MSCI152: Introduction to Business Intelligence and Analytics

Lecture 16: Introduction to Descriptive Data Mining

Dr Anna Sroginis

Lancaster University Management School

Agenda

1. Recap
2. Introduction to data mining
3. Cluster Analysis
4. k-Means clustering
5. Hierarchical clustering

More details can be found

- Camm et al. Chapters 5.1, 5.2
- James et al. An Introduction to Statistical Learning Chapters 1 and 12 (you can download a free copy!)

Agenda

1. Recap
2. Introduction to data mining
3. Cluster Analysis
4. k-Means clustering
5. Hierarchical clustering

What do you know about **data mining**?

Question: What is Data Mining? In a single word/phrase

Vote here:

or

www.wooclap.com/SPPKAY

Unsupervised vs supervised statistical learning

Statistical Learning

refers to a set of tools for making sense of complex datasets (understanding data).

Supervised SL

involves building a statistical model for predicting, or estimating, an output based on one or more inputs.

Unsupervised SL

is the situation where there are inputs but no supervising output. There is no outcome/response variable to predict that can "supervise" our analysis.

Unsupervised vs supervised statistical learning

Unsupervised statistical learning

- The goal of unsupervised learning techniques is to use the variable values to identify relationships between observations.
- It is high-dimensional descriptive analytics, designed to describe patterns and relationships in large data sets with many observations of many variables.
- But without explicit outcome/response variable!

Unsupervised Statistical Learning = Descriptive Data Mining

Agenda

Clustering Problem

Clustering

Given a set of features $X_1, X_2, .., X_p$ measured on n observations, **the goal is to divide these objects into groups ("clusters")** such that objects within a group tend to be more similar to one another as compared to objects belonging to different groups.

- Is there an informative way to visualize the data?
- Can we discover subgroups among the variables or among the observations?

Clustering Problem

- **Task**: Can we discover subgroups among the variables or among the observations?
- **Difference to Classification**: Class labels are unknown $\Rightarrow$ Similarity depends on application
- **No unique definition of a cluster**
- A part of an **exploratory data analysis** (to explore and characterise data before supervised modelling)
- Quite **subjective** and depends on the application: there is no way to check our work because we don't know the true answer!

Clustering problem: examples

1. A cancer research: investigating groups and subgroups among the breast cancer samples;
2. An online shopping site: identifying groups of shoppers with similar browsing and purchase histories (as in **market segmentation**);

Basic steps

1. **Feature Selection**: Features selected to encode as much information as possible concerning task
2. **Proximity measure**: Quantifies how similar or dissimilar feature vectors are
3. **Clustering criterion**: Determines what is a sensible type of cluster for application
4. **Clustering algorithm**: Determined by previous 2 choices
5. **Validation and interpretation of results**

Two main clustering techniques

k-means clustering

a method which iteratively assigns each observation to one of k clusters in an attempt to achieve clusters that contain observations as similar to each other as possible

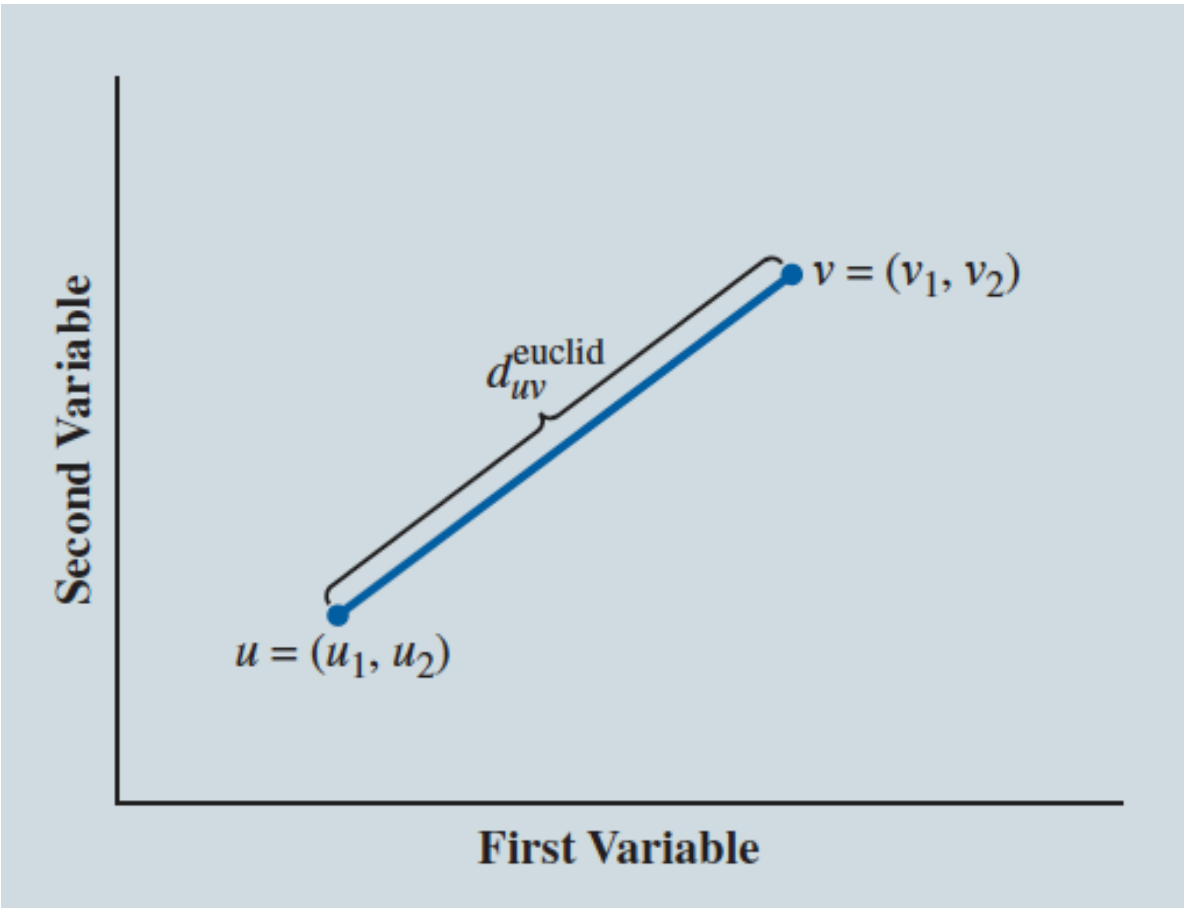
Hierarchical clustering

starts with each observation belonging to its own cluster and then sequentially merges the most similar clusters to create a series of nested clusters.

Measurements of dissimilarity

Euclidean distance is a common method to measure dissimilarity between observations:

$$d = \sqrt{(u_1 - v_1)^2 + (u_2 - v_2)^2} \tag{1}$$



Euclidean distance: example

Let

- $u = (23, \$20,375)$ correspond to a 23-year-old customer with an annual income of \$20,375
- $v = (48, \$19,475)$ correspond to a 48-year-old with an annual income of \$19,475

As measured by Euclidean distance, the dissimilarity between these two observations is

$$d = \sqrt{(23-48)^2 + (20,375-19,475)^2} = \sqrt{625 + 810,000} = 900 \quad (2)$$

In this case, the amount of dissimilarity between observations is dominated by the Income variable because of the difference in scale.

To avoid that, we need to standardize the units $\Rightarrow$ replace the original values by $(x - \bar{x})/s_x$

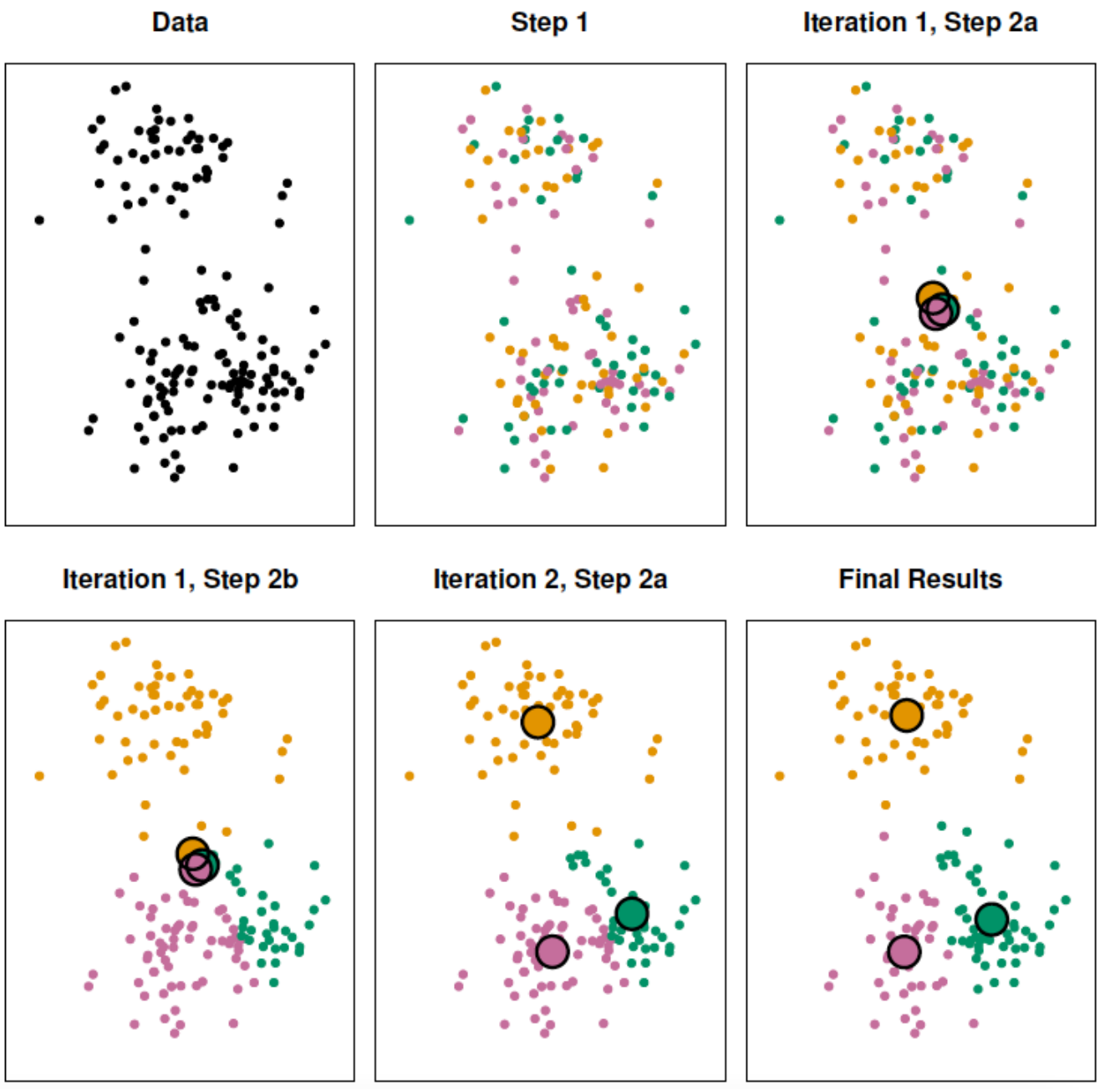
Agenda

k-Means clustering

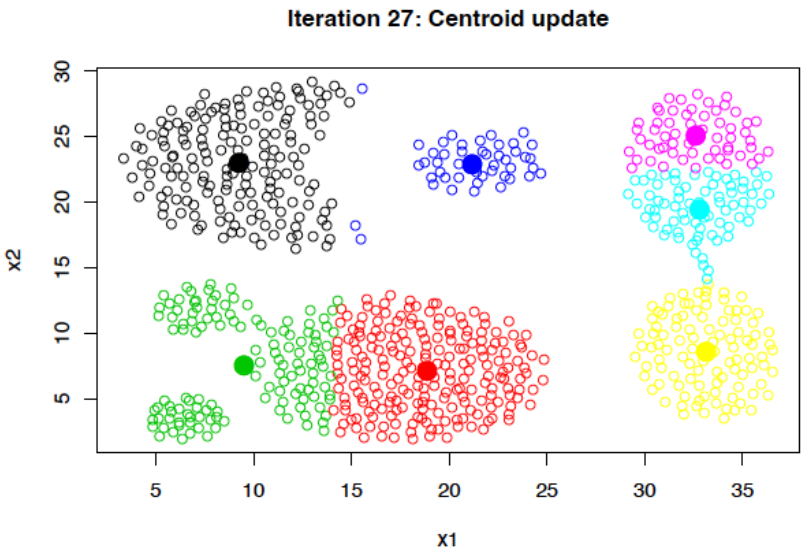
Several steps:

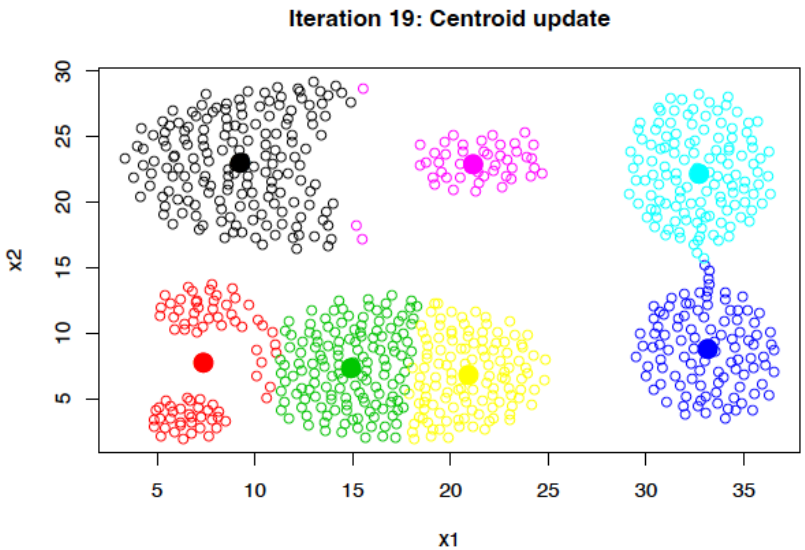
1. Specify the number of clusters, k
2. Randomly assign each observation to one of the k clusters
3. After all observations have been assigned to a cluster, the resulting cluster centroids are calculated (these cluster centroids are the "means" of k-means clustering)
4. Using the updated cluster centroids, all observations are reassigned to the cluster with the closest centroid (where Euclidean distance is the standard metric).
5. Repeat this process (calculate cluster centroid, assign each observation to the cluster with nearest centroid)
    - until there is no change in the clusters or a specified maximum number of iterations is reached

k-Means clustering



k-Means clustering: limitations





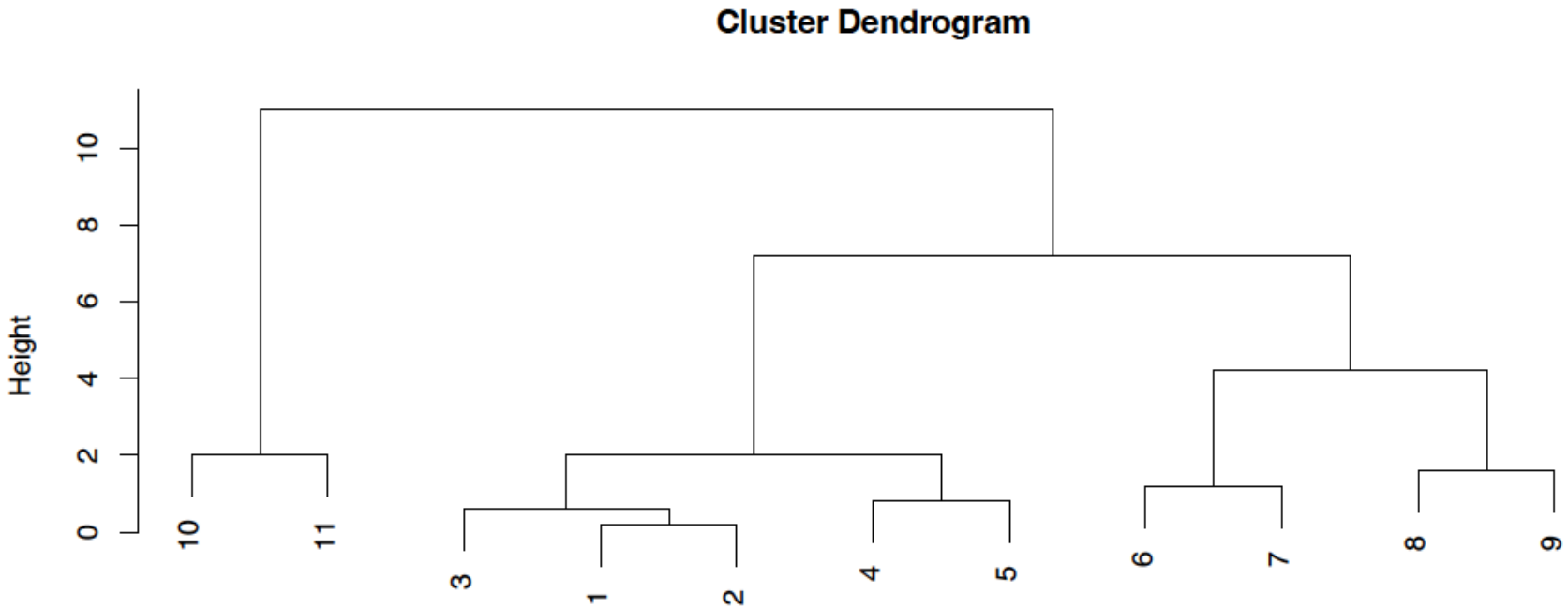
- Sensitive to initialisation of centroids
- k-means designed to identify spherical clusters: Elongated clusters, and clusters of different size cause problems

Agenda

1. Recap
2. Introduction to data mining
3. Cluster Analysis
4. k-Means clustering
5. Hierarchical clustering

Hierarchical clustering

- Pair of clusters with lowest dissimilarity merged recursively
- Height of connecting lines reflects dissimilarity
- **Nested structure**: Each cluster at a lower level is a subset of a cluster at a higher level
  - Be aware that not all data can be structured this way!

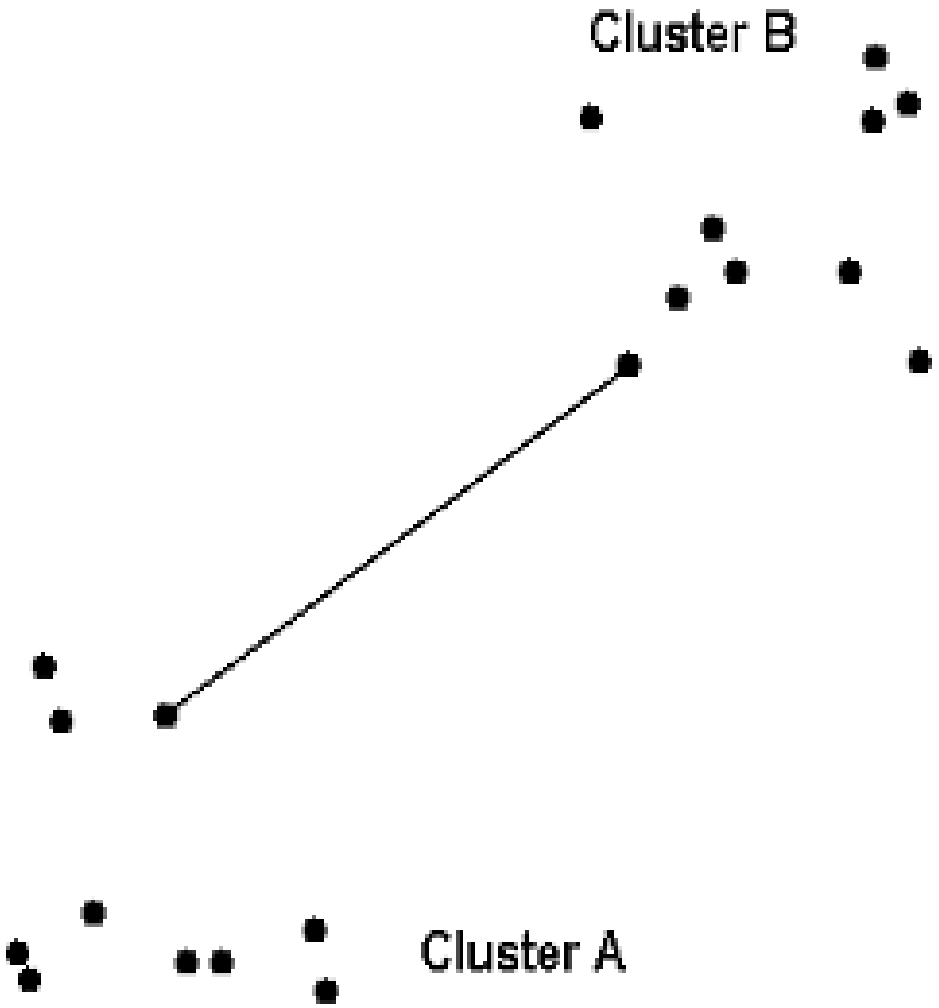


Hierarchical clustering

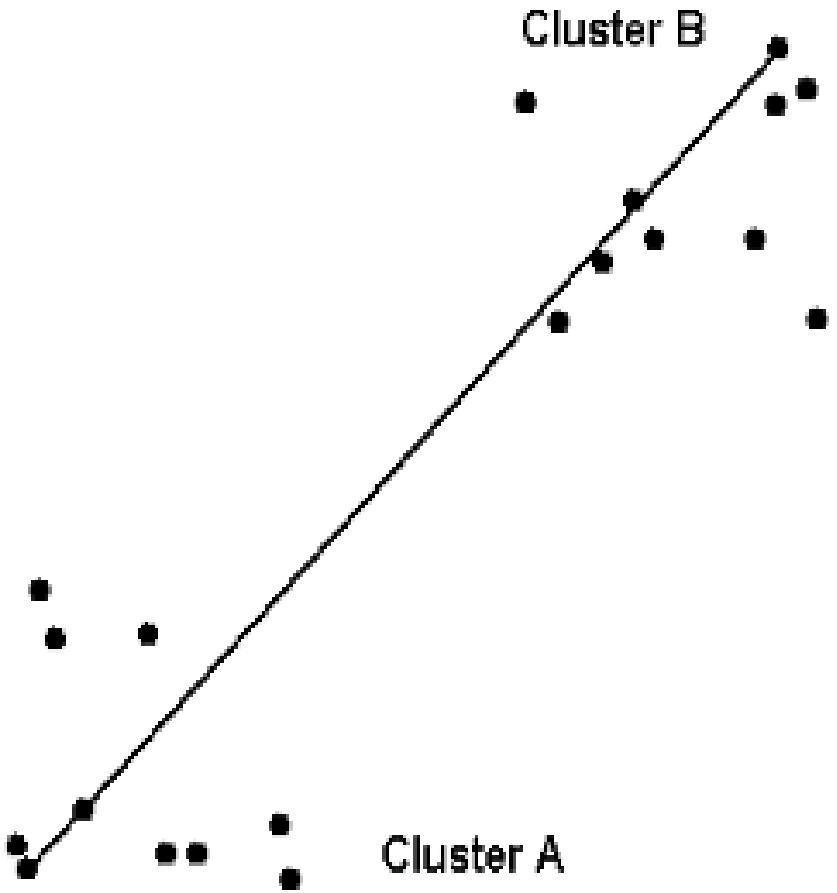
Several steps:

1. At the beginning each respondent is their own cluster;
2. Measure distances between all pairs of respondents;
3. Construct a distance (or dissimilarity) matrix;
4. Join two closest objects, either by forming new group or joining to the old one;
5. Recalculate dissimilarities matrix based on linkage:
   - **Single linkage**: use smallest distance
   - **Complete linkage**: use largest distance
   - **Average linkage**: average distance between all pairs in clusters
   - **Centroid linkage**: distance between cluster centroids (mean)
   - etc...
6. Repeat steps 4 and 5 until all objects are clustered

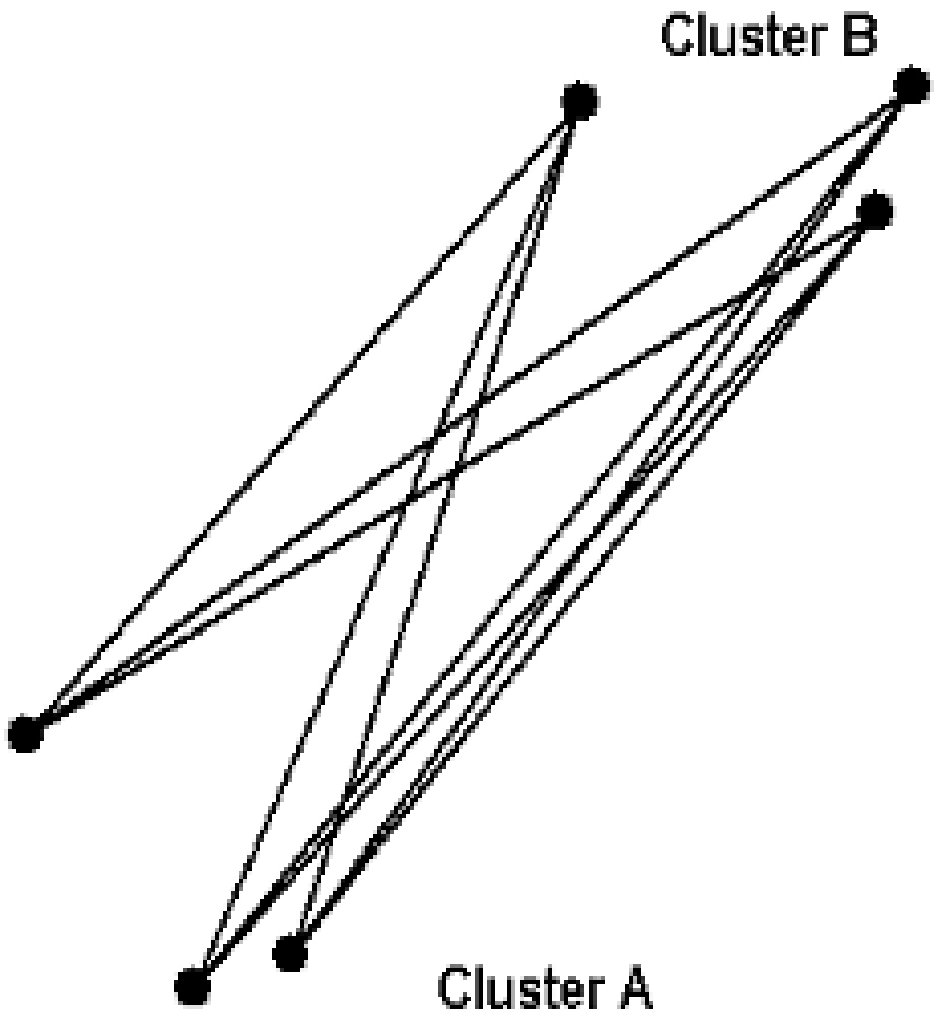
Simple linkage: use smallest distance



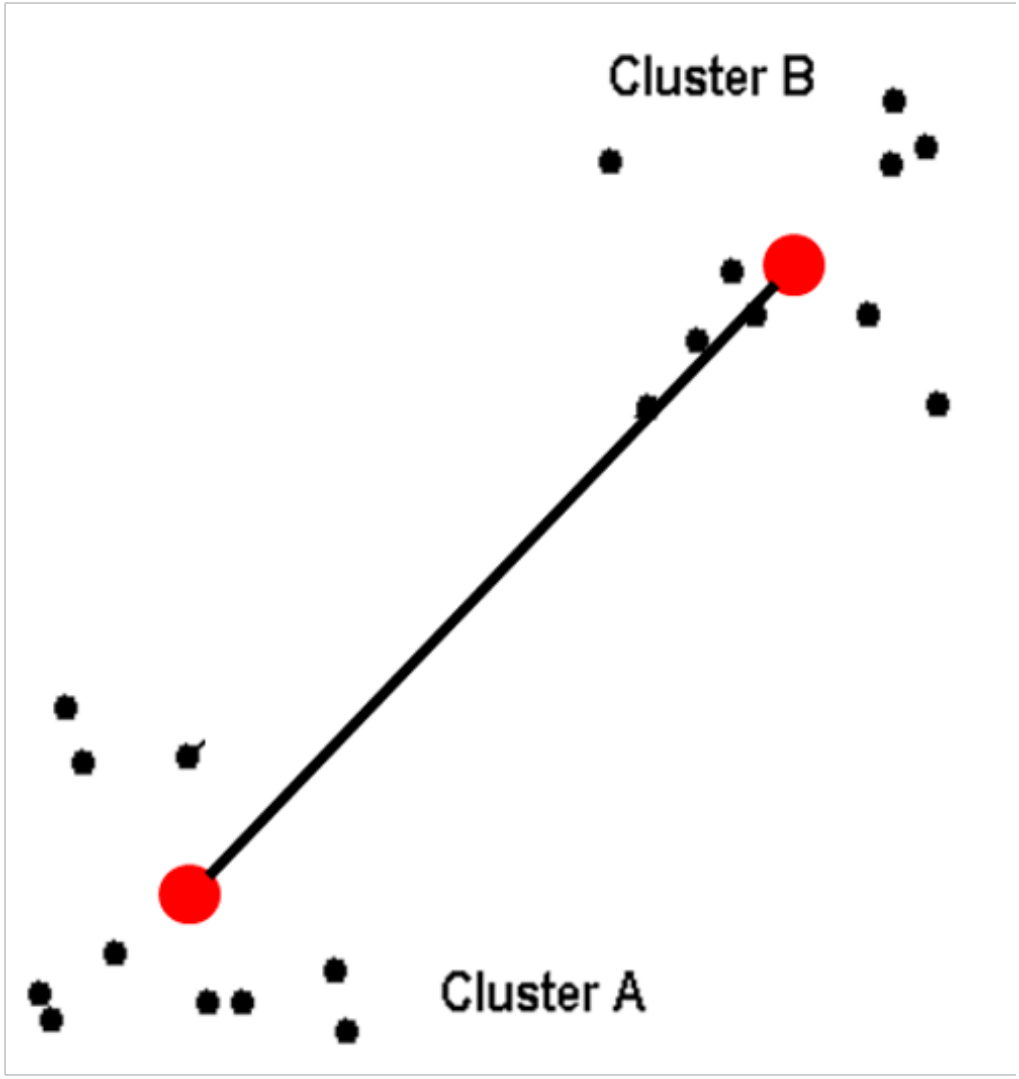
Complete linkage: use largest distance



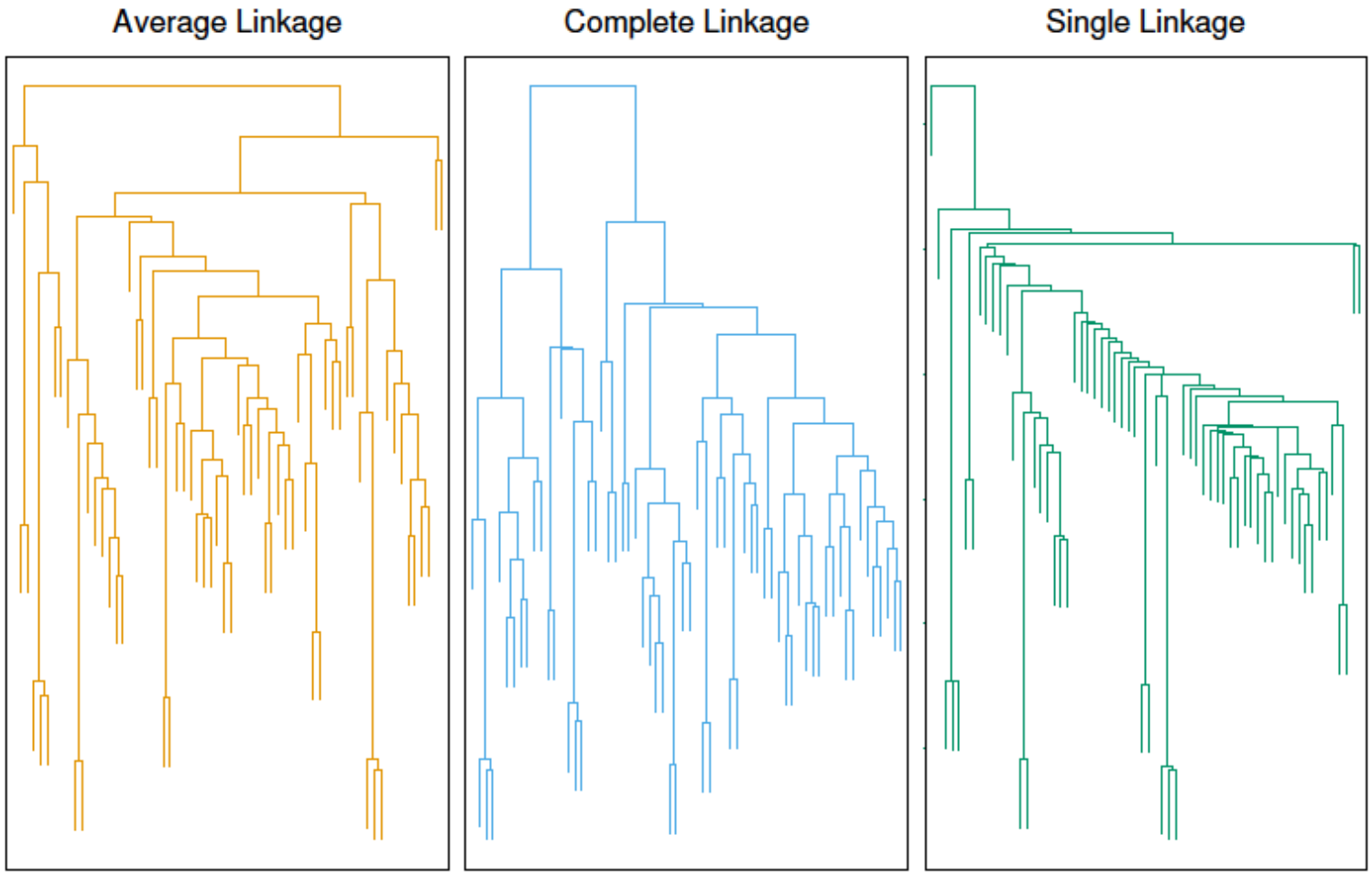
Average linkage: average distance between all pairs in clusters



Centroid linkage: distance between cluster centroids (mean)



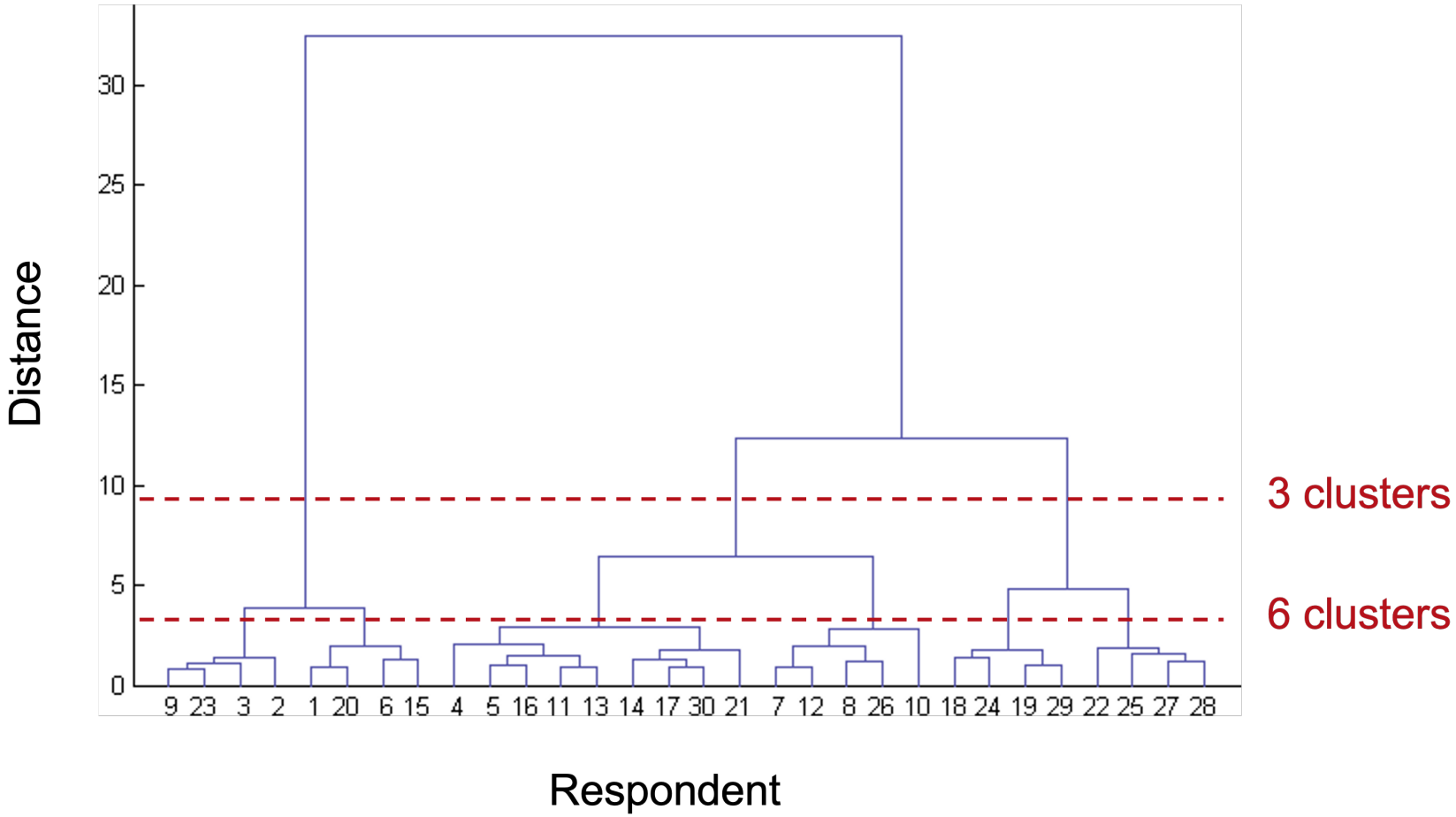
Hierarchical clustering



Average and complete linkage tend to yield more balanced clusters

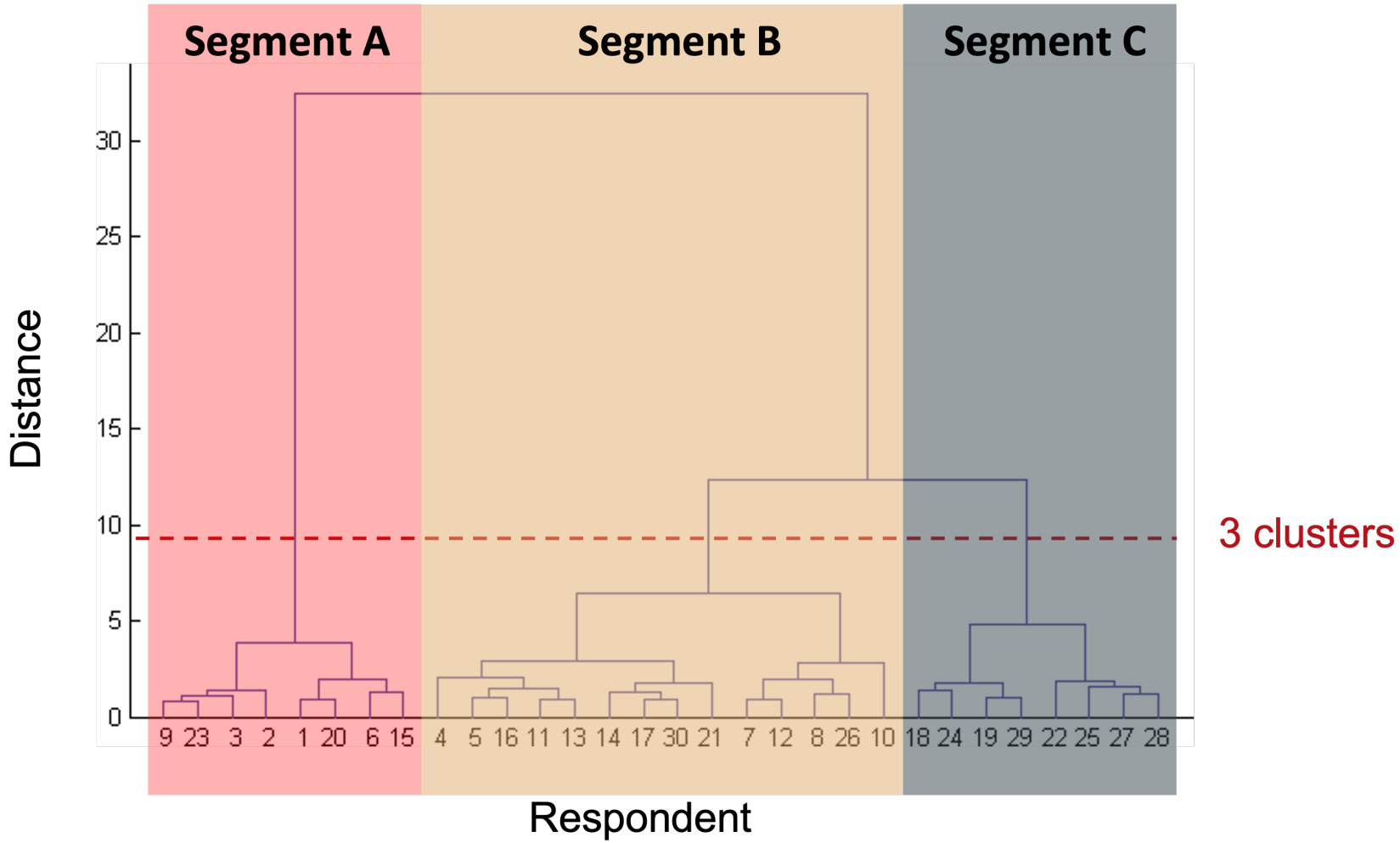
Hierarchical clustering

Construct a dendrogram in order to visualise the result

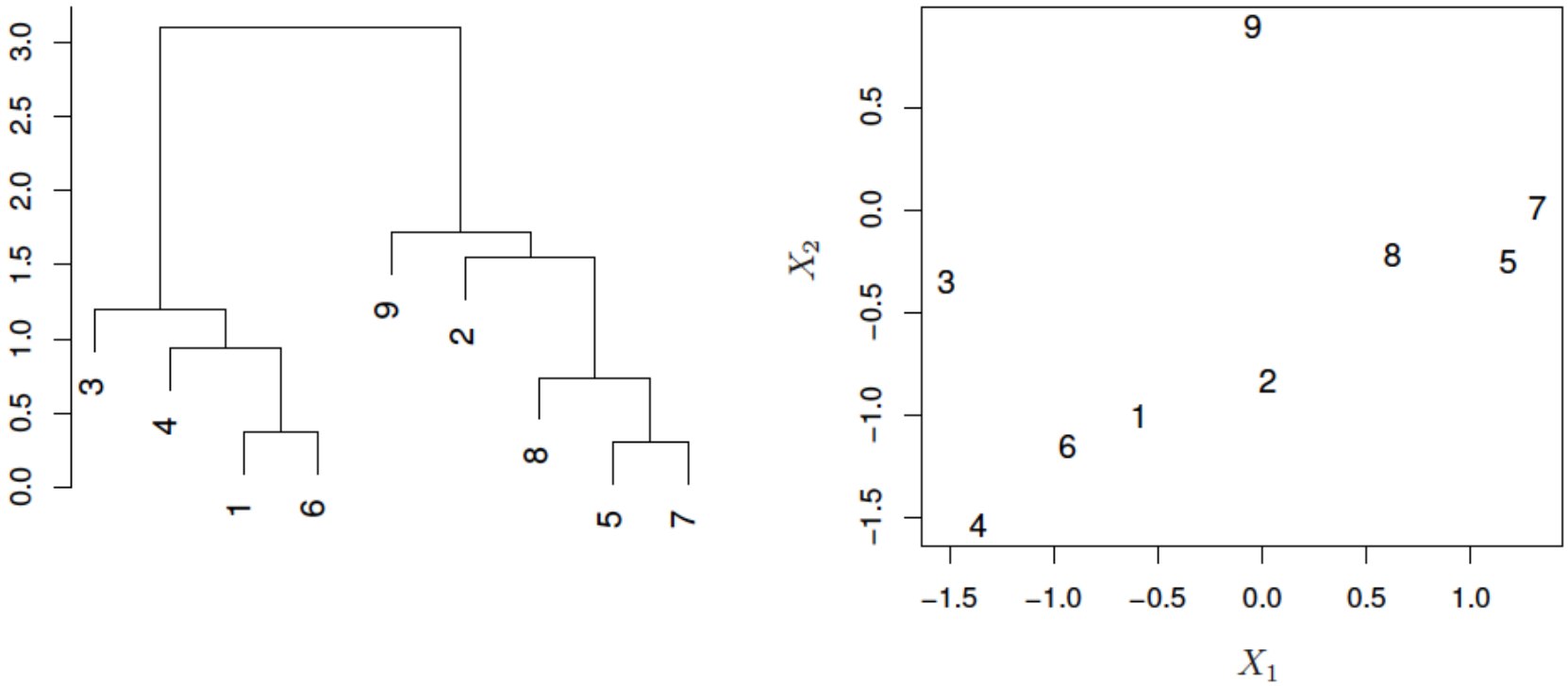


Hierarchical clustering

Construct a dendrogram in order to visualise the result



Hierarchical clustering: example



Interpretation:

- Observations 5 and 7 are quite similar to each other, as are observations 1 and 6.
- Observation 9 is no more similar to observation 2 than it is to observations 8, 5, and 7,
  - even though observations 9 and 2 are close together in terms of horizontal distance

Hierarchical clustering

Advantages:

- It aligns well with the idea of market segmentation
- Very suitable for specific applications: Biology, Medicine, Social sciences, Text mining;
- Segments in segments;
- Does not assume much.

Disadvantages:

- Number of clusters? Open question;
- Might be heavy for big datasets;
- Does not guarantee finding global optimum.

Clustering: practical issues

- Should the observations or features first be standardized in some way?
- 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?
- In the case of K-means clustering, how many clusters should we look for in the data?

With these methods, there is no single right answer!

Wrap up

Today we:

- Covered two main unsupervised statistical learning techniques: k-Means and hierarchical clustering

Next time:

- Report Writing by Chris.

Bonus question: Linear regression is an approach for supervised or unsupervised learning?