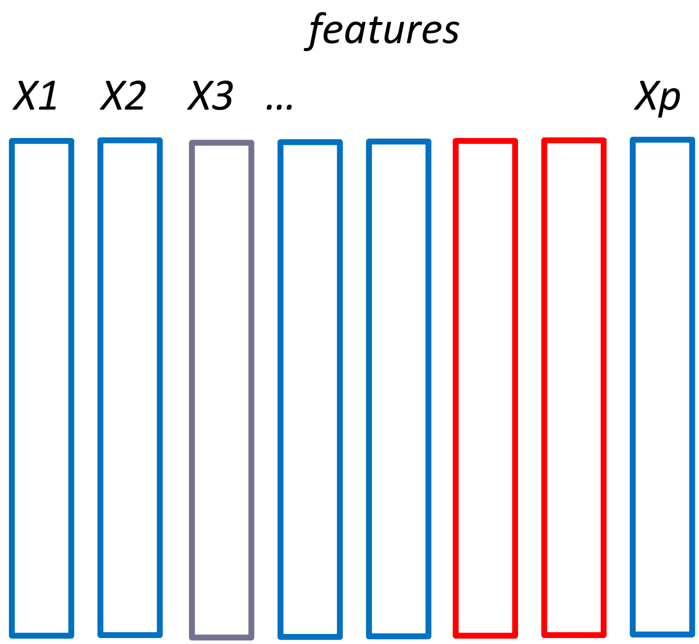
Clustering

- Methods for finding groups, or clusters, from a population
- Groups with common characteristics, attributes
- A good clustering is one where the observations *within* a group are similar but observations *between* groups are different

Supervised Learning



Unsupervised Learning



No Response

Outline

- Supervised learning
  - Classification Problem
    - - KNN

- Unsupervised learning
  - Clustering
    - - K Means
    - - Hierarchical clustering

Outline

- Supervised learning

  Classification Problem

  - KNN

- Unsupervised learning

  Clustering (distance-based methods)

  - K Means
  - Hierarchical clustering

Outline

- Classification
  - Categories are known
  - Predict the category of new observations
- Clustering
  - Discover categories (clusters) of a new categorical variable

K-MEANS CLUSTERING

K-Means Clustering

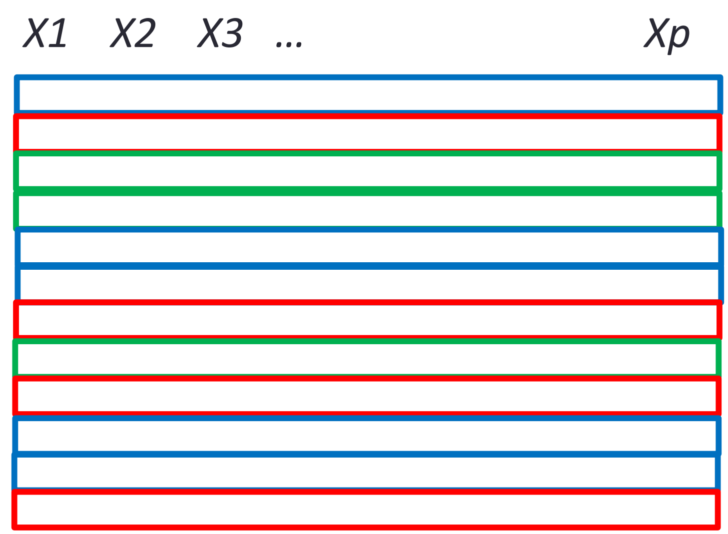
- Example

    Want to find clusters from a data set

K-Means Clustering

- Example

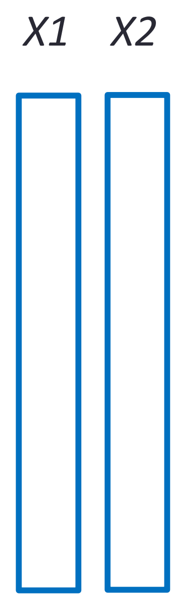
Want to find clusters from a data set



K-Means Clustering

- Example

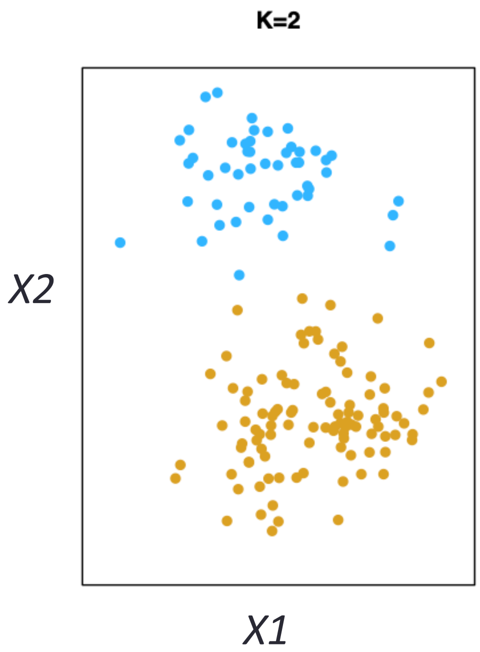
Want to find clusters from a data set with two features only

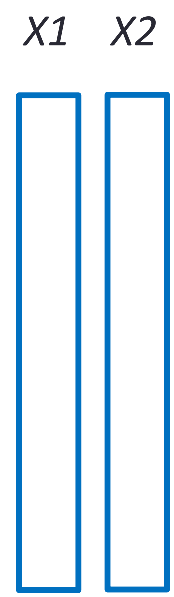


K-Means Clustering

- Example

 Want to find K clusters from a data set with two features only

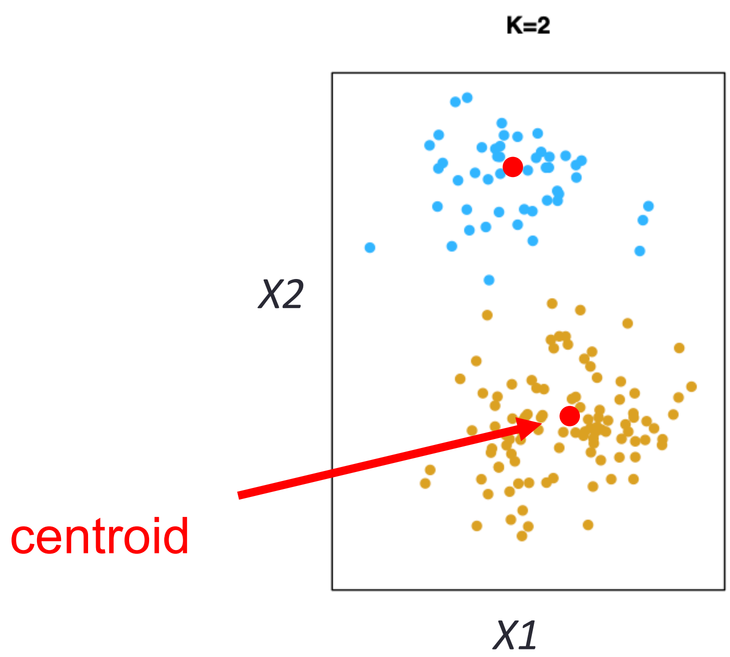


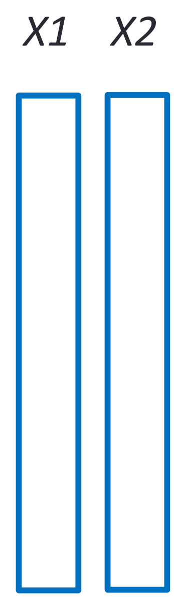


K-Means Clustering

- Example

Want to find K clusters from a data set with two features only

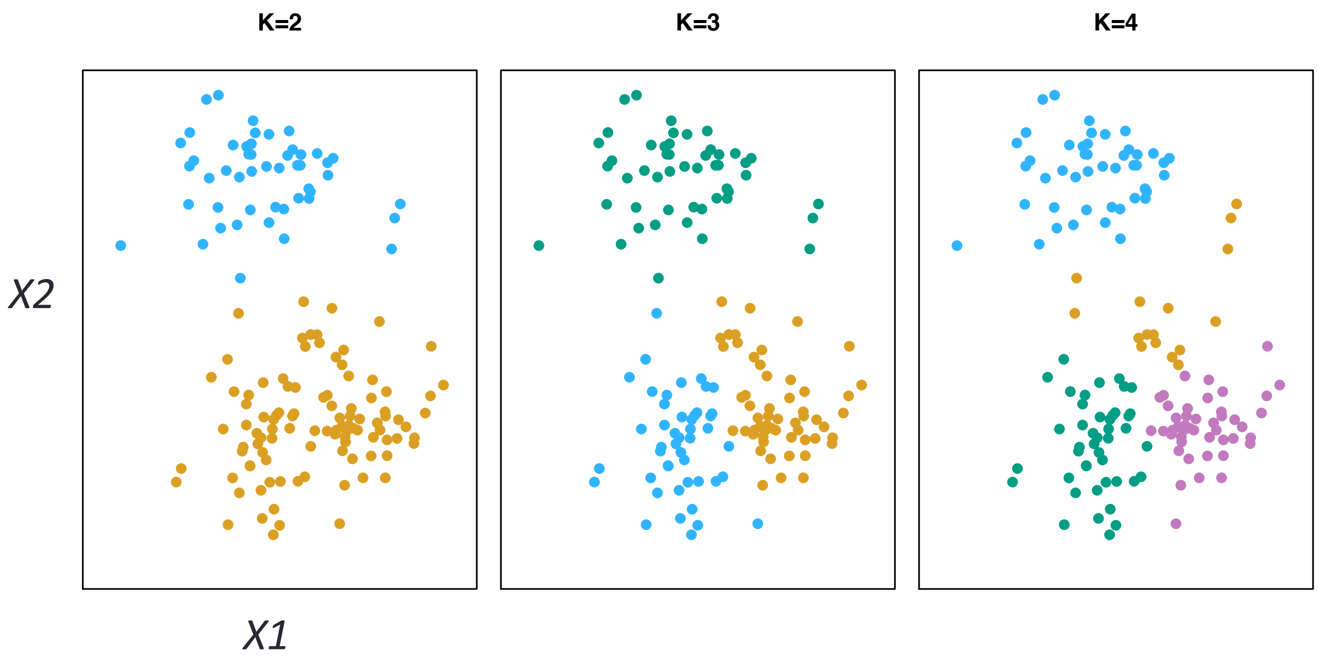




K-Means Clustering

- Example

Want to find K clusters from a data set with two features only



K-Means Clustering

- A good clustering is one where the observations *within* a group are similar but observations *between* groups are different
- A good clustering provides smallest *within-cluster variation*
- Because observations *within* a group are deemed to be similar
- How to measure *within-cluster variation*?

Within-cluster variation

- Find the squared-distance between observations 1 and 2

X_1	X_2	X_3	$\ldots$	X_p
x_{11}	x_{12}	x_{13}	$\ldots$	x_{1p}
x_{21}	x_{22}	x_{23}	$\ldots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_{n1}	x_{n2}	x_{n3}	$\ldots$	x_{np}

Within-cluster variation

- Find the squared-distance between observations 1 and 2

$$
\begin{aligned}
d_{12}^2 &= (x_{21} - x_{11})^2 + (x_{22} - x_{12})^2 + \cdots + (x_{2p} - x_{1p})^2 \\
&= \sum_{m=1}^{p} (x_{2m} - x_{1m})^2
\end{aligned}
$$

X_1	X_2	X_3	$\dots$	X_p
$\boxed{x_{11}}$	$\boxed{x_{12}}$	$\boxed{x_{13}}$	$\boxed{\dots}$	$\boxed{x_{1p}}$
$\boxed{x_{21}}$	$\boxed{x_{22}}$	$\boxed{x_{23}}$	$\boxed{\dots}$	$\boxed{x_{2p}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_{n1}	x_{n2}	x_{n3}	$\dots$	x_{np}

Within-cluster variation

- Find the squared-distance between *all* observations

$$
\begin{aligned}
d_{12}^2 &= (x_{21} - x_{11})^2 + (x_{22} - x_{12})^2 + \cdots + (x_{2p} - x_{1p})^2 \\
&= \sum_{m=1}^{p} (x_{2m} - x_{1m})^2
\end{aligned}
$$

X_1	X_2	X_3	$\ldots$	X_p
x_{11}	x_{12}	x_{13}	$\ldots$	x_{1p}
x_{21}	x_{22}	x_{23}	$\ldots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_{n1}	x_{n2}	x_{n3}	$\ldots$	x_{np}

$$
d^2 = \sum_{\text{all pairs i,j}} \sum_{m=1}^{p} (x_{im} - x_{jm})^2
$$

Within-cluster variation

- Notation for clusters

X_1	X_2	X_3	$\dots$	X_p
x_{11}	x_{12}	x_{13}	$\dots$	x_{1p}
x_{21}	x_{22}	x_{23}	$\dots$	x_{2p}
x_{31}	x_{32}	x_{33}	$\dots$	x_{3p}
x_{41}	x_{42}	x_{43}	$\dots$	x_{4p}
x_{51}	x_{52}	x_{53}	$\dots$	x_{5p}
x_{61}	x_{62}	x_{63}	$\dots$	x_{6p}
x_{71}	x_{72}	x_{73}	$\dots$	x_{7p}
x_{81}	x_{82}	x_{83}	$\dots$	x_{8p}
x_{91}	x_{92}	x_{93}	$\dots$	x_{9p}

$$
\begin{aligned}
C_1 &= \{2, 3, 9\} & \qquad |C_1| = 3 \\
C_2 &= \{4, 6, 7, 8\} & \qquad |C_2| = 4 \\
C_3 &= \{1, 5\} & \qquad |C_3| = 2
\end{aligned}
$$

Within-cluster variation

Consider K clusters $C_1, C_2, \ldots, C_K$

For the r^{th} cluster, with $|C_r|$ observations, the within-cluster variation is

$$WCV_r \quad = \quad \frac{1}{|C_r|} \sum_{i,j \,\epsilon\, C_r} \sum_{m=1}^{p} (x_{im} - x_{jm})^2$$

For all clusters the within-cluster variation is $$TWCV = \sum_{k=1}^{K} WCV_k$$

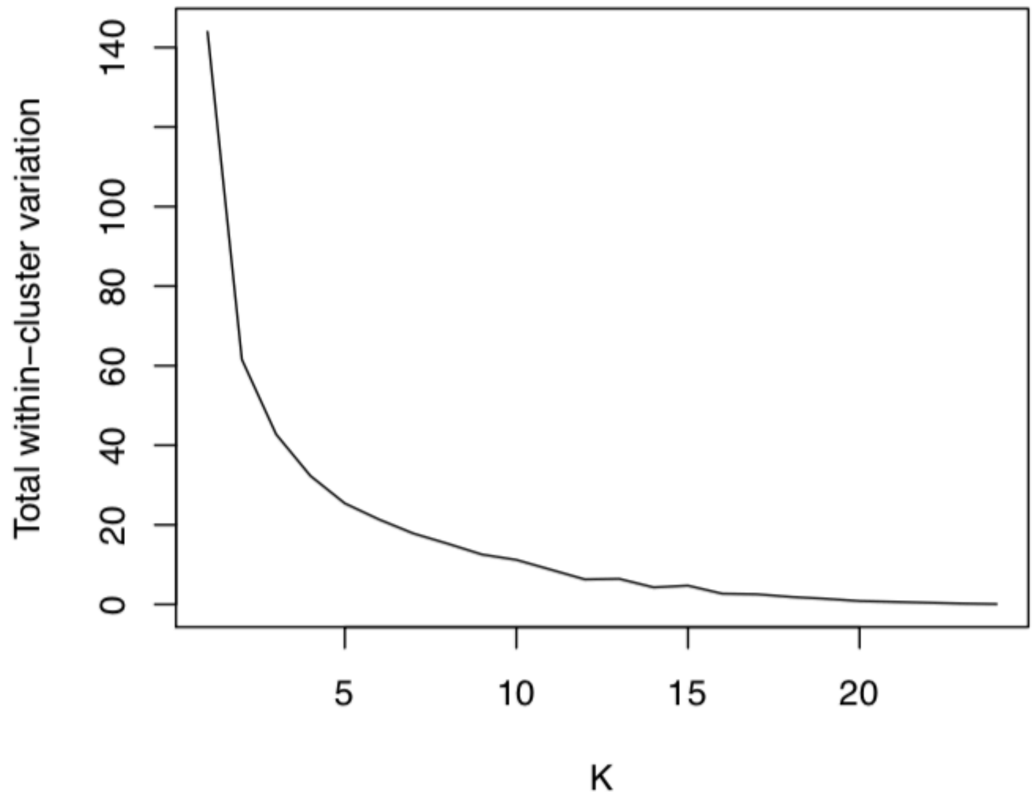
Within-cluster variation

How to find clusters C_1, C_2, …, C_K that result in the smallest

$$TWCV = \sum_{k=1}^{K} WCV_k$$

Within-cluster variation

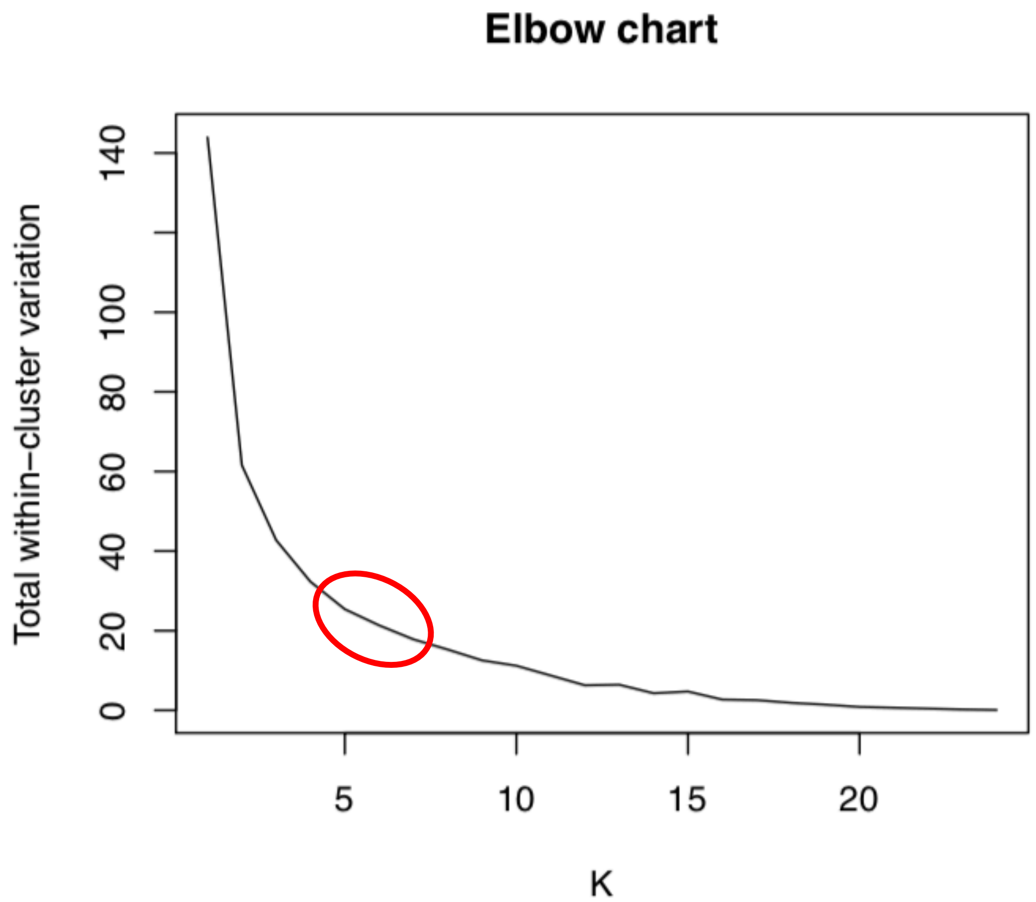
How to choose *K*?



Within-cluster variation

How to choose *K*?

Identify the point when the TWCV starts decreasing slowly



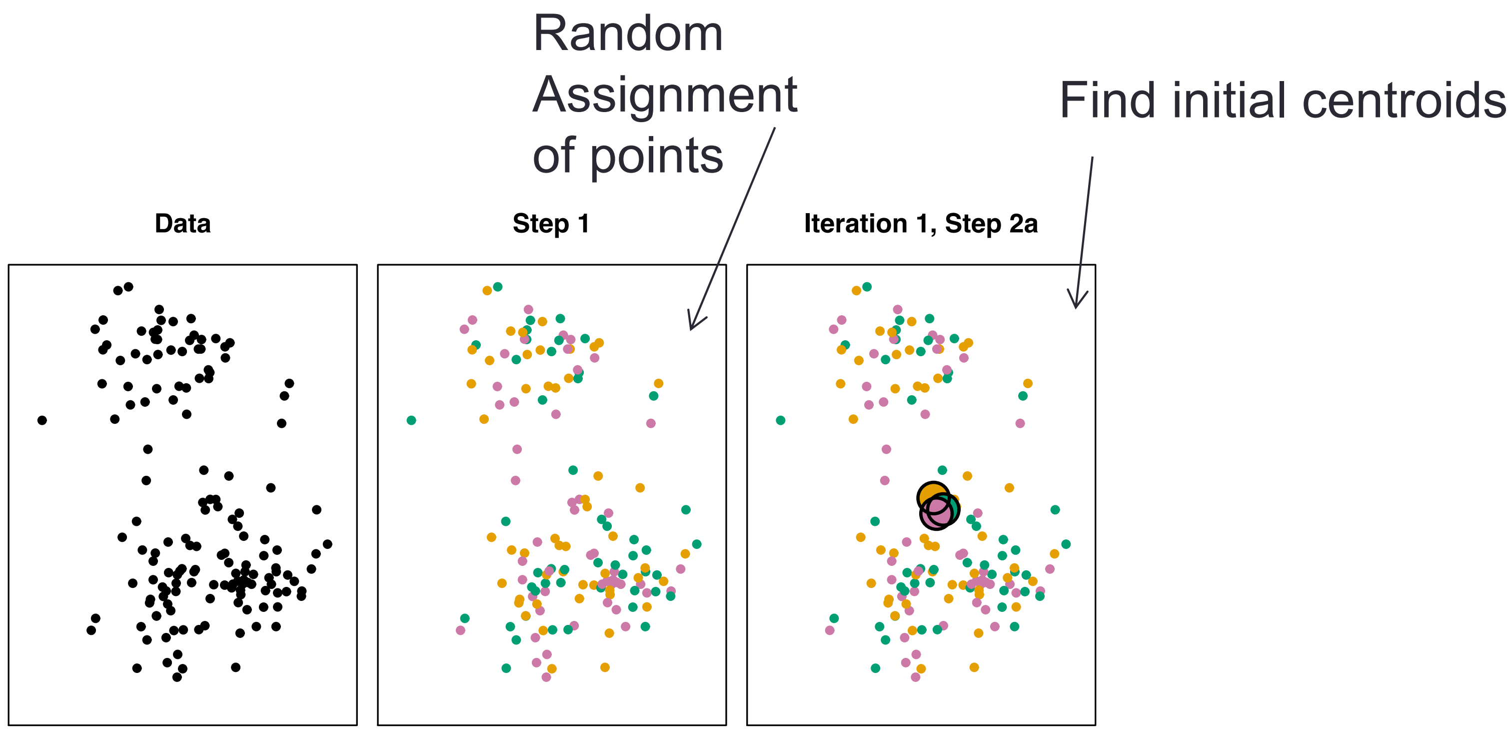
K-Means Algorithm

- Fix K
- Randomly assign an integer (1 to K)
  to each observation (row)
- These assignments are the initial cluster assignments

K-Means Algorithm

- Fix K
- Randomly assign an integer (1 to K)
  to each observation (row)
- These assignments are the initial cluster assignments
- Repeat
  - Find the centroid of each cluster
  - For each observation, find the distance to each centroid
  - Assign the observation to the closest centroid
- Finish when the cluster assignments stop changing

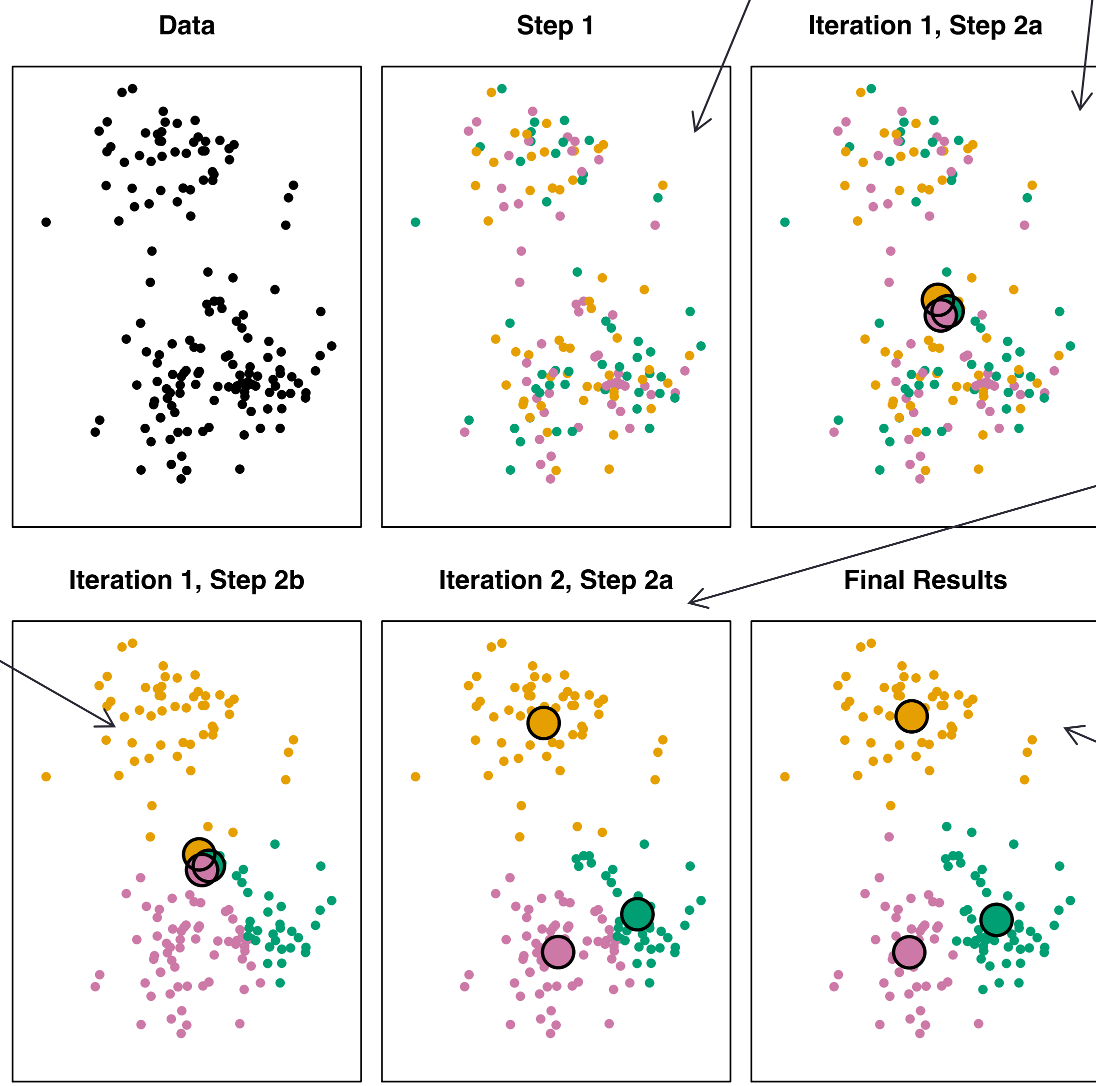
The K-Means Algorithm



The K-Means Algorithm

Random Assignment of points

Find initial centroids



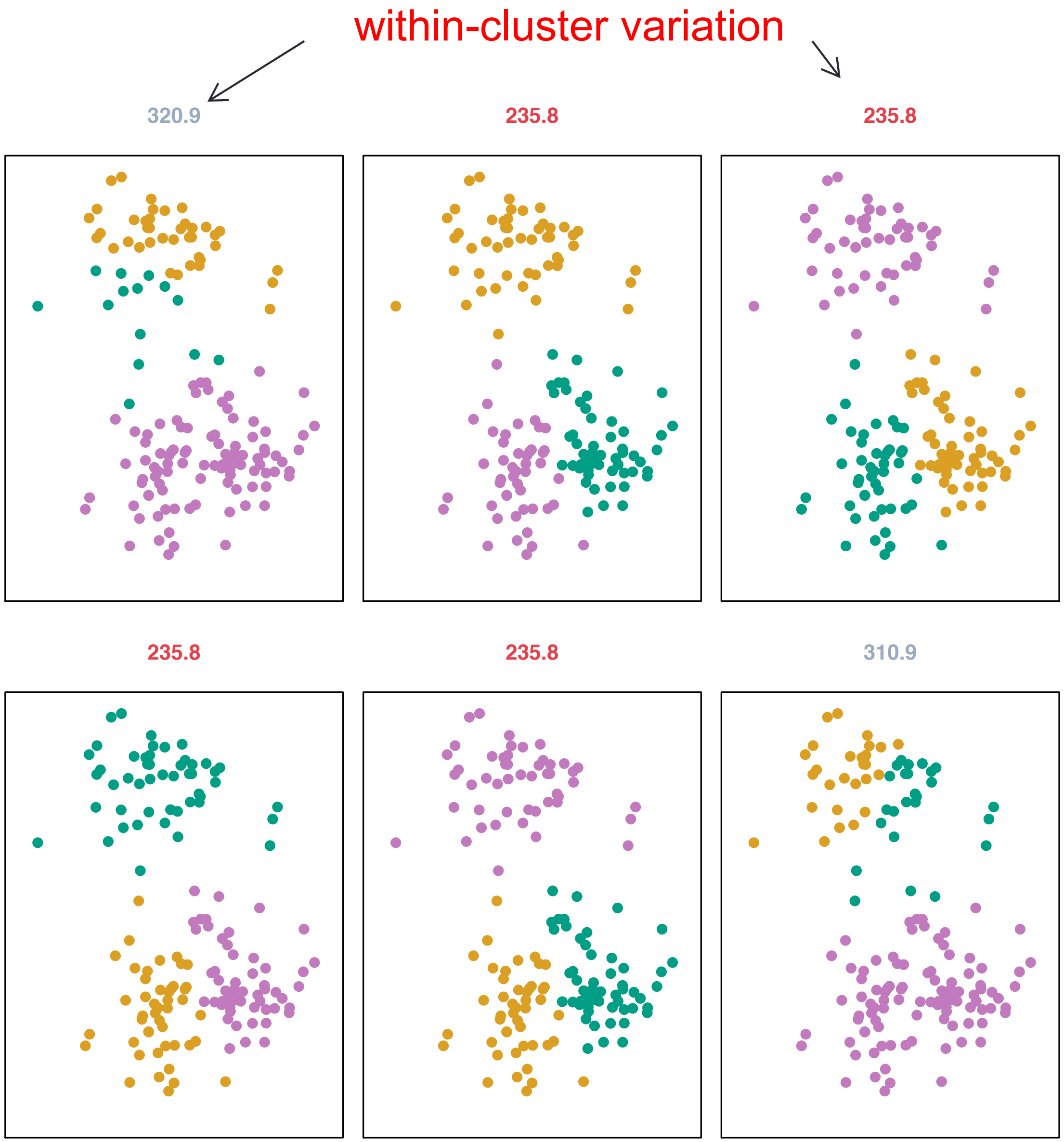
Find new cluster centroids

Reassign points to the closest centroid

if there is no further change, stop

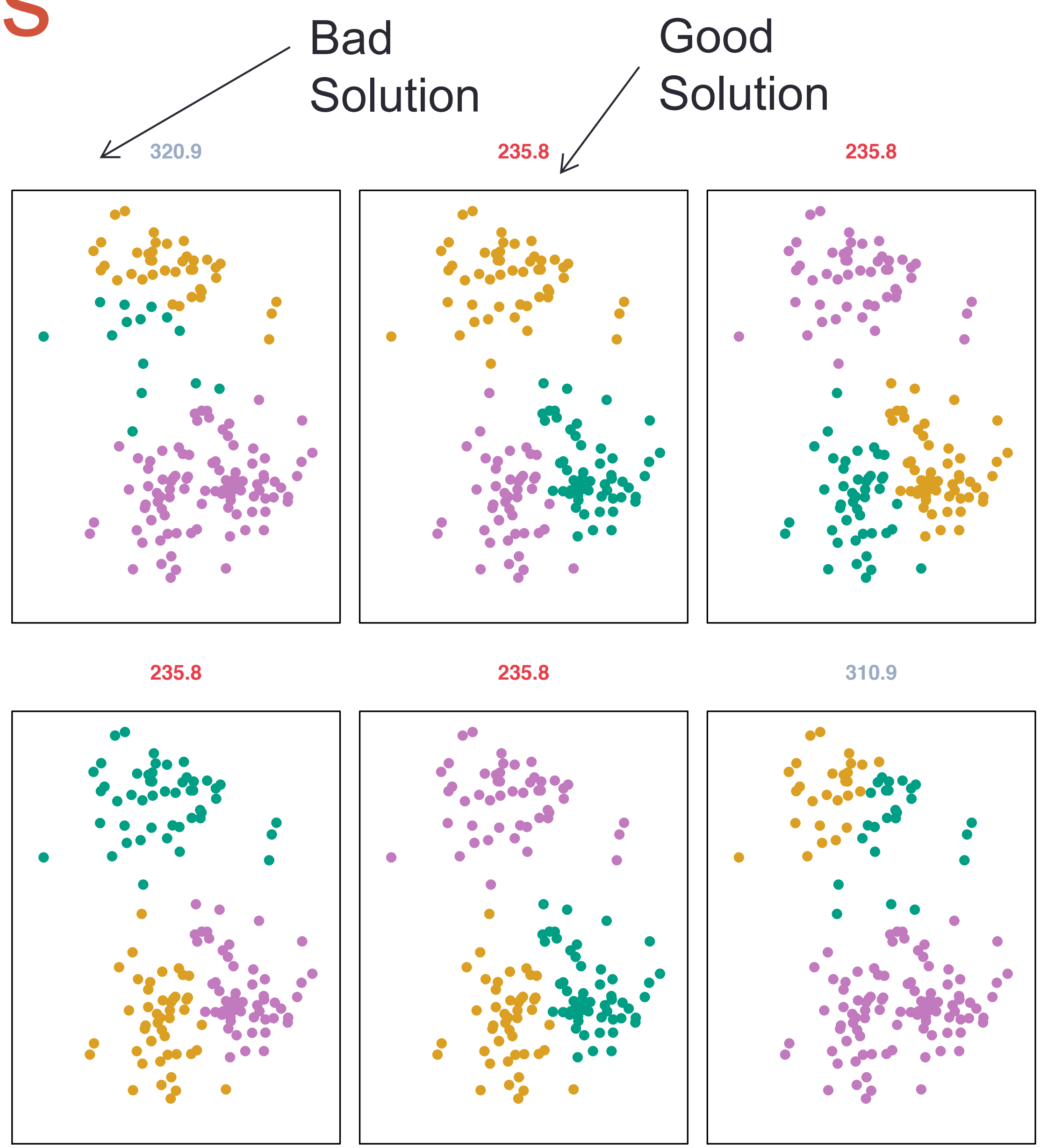
Local Optimums

- K-means always find a solution that depends on the initial assignment
- We must run the algorithm with many different initial assignments
- Select the solution with the smallest within-cluster variation



Local Optimums

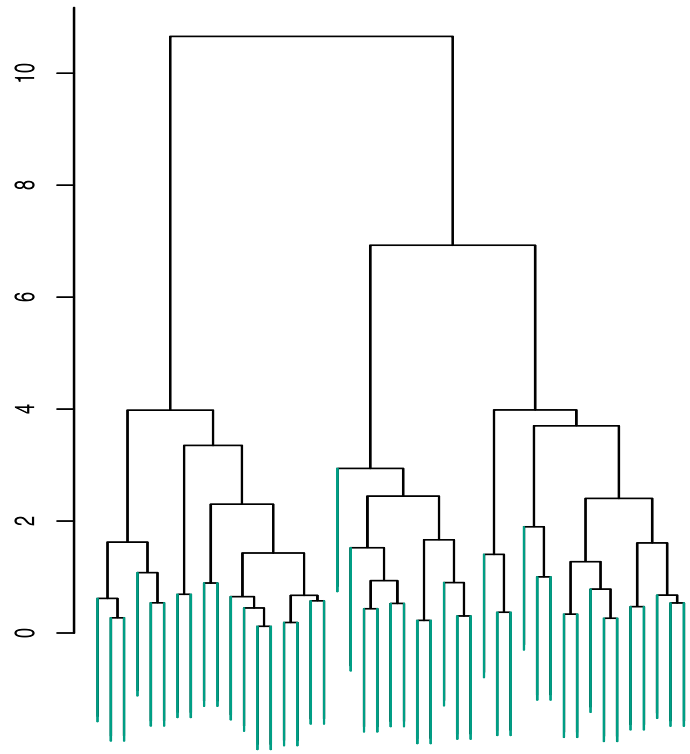
- K-means always find a solution that depends on the initial assignment
- We must run the algorithm with many different initial assignments
- Select the solution with the smallest within-cluster variation



HIERARCHICAL CLUSTERING

Hierarchical Clustering

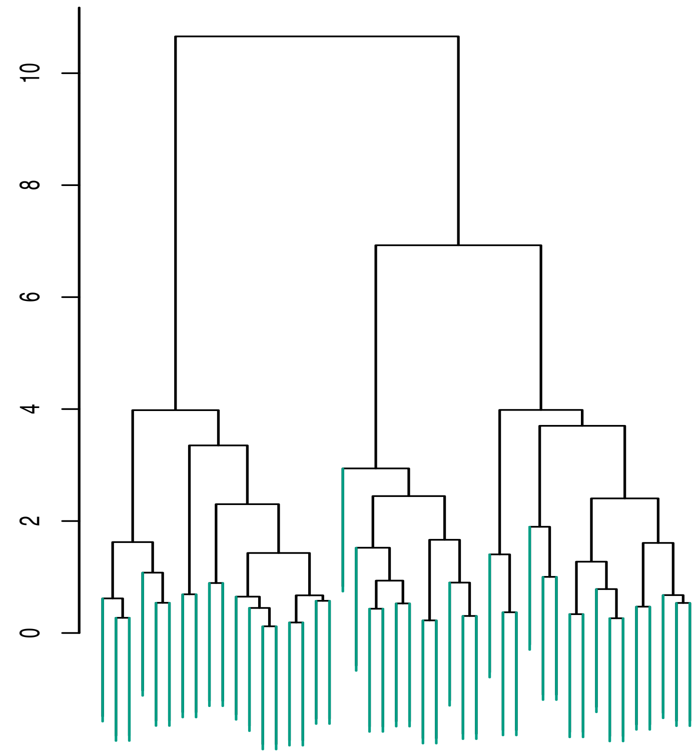
- It results in a tree plot
- Observations are shown as the leaves (bottom)
- As we move up they are combined into clusters
- Based on a distance measure (dissimilarity)



dendrogram

Hierarchical Clustering

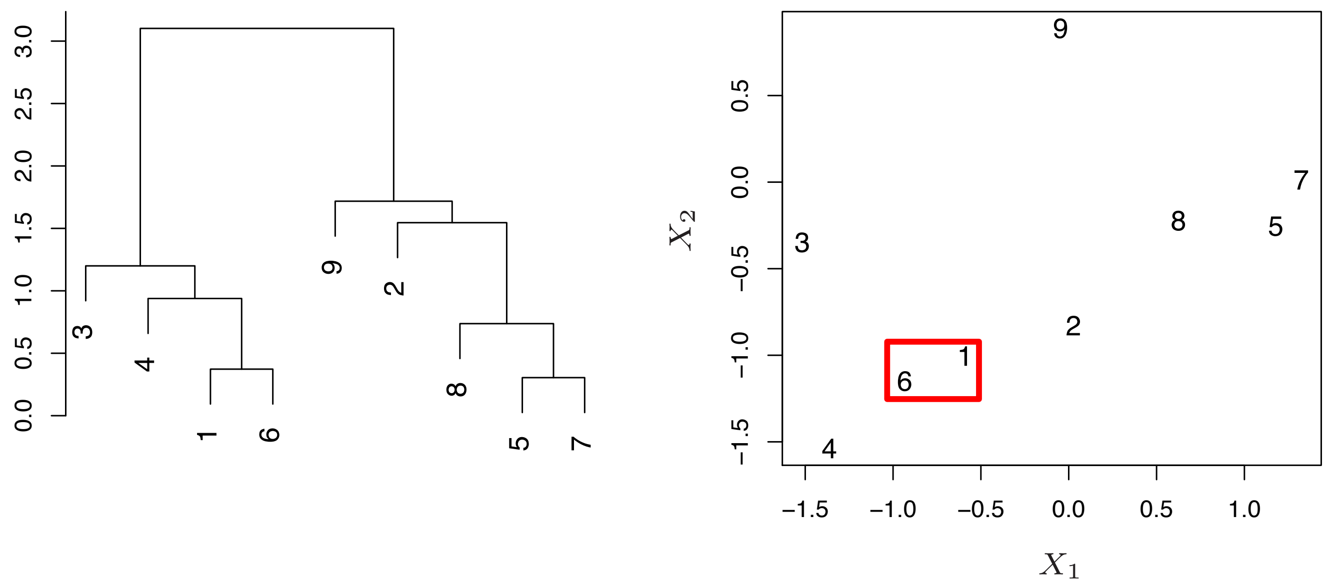
- No need to fix the number of clusters in advance
- Categorical features must be converted to numeric



dendrogram

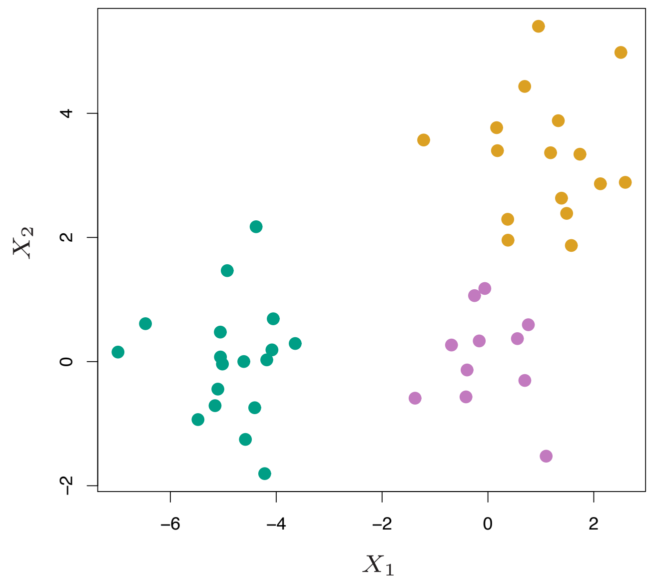
Dendrogram

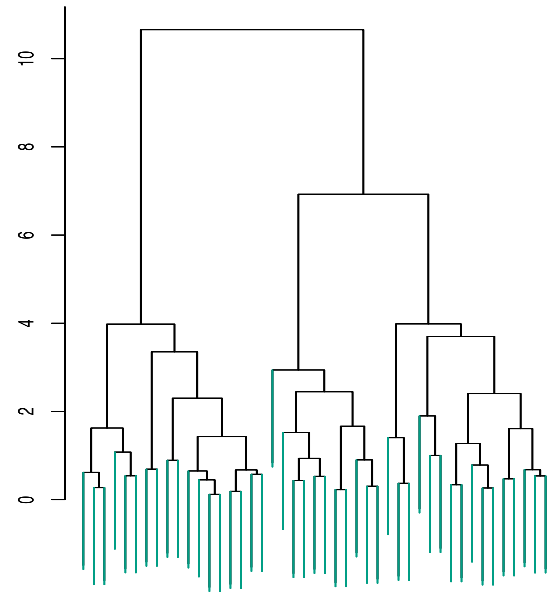
- Vertical axis shows the distance separating observations/clusters
- It indicates how dissimilar the points are
- 1 and 6 dissimilarity is small (~0.4) since they are close
- After the points are fused they are treated as a single observation and the algorithm continues



Dendrogram

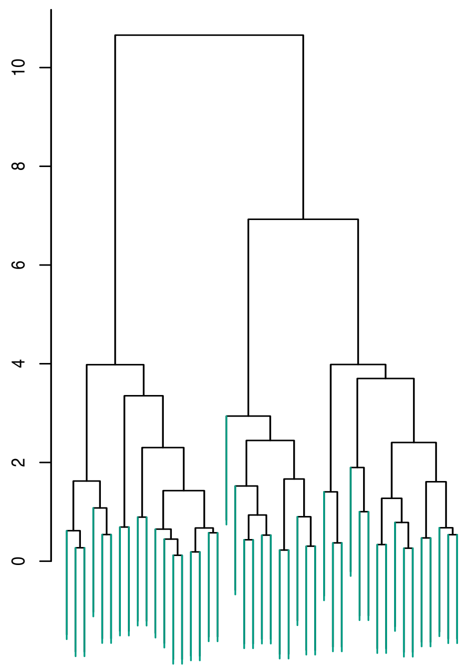
- At the bottom, each “leaf” of the dendrogram represents one of the 45 observations
- As we move up the tree, some leaves begin to fuse.
- These are observations that are similar to each other.
- As we move higher up the tree, an increasing number of observations have fused.
- Observations that fuse later are less similar



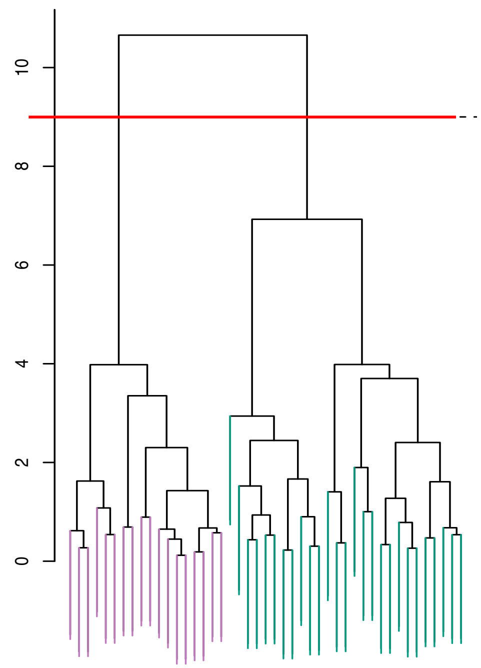


Choosing Clusters

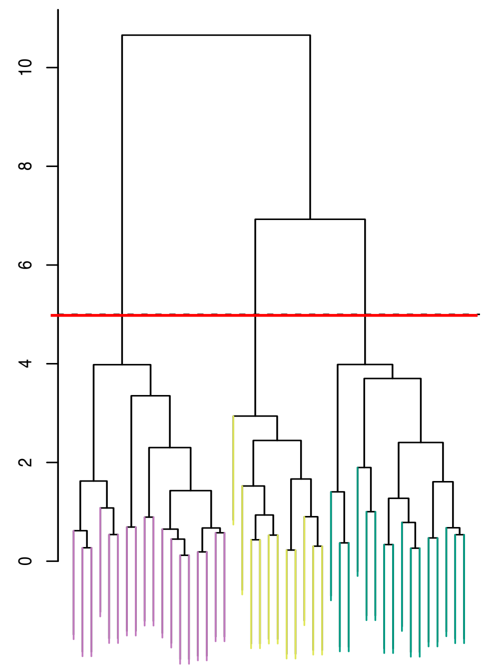
Cut the dendrogram to choose the number of clusters



One Cluster



Two Clusters



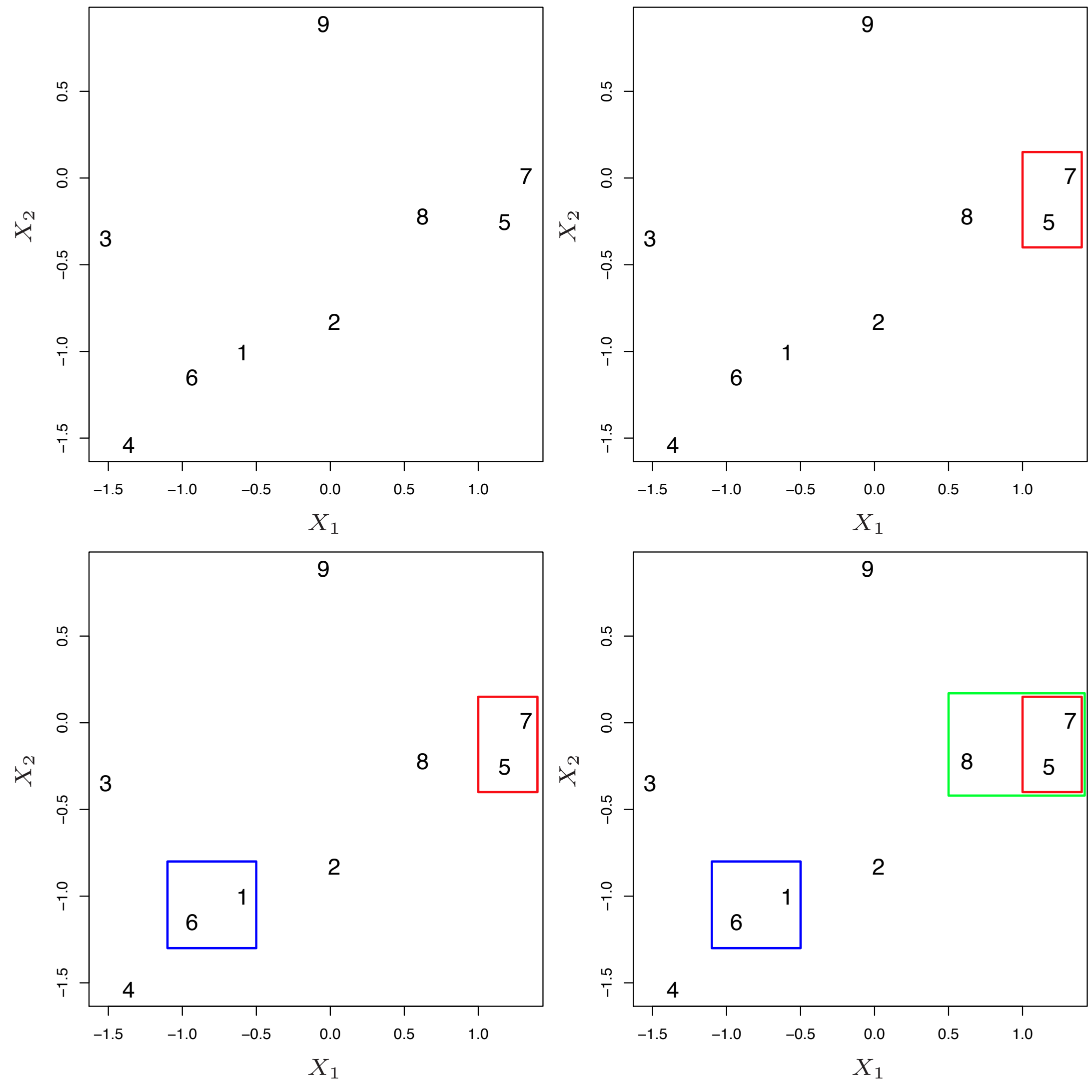
Three Clusters

Algorithm (Agglomerative Approach)

- Start with each point as a separate cluster (*n* clusters)
- Calculate the distance (or dissimilarity) between all points/clusters
- Fuse the two clusters that are most similar so that there are now *n-1* clusters
- Fuse next two most similar clusters so there are now *n*-2 clusters
- Continue until there is only 1 cluster

Example

- Start with 9 clusters
- Fuse 5 and 7
- Fuse 6 and 1
- Fuse the (5,7) cluster with 8
- Continue until all observations are fused



How is dissimilarity defined?

- Implementing hierarchical clustering requires defining a dissimilarity measure
- Also called *linkage*
- How do we define the dissimilarity, or linkage, between two clusters?
- There are four options:

  Complete Linkage

  Single Linkage

  Average Linkage

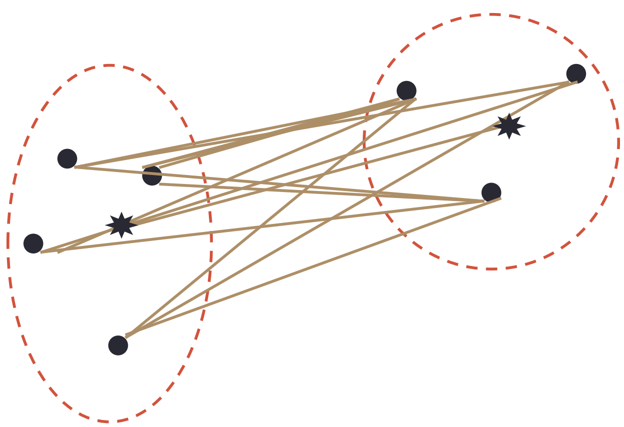
  Centriod Linkage

Distance Between Clusters

There are many possible distances between two clusters.

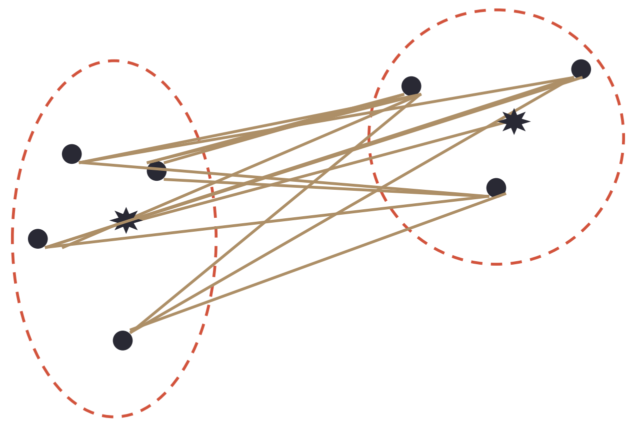
The largest, smallest, or average distance can be used as a dissimilarity measure

The centroid of each cluster can also be found. Then the distance between these two is a measure of dissimilarity too



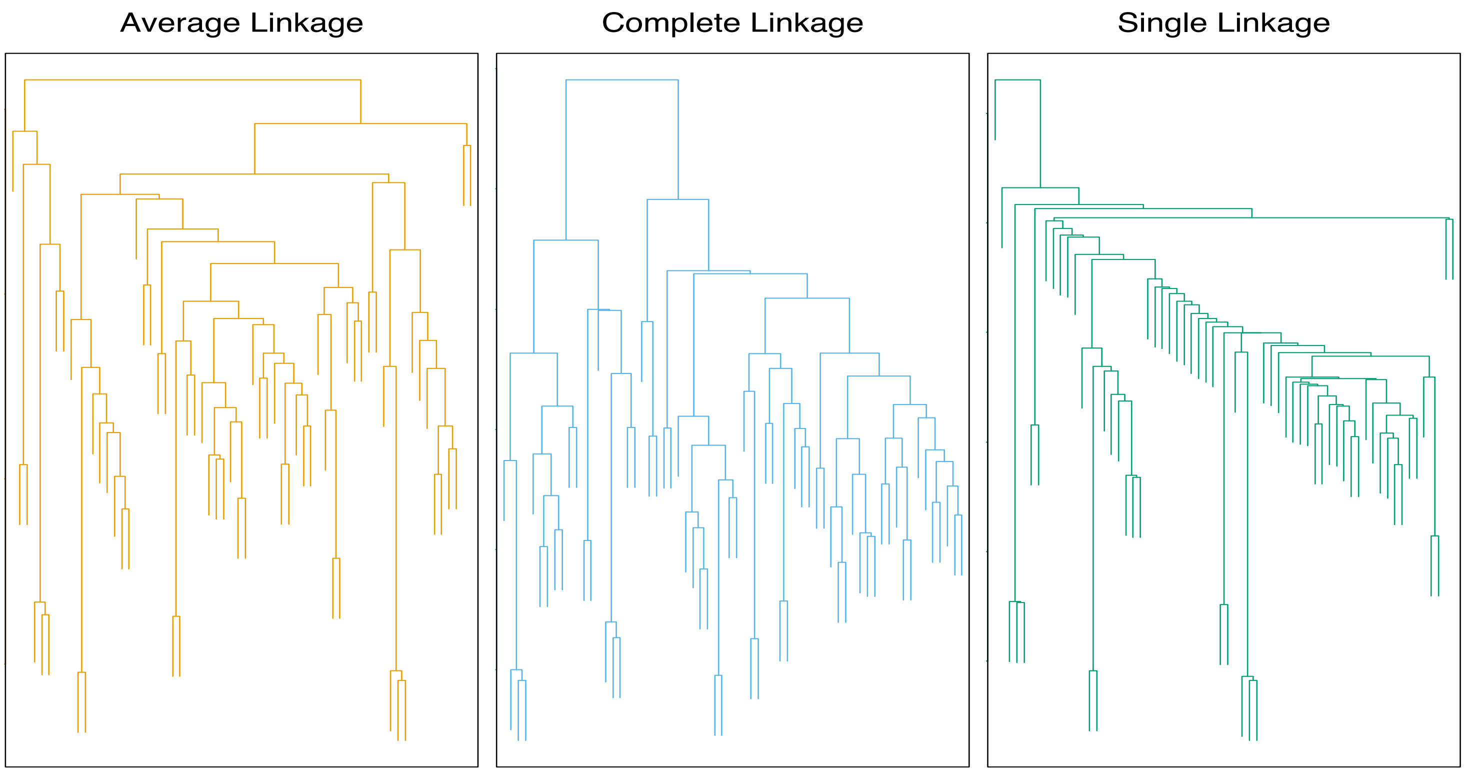
Distance Between Clusters

- Complete Linkage: Largest distance between observations
- Single Linkage: Smallest distance between observations
- Average Linkage: Average distance between observations
- Centroid: Distance between the two centroids



Linkage Can be Important

- Linkage method may result in very different clusters
- Complete and average linkage tend to yield evenly sized clusters
- Single linkage tends to yield extended clusters to which single leaves are fused one by one



Example – complete linkage

- Dataset with 5 observations
- Distance matrix is shown
- Merge obs 3 and 5 into cluster (35)
- Find distances from (35) to obs 1, 2, and 4

$$
\begin{array}{c}
\\ 1 \\ 2 \\ 3 \\ 4 \\ 5
\end{array}
\begin{array}{c}
\begin{array}{ccccc} \;1 & \;\;2 & \;\;3 & \;4 & \;5 \end{array} \\
\begin{bmatrix}
0 & & & & \\
9 & 0 & & & \\
3 & 7 & 0 & & \\
6 & 5 & 9 & 0 & \\
11 & 10 & 2 & 8 & 0
\end{bmatrix}
\end{array}
$$

Example – complete linkage

- Dataset with 5 observations
- Distance matrix is shown
- Merge obs 3 and 5 into cluster (35)
- Find distances from (35) to obs 1, 2, and 4

$$
\begin{array}{c c}
 & \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\
\begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} &
\begin{bmatrix}
0 & & & & \\
9 & 0 & & & \\
3 & 7 & 0 & & \\
6 & 5 & 9 & 0 & \\
11 & 10 & 2 & 8 & 0
\end{bmatrix}
\end{array}
$$

For complete linkage use max{}

$$d_{(35)1} = \max\{d_{31}, d_{51}\} = \max\{3, 11\} = 11$$

$$d_{(35)2} = \max\{d_{32}, d_{52}\} = 10$$

$$d_{(35)4} = \max\{d_{34}, d_{54}\} = 9$$

Example – complete linkage

- Dataset with 5 observations
- Distance matrix is shown
- Merge obs 3 and 5 into cluster (35)
- Find distances from (35) to obs 1, 2, and 4

$$
\begin{array}{c|c}
 & \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\
\begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} &
\begin{bmatrix}
0 & & & & \\
9 & 0 & & & \\
3 & 7 & 0 & & \\
6 & 5 & 9 & 0 & \\
11 & 10 & 2 & 8 & 0
\end{bmatrix}
\end{array}
$$

$$
\begin{aligned}
d_{(35)1} &= \max\{d_{31}, d_{51}\} = \max\{3, 11\} = \boxed{11} \\
d_{(35)2} &= \max\{d_{32}, d_{52}\} = \boxed{10} \\
d_{(35)4} &= \max\{d_{34}, d_{54}\} = \boxed{9}
\end{aligned}
$$

$$
\begin{array}{c|c}
 & \begin{array}{cccc} (35) & 1 & 2 & 4 \end{array} \\
\begin{array}{c} (35) \\ 1 \\ 2 \\ 4 \end{array} &
\begin{bmatrix}
0 & & & \\
11 & 0 & & \\
10 & 9 & 0 & \\
9 & 6 & 5 & 0
\end{bmatrix}
\end{array}
$$

Example – complete linkage

- Merge obs 2 and 4 into cluster (24)
- Find distance from (24) to cluster (35)

$$
\begin{array}{c}
 \\ (35) \\ 1 \\ 2 \\ 4
\end{array}
\begin{array}{c}
\begin{array}{cccc} (35) & 1 & 2 & 4 \end{array} \\
\begin{bmatrix}
0 & & & \\
11 & 0 & & \\
10 & 9 & 0 & \\
9 & 6 & 5 & 0
\end{bmatrix}
\end{array}
$$

$$d_{(24)(35)} = \max\{d_{2(35)}, d_{4(35)}\} = \max\{10, 9\} = 10$$

Example – complete linkage

- Merge obs 2 and 4 into cluster (24)
- Find distance from (24) to cluster (35)

$$\begin{array}{c} \\ (35) \\ 1 \\ 2 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} (35) & 1 & 2 & 4 \end{array} \\ \begin{bmatrix} 0 & & & \\ 11 & 0 & & \\ 10 & 9 & 0 & \\ 9 & 6 & 5 & 0 \end{bmatrix} \end{array}$$

$$d_{(24)(35)} = \max\{d_{2(35)}, d_{4(35)}\} = \max\{10, 9\} = 10$$

- Find distance from (24) to obs 1

$$\begin{array}{c} \\ (35) \\ 1 \\ 2 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} (35) & 1 & 2 & 4 \end{array} \\ \begin{bmatrix} 0 & & & \\ 11 & 0 & & \\ 10 & \boxed{9} & 0 & \\ 9 & \boxed{6} & 5 & 0 \end{bmatrix} \end{array}$$

$$d_{(24)1} = \max\{d_{21}, d_{41}\}$$

$$= \max\{9, 6\} = 9$$

Example – complete linkage

- Merge (24) with 1

$$d_{(24)1} = \max\{d_{21}, d_{41}\} = 9$$

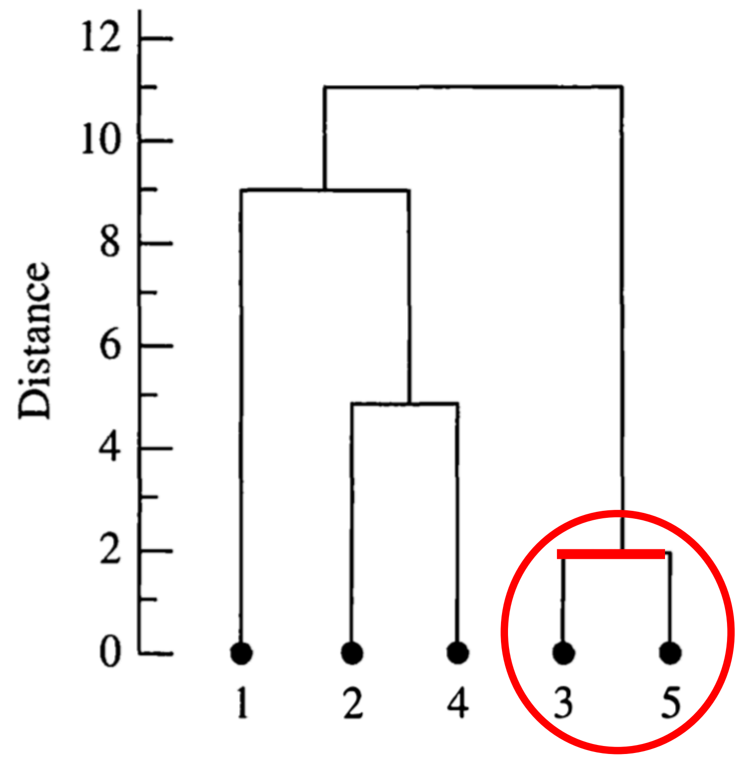
$$\begin{array}{cc} & \begin{array}{ccc} (35) & (24) & 1 \end{array} \\ \begin{array}{c} (35) \\ (24) \\ 1 \end{array} & \begin{bmatrix} 0 & & \\ 10 & 0 & \\ 11 & 9 & 0 \end{bmatrix} \end{array}$$

- Finally merge clusters (124) with (35) into a single cluster (12345) at the distance

$$d_{(124)(35)} = \max\{d_{1(35)}, d_{(24)(35)}\} = \max\{11, 10\} = 11$$

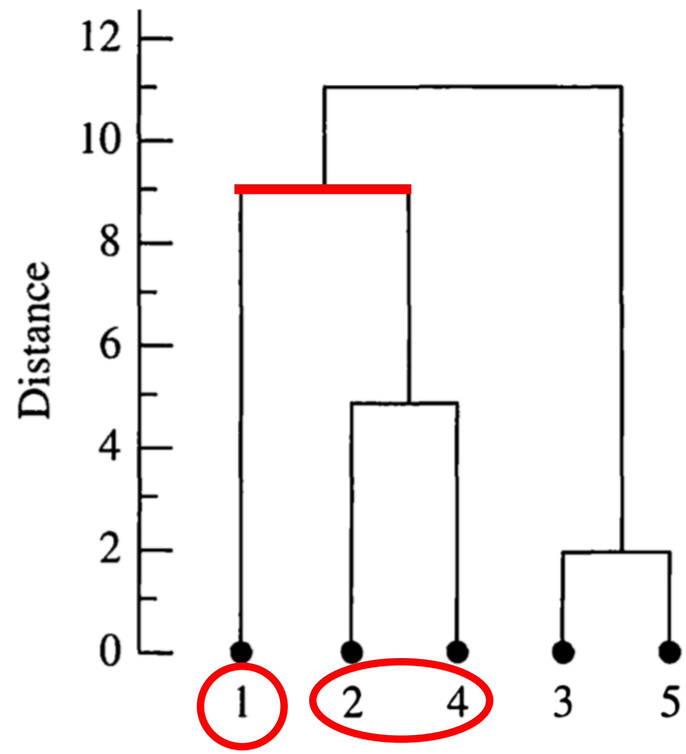
Example – complete linkage

$$
\begin{array}{c c}
 & \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\
\begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} &
\begin{bmatrix}
0 & & & & \\
9 & 0 & & & \\
3 & 7 & 0 & & \\
6 & 5 & 9 & 0 & \\
11 & 10 & 2 & 8 & 0
\end{bmatrix}
\end{array}
$$



Example – complete linkage

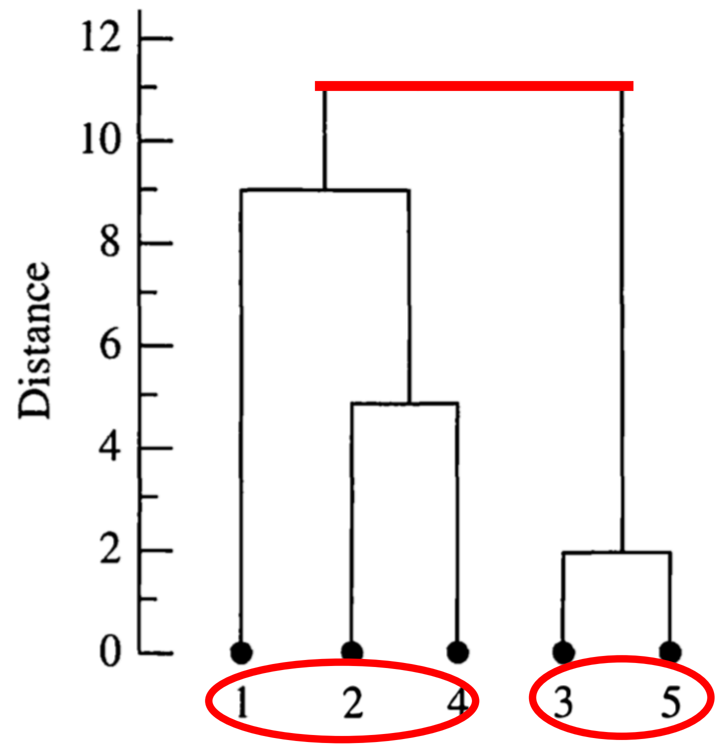
$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & 2 & 8 & 0 \end{bmatrix} \end{array}$$



$$d_{(24)1} = \max\{d_{21}, d_{41}\} = 9$$

Example – complete linkage

$$
\begin{array}{c c}
 & \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\
\begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} &
\begin{bmatrix}
0 & & & & \\
9 & 0 & & & \\
3 & 7 & 0 & & \\
6 & 5 & 9 & 0 & \\
11 & 10 & 2 & 8 & 0
\end{bmatrix}
\end{array}
$$



$$d_{(24)1} = \max\{d_{21}, d_{41}\} = 9$$

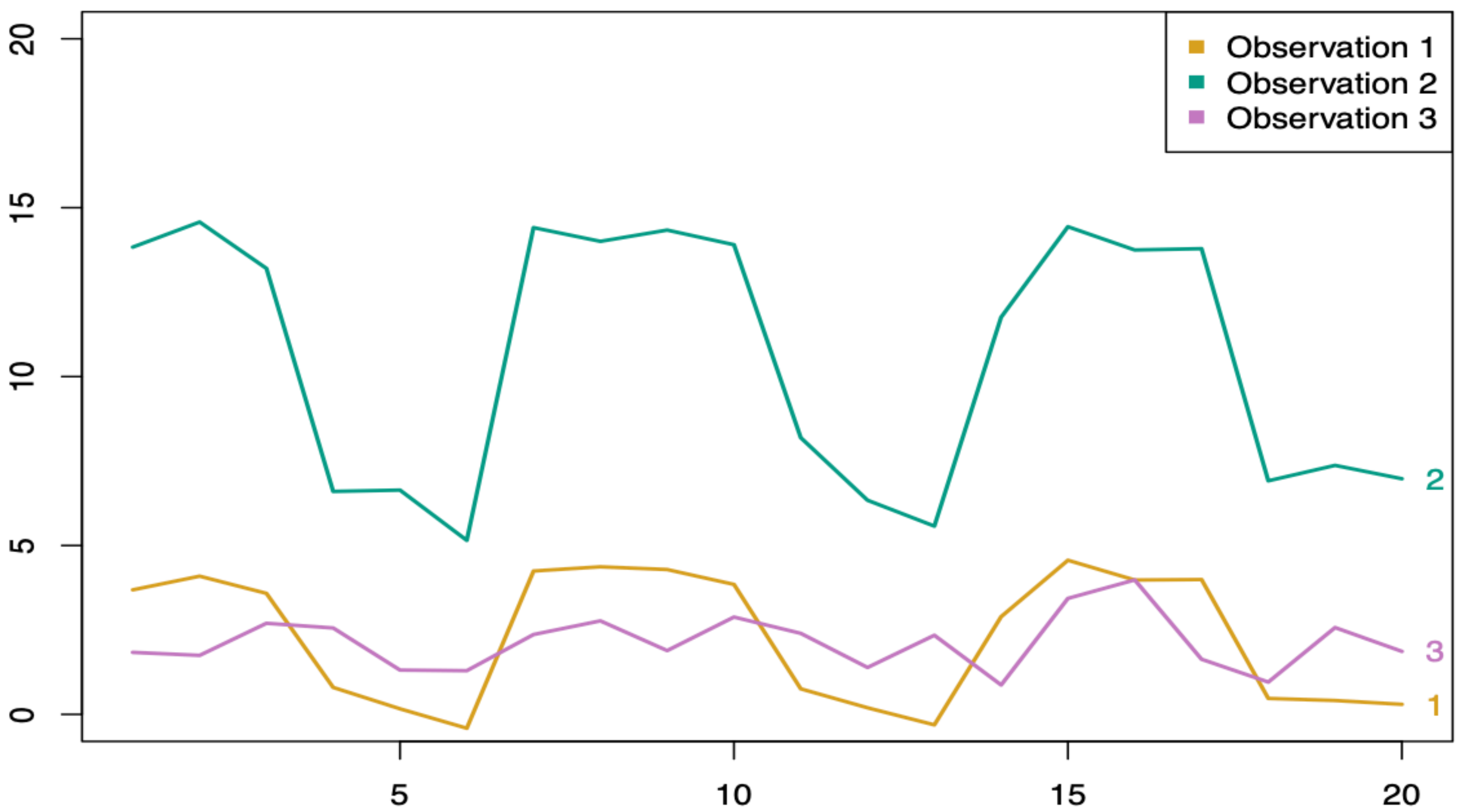
$$d_{(124)(35)} = \max\{d_{1(35)}, d_{(24)(35)}\} = \max\{11, 10\} = 11$$

Choice of Dissimilarity Measure

- So far, we have considered using *Euclidean* distance as the dissimilarity measure
- An alternative measure that could make sense in some cases is the *correlation-based* distance

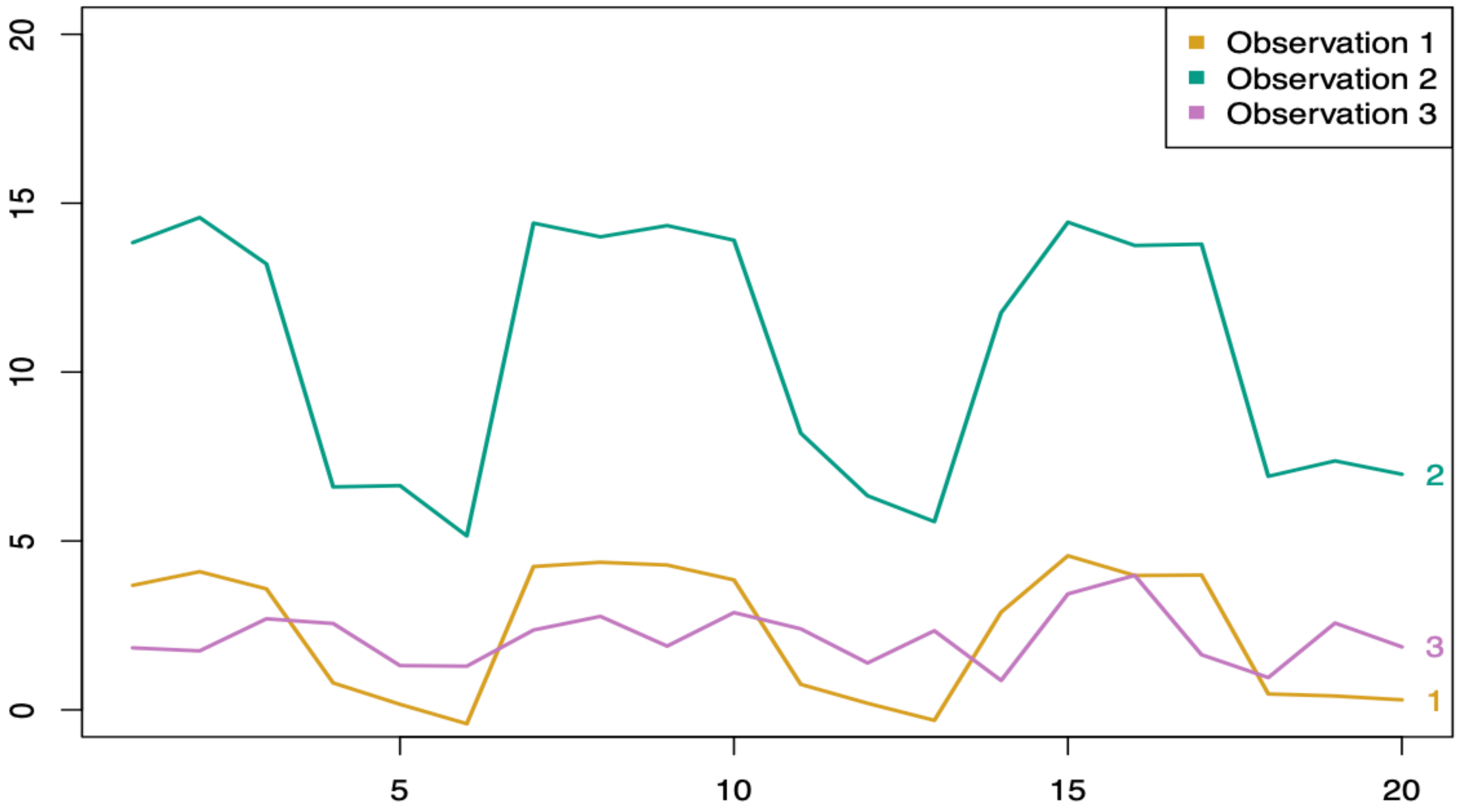
Euclidian vs Correlation-based distance

- Consider 3 observations with p = 20 features each
- Observations 1 and 3 have similar values for each feature, therefore there is a small distance between them



Euclidian vs Correlation-based distance

- Observations 1 and 3 are weakly correlated, therefore they should have a large correlation-based distance
- Observations 1 and 2 are highly correlated, and would be considered similar in terms of correlation measure



Euclidian vs Correlation-based distance

- Suppose we record the number of purchases of each item (columns) for many customers (rows)
- Using Euclidean distance, customers who have purchases of similar dollar amount would be clustered together
- Using correlation measure, customers who tend to purchase the same types of products will be clustered together even if the magnitude of their purchase may be different

Practical Issues in Clustering

- Should the features be scaled?
- Hierarchical clustering
  - What dissimilarity measure?
  - What type of linkage?
  - Where to cut the dendrogram to choose *K*?
- K-means clustering
  - How many clusters?