Machine Learning for Bioinformatics

Unsupervised learning

Irina Mohorianu (CSCI)

What is Machine Learning?

What do you see?

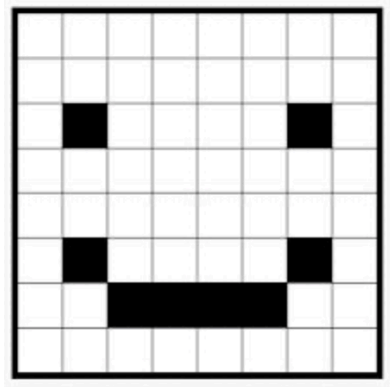
We need to classify the pictures with a dog and the ones with the muffin.

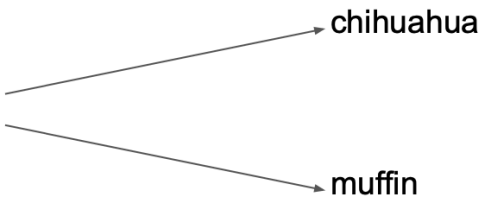
Or do we need to partition the images and manually annotate some of them?

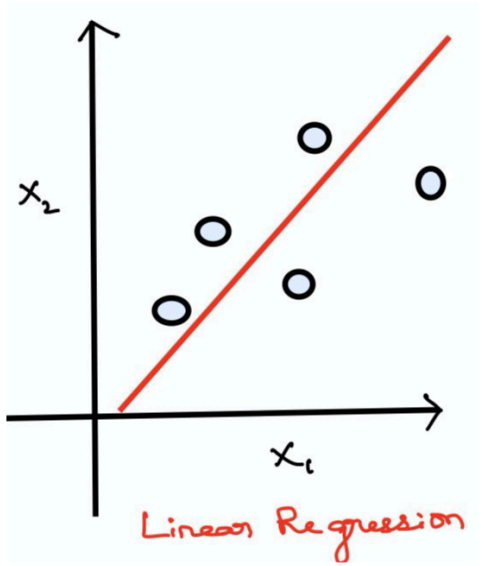
What is Machine Learning?

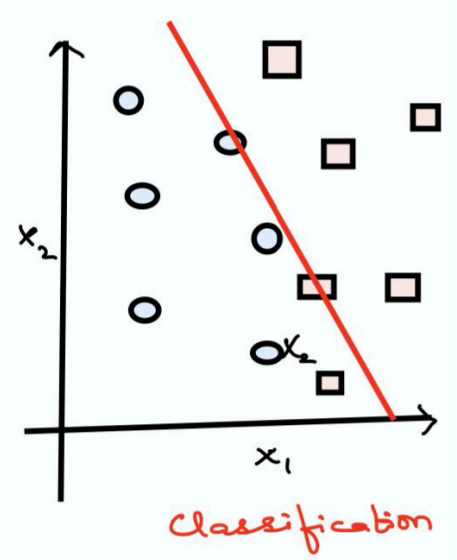
New York Times reported in 1958 that the invention was the beginning of a computer that would “be able to walk, talk, see, write, reproduce itself and be conscious of its existence”

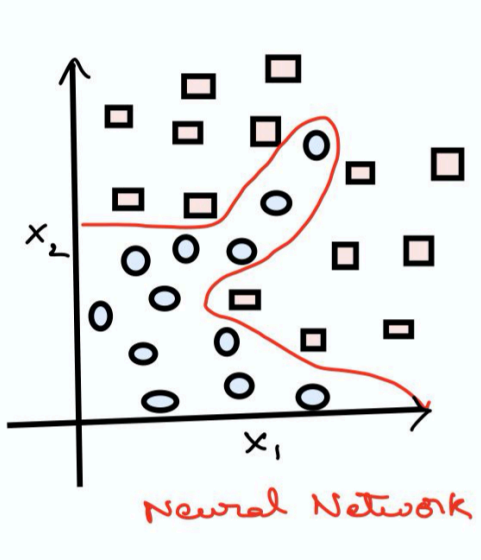
What is Machine Learning? Behind the hood











What is Machine Learning?

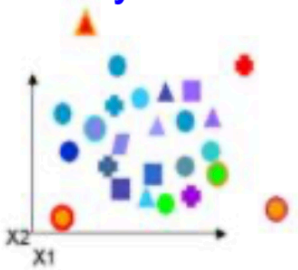
Machine Learning (ML) is generating abstract hypotheses (models) based on data which can be later used predict results on new data.

Types of ML:

[1] Unsupervised – Tue lecture
[2] Supervised – Thu lecture
[3] Semi-supervised
[4] Reinforcement learning

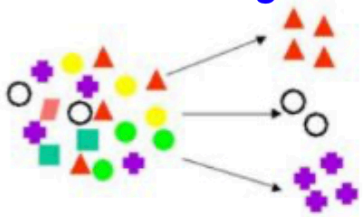
Common models used in Machine Learning

Anomaly detection



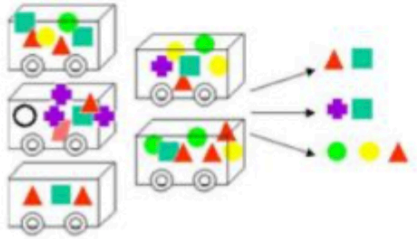
Identification of outliers

Clustering



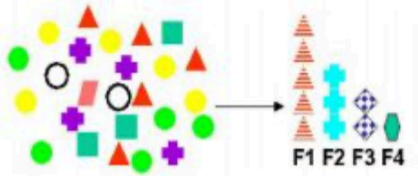
Finding groups/ structure in the data

Association



Finding rules associated with naturally co-occurring terms

Feature selection

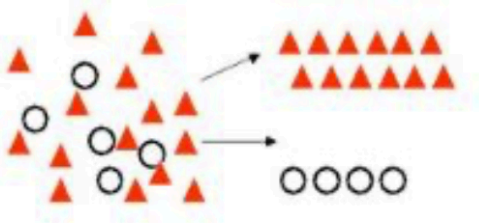


Selection of best features to describe a target attribute

Common models used in Machine Learning

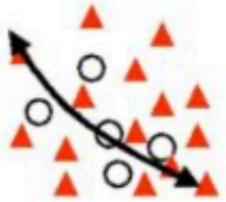
Techniques

Classification



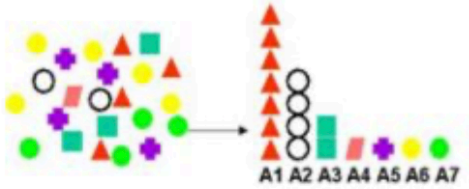
Separation of data point into classes.
Typical example: yes/no answers

Regression



Continuous numerical outcome.
Prediction of future values based on observed ones

Attribute Ranking



Identification of important/ discriminative features based on their relationship with the target attribute.

ML unsupervised. Outline

Dimensionality reduction

- Principal component analysis

Clustering approaches

- Hierarchical clustering
- k means clustering

Dimensionality reduction. PCA

dimensionality reduction = modelling approach that reduces/ summarises the number of variables in a dataset to a few highly informative or representative ones.

Why is this necessary: large datasets with many variables are inherently difficult to summarise.

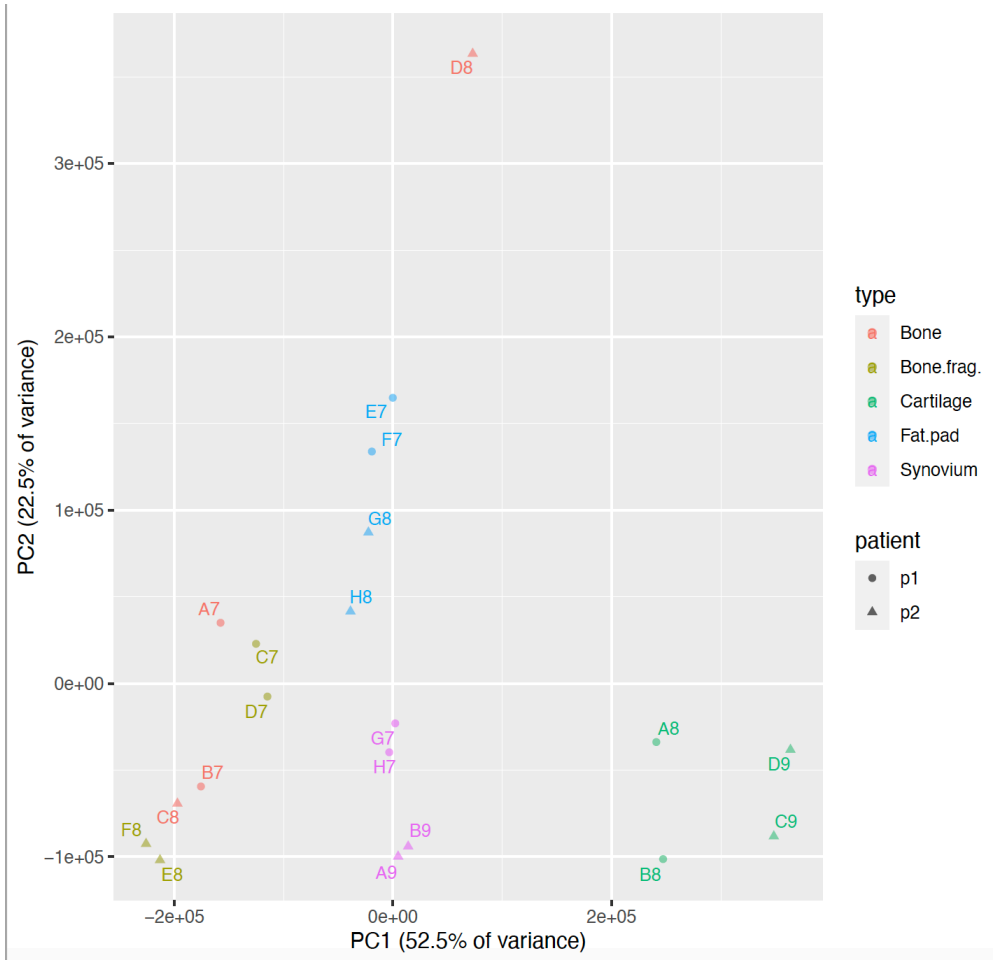
bulk mRNAseq experiments provide observations for many variables (i.e. many genes) but with relatively few samples e.g., few time points or conditions. The imbalance between the number of variables and the number of observations is referred to as **large p, small n**, and makes statistical analysis difficult. Dimensionality reduction techniques provide useful information on the relations between samples.

For single cell RNA-sequencing (scRNA-seq) the scale of datasets has shifts away from large p, small n, towards providing measurements of many variables (thousands of genes) but with a corresponding large number of observations (large n) albeit from potentially heterogeneous populations.

scRNA-sequencing was largely driven by the need to investigate the transcriptomes of cells that were limited in quantity, such as embryonic cells, with early applications in mouse blastomeres.

Dimensionality reduction. PCA

We are exceptionally good at identifying patterns in two and three-dimensional spaces. To illustrate this, note the Great Britain shaped cloud in the image (presumably drifting away from an EU shaped cloud, not shown). Golcar Matt/Weatherwatchers BBC News

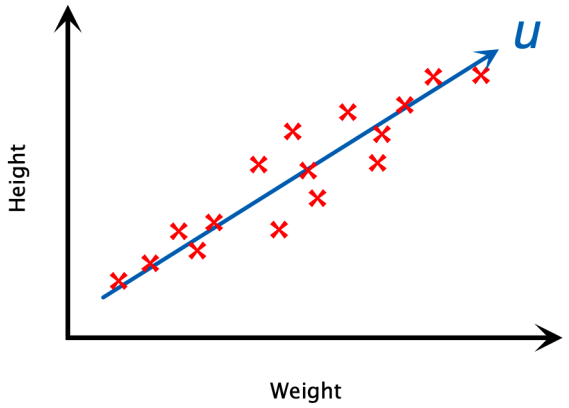


PCA on bulk mRNAseq data.
mRNAseq data analysis lecture.

Dimensionality reduction

dimensionality reduction transforms a n-dimensional dataset to a k-dimensional dataset with k < n
- dataset compression
- less memory storage consumption
- machine learning algorithms run faster on lowdimensional data
- data visualization: high-dimensional data can be transformed to 2D or 3D for plotting

PCA: we project the data on k orthogonal bases vectors u that minimize the projection error



Original dataset containing the features weight and height

Projection on the vector u

Dimensionality reduction

input: $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$

preprocessing:

- mean normalization
  1. compute mean of each feature *j*

$$\mu_j = \frac{1}{m}\sum_{i=1}^{m} x_j^{(i)}$$

  2. subtract the mean from data

$$x_j^{(i)} \leftarrow x_j^{(i)} - \mu_j$$

- feature scaling

$$x_j^{(i)} \leftarrow a_j x_j^{(i)}$$

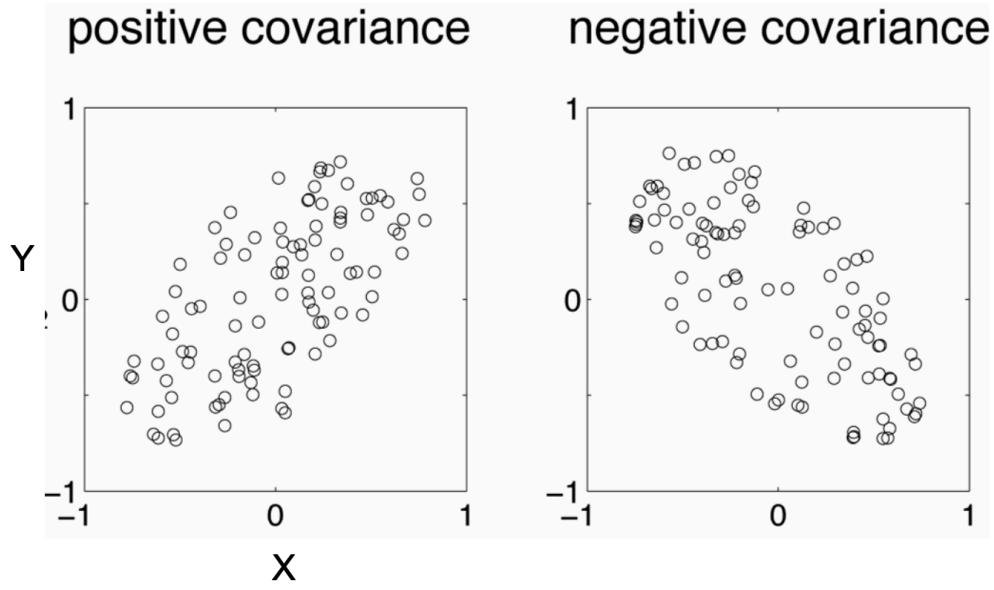
Aim: estimate the variance and covariance as a measure of the "spread" of a set of points around their mean

Variance: measure of the deviation from the mean for points in one dimension e.g. heights

Covariance: measure of how much each of the dimensions vary from the mean with respect to each other e.g. variation of height wrt weight

Covariance is measured between 2 dimensions to see if there is a relationship between the 2 dimensions e.g. number of hours studied & marks obtained.

Dimensionality reduction



PCA
is a linear transformation that chooses a new coordinate system for the data set such that greatest variance by any projection of the data set comes to lie on the first axis (then called the first principal component), the second greatest variance on the second axis, and so on.

Choosing *k*

average squared projection error: $\frac{1}{m}\sum_{i=1}^{m}\left\|x^{(i)} - x_{approx}^{(i)}\right\|^2$

total variation in the data: $\frac{1}{m}\sum_{i=1}^{m}\left\|x^{(i)}\right\|^2$

PCA Algorithm

compute covariance matrix $\Sigma = \frac{1}{m}\sum_{i=1}^{m} x^{(i)} x^{(i)T}$

diagonalize covariance matrix (using SVD)

$S = U^{-1}\Sigma U$

U is the matrix of Eigenvectors
S a diagonal matrix containing the Eigenvalues

dimensionality reduction from *n* to *k* dimensions:
project the data onto the Eigenvectors corresponding to the *k* largest Eigenvalues

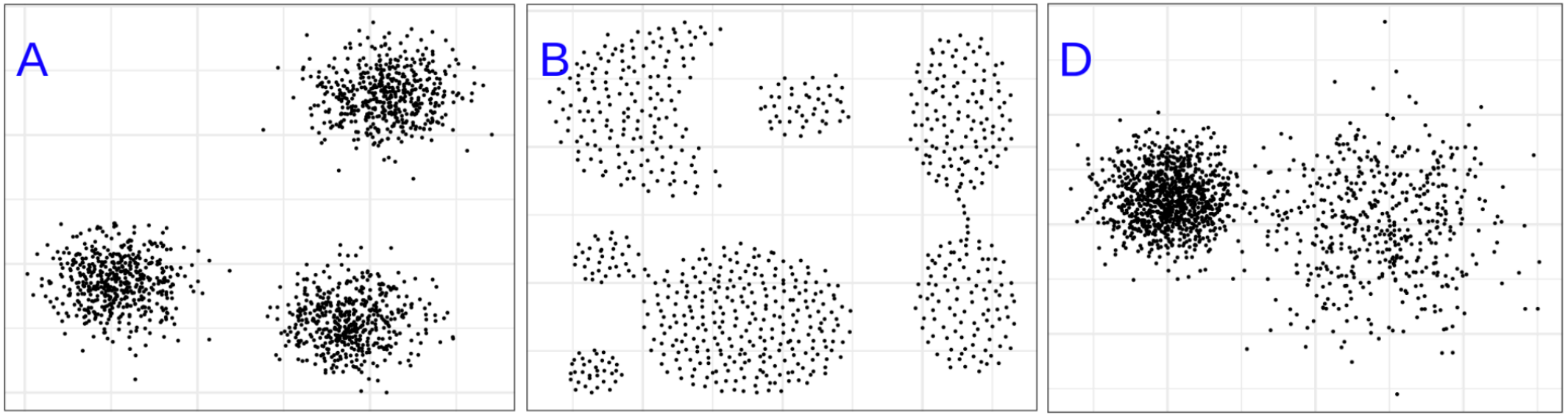
$$z^{(i)} = U_{reduce}^{T} x^{(i)}$$

Clustering.

Clustering: method for finding groups (clusters) of similar objects.
The members of a cluster should be more similar to each other, than to objects in other clusters.
Clustering algorithms aim to **minimize intra-cluster variation** and **maximize inter-cluster variation**.

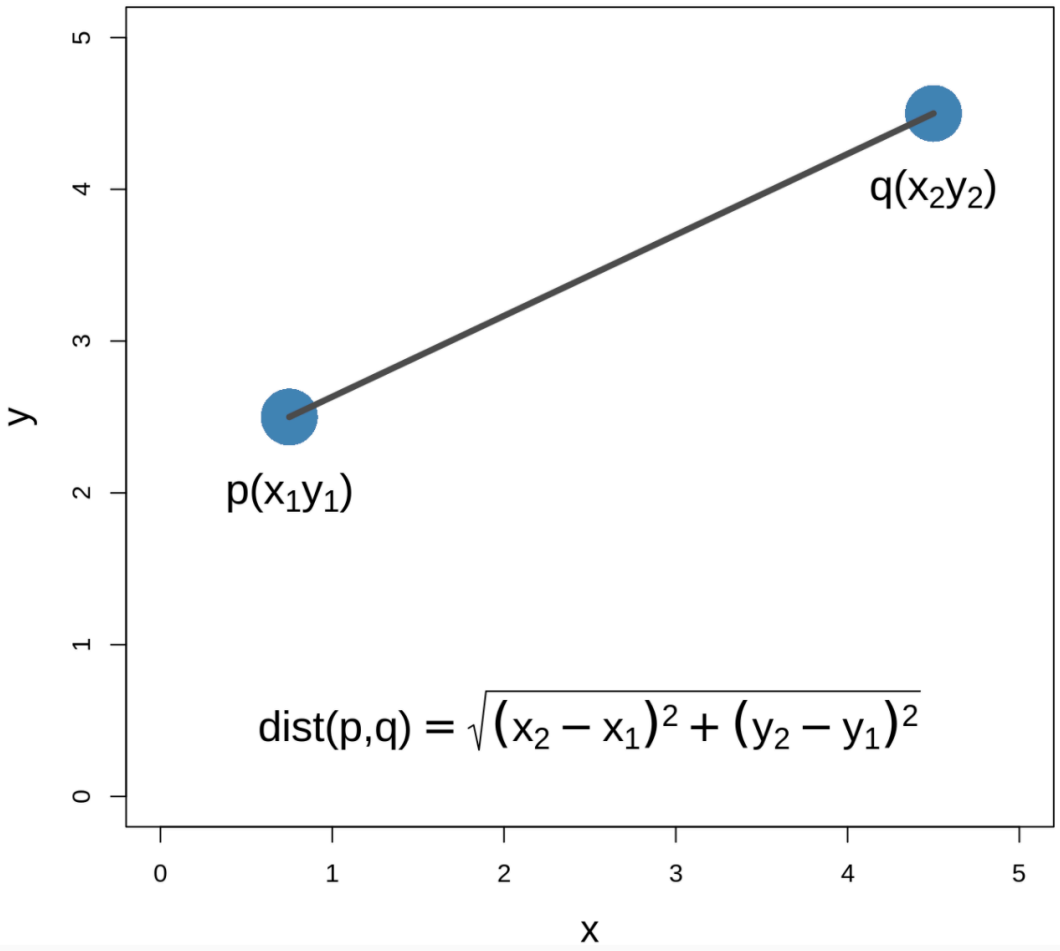
Methods of clustering:
[a] Hierarchic techniques produce dendrograms (trees) through a process of division or agglomeration.

[b] Partitioning algorithms divide objects into non-overlapping subsets (examples include k-means and DBSCAN)



Clustering. Distances

$$distance\,(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$



Manhattan distance:

$$distance\,(p,q) = \sum_{i=1}^{n}|p_i - q_i|$$

There are other metrics to measure the distance between observations. For example, the Minkowski distance is a generalization of the Euclidean and Manhattan distances and is defined as

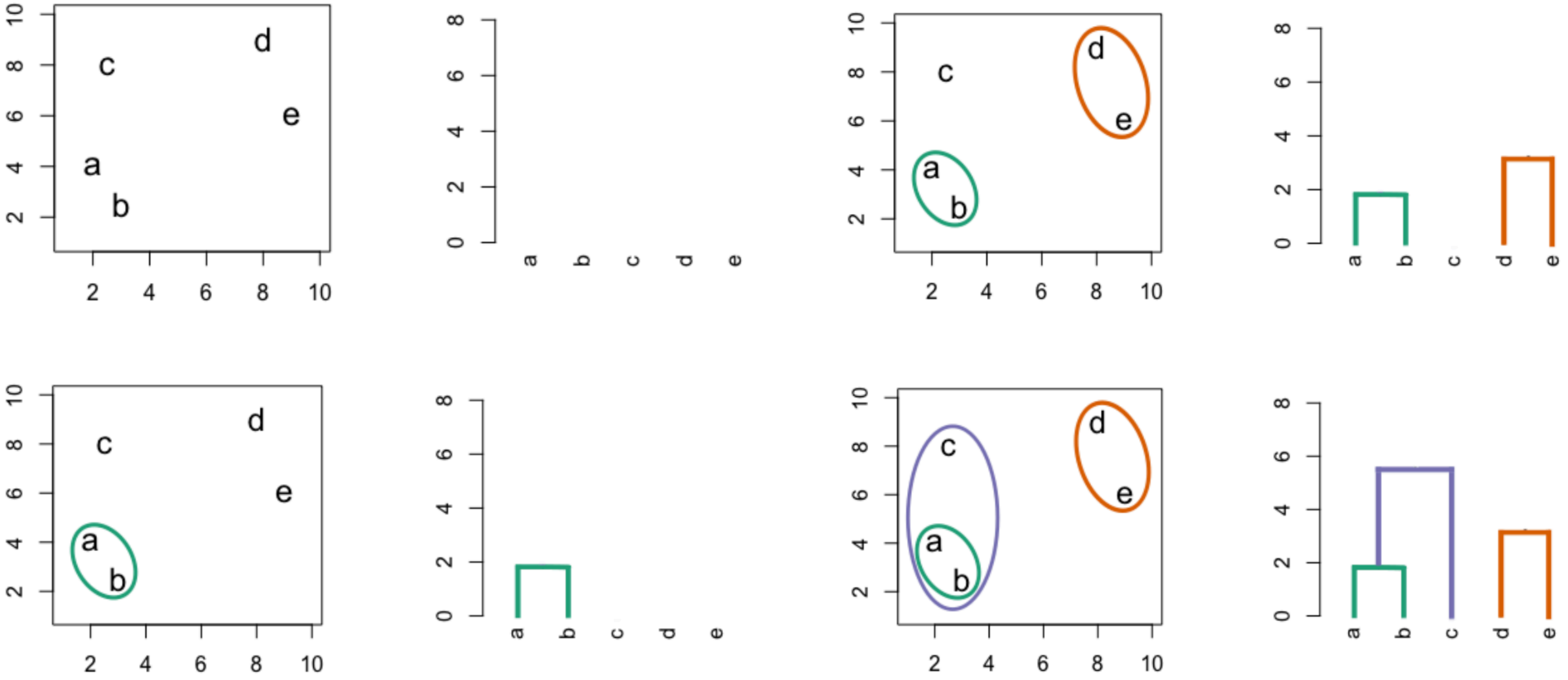
Minkowski distance:

$$distance\,(p,q) = \sqrt[p]{\sum_{i=1}^{n}(p_i - q_i)^p}$$

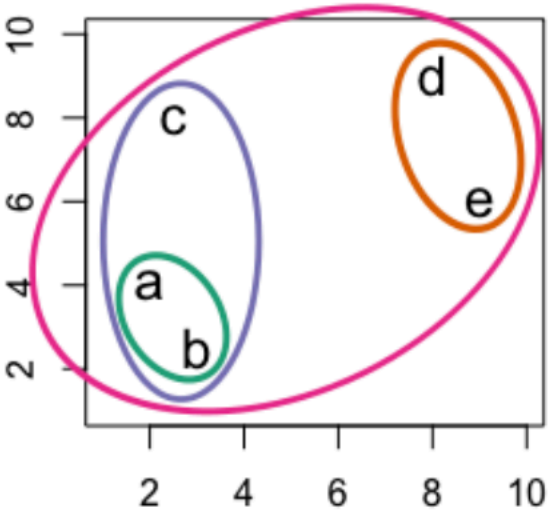
where p>0 (Han, Pei, and Kamber 2011). When p=2 the Minkowski distance is the Euclidean distance and when q=1 it is the Manhattan distance

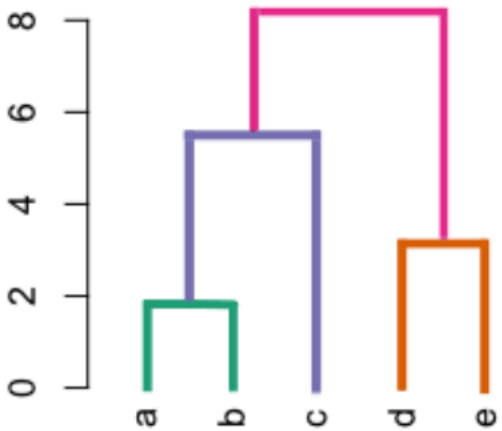
Alternative distances are correlation based distances.

Clustering. Hierarchical clustering



Clustering. Hierarchical clustering





Single linkage - nearest neighbours linkage

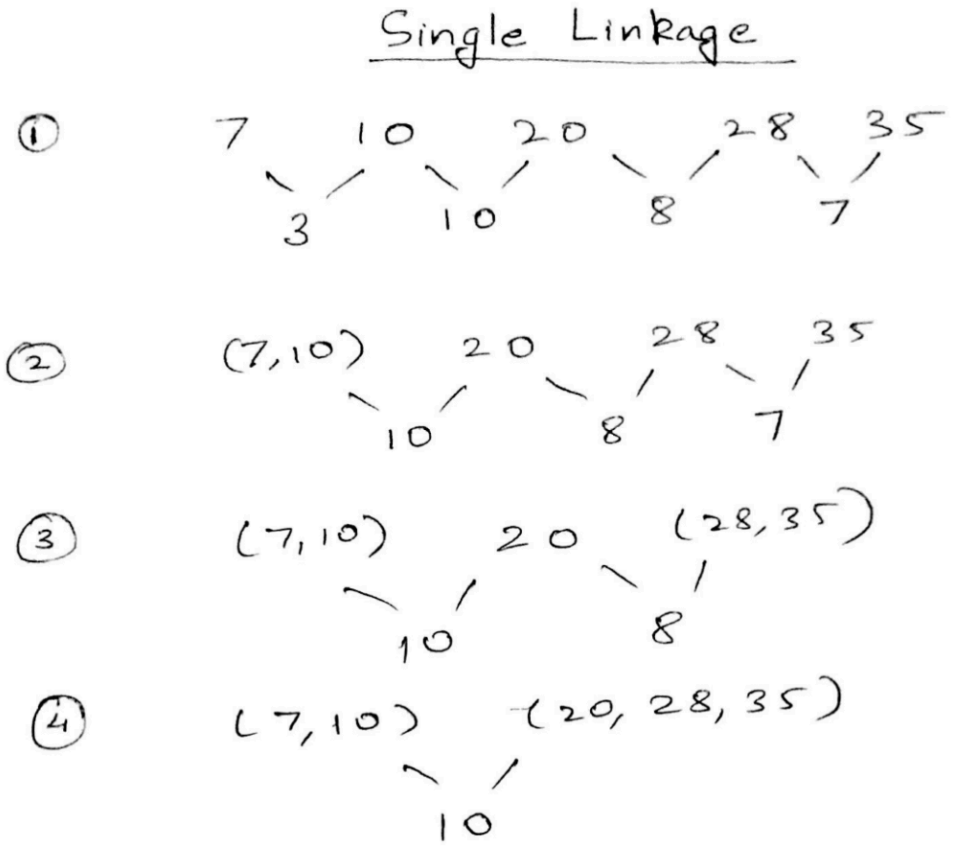
Complete linkage - furthest neighbours linkage

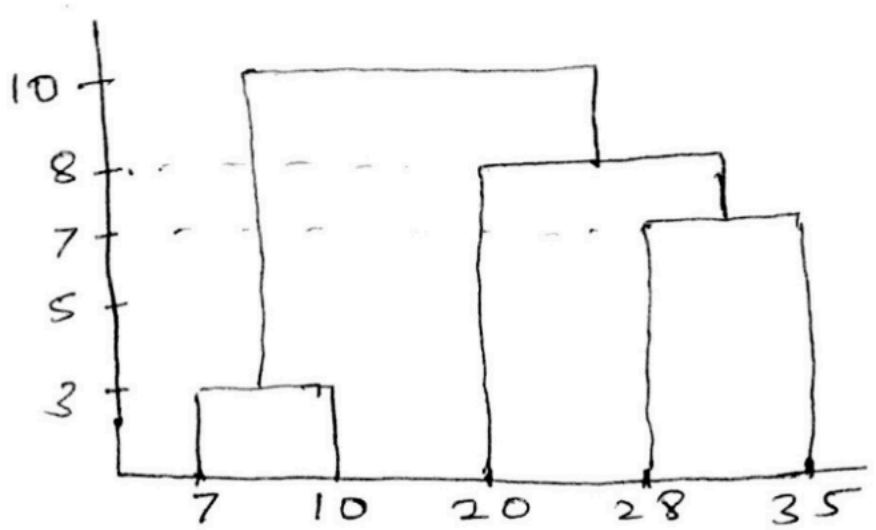
Average linkage - UPGMA (Unweighted Pair Group Method with Arithmetic Mean)

Clustering. Linkage

Single linkage - nearest neighbours linkage
Complete linkage - furthest neighbours linkage
Average linkage - UPGMA (Unweighted Pair Group Method with Arithmetic Mean)

For the data set **{7,10,20,28,35}**, perform hierarchical clustering and plot the dendrogram to visualize it.

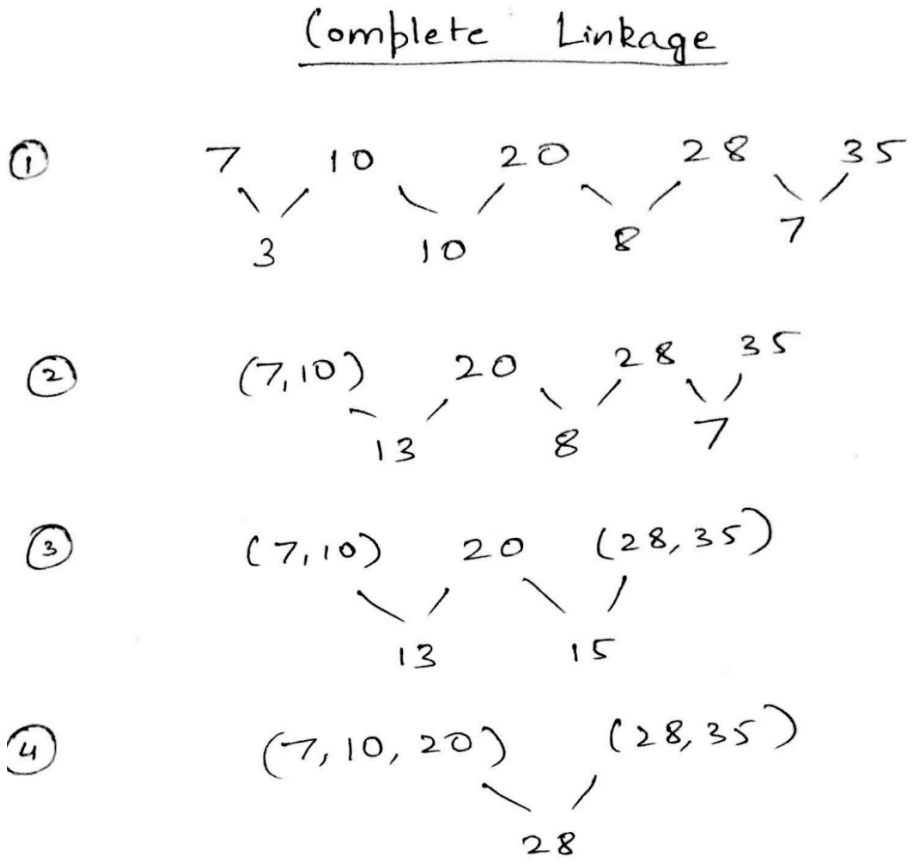


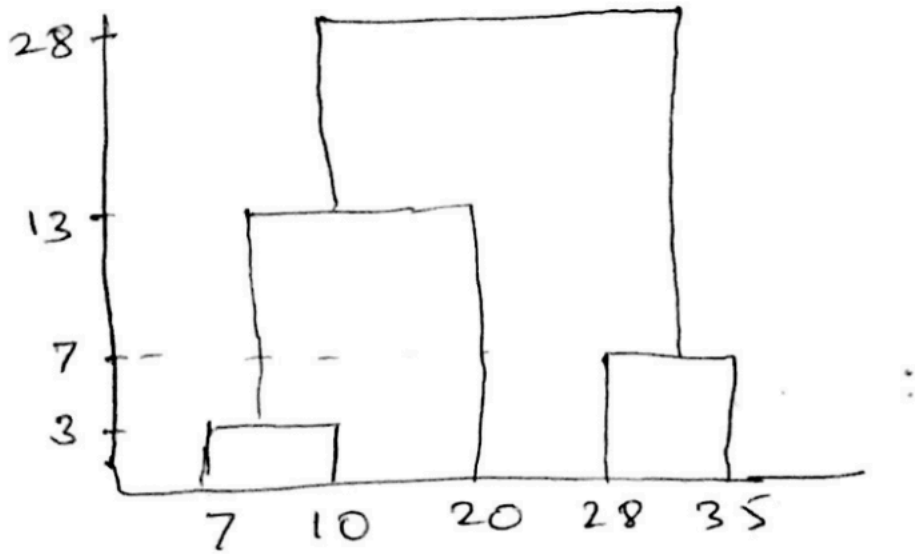


Clustering. Linkage

Single linkage - nearest neighbours linkage
Complete linkage - furthest neighbours linkage
Average linkage - UPGMA (Unweighted Pair Group Method with Arithmetic Mean)

For the data set **{7,10,20,28,35}**, perform hierarchical clustering and plot the dendrogram to visualize it.

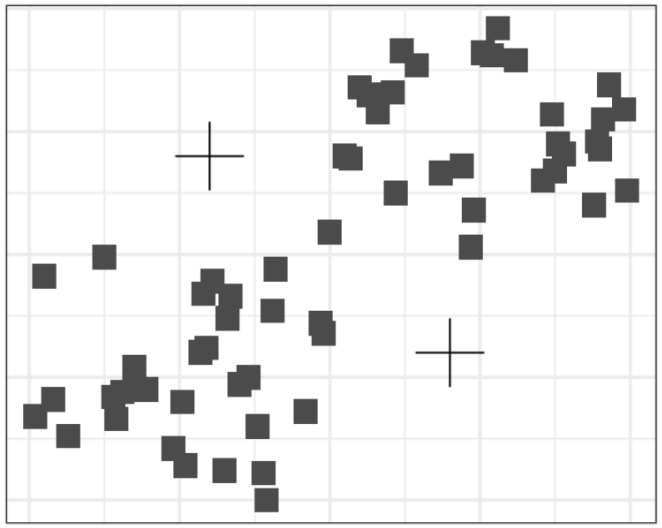


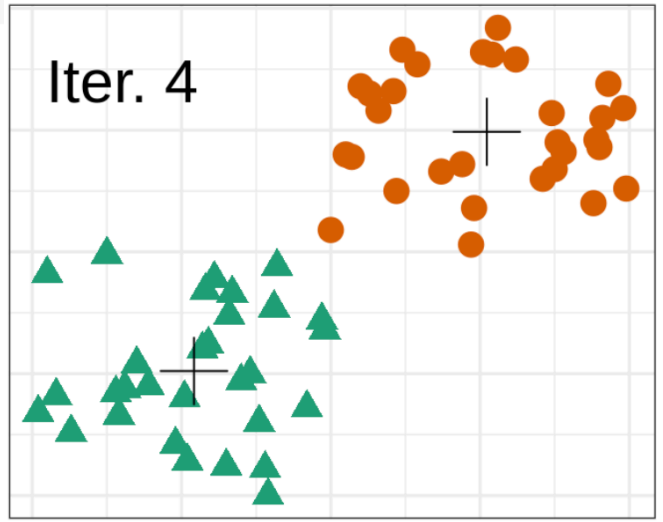


Clustering. K means clustering

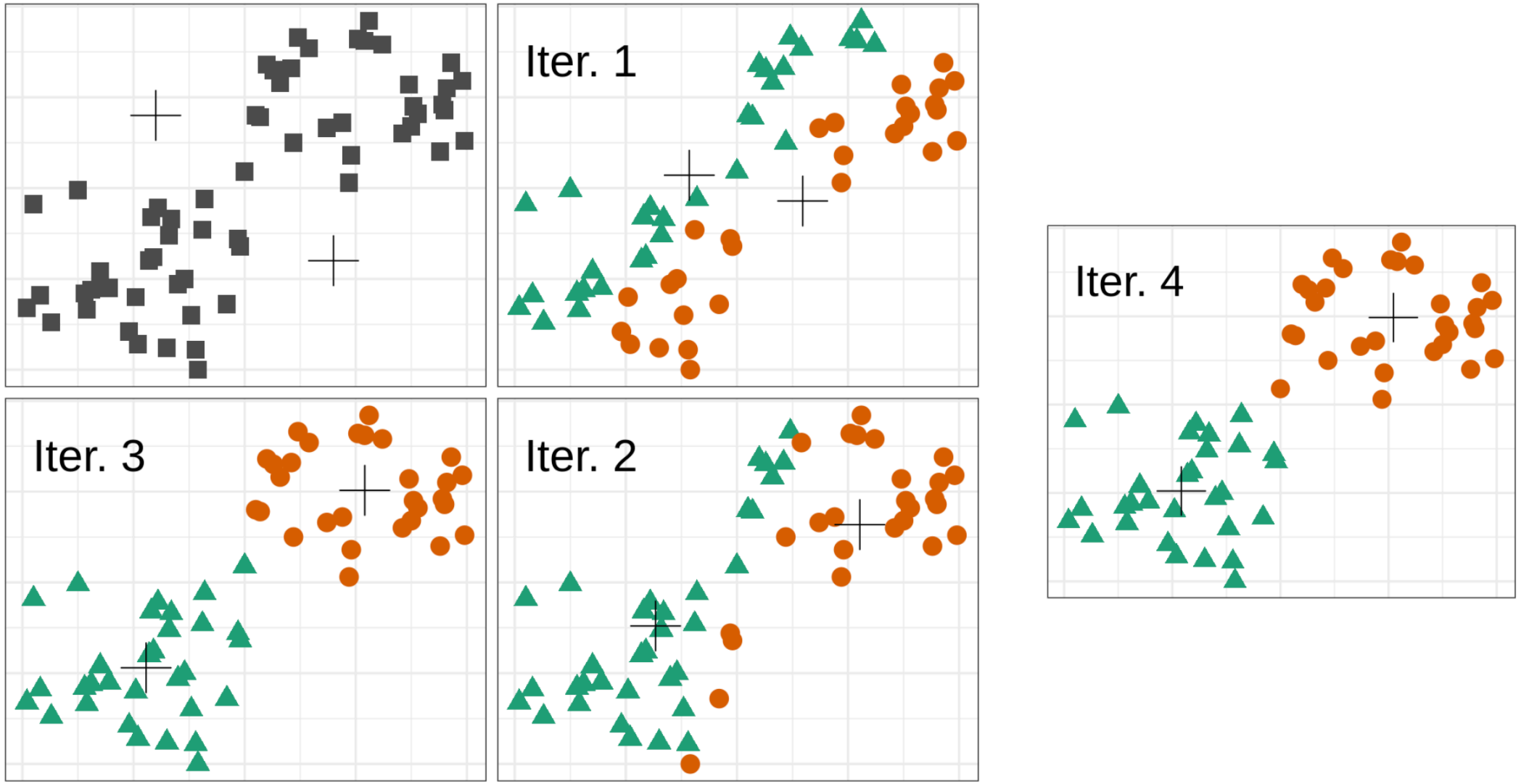
Pseudocode for the K-means algorithm

```
randomly choose k objects as initial centroids
while true:
  1. create k clusters by assigning each object to closest centroid
  2. compute k new centroids by averaging the objects in each cluster
  3. if none of the centroids differ from the previous iteration:
        return the current set of clusters
```





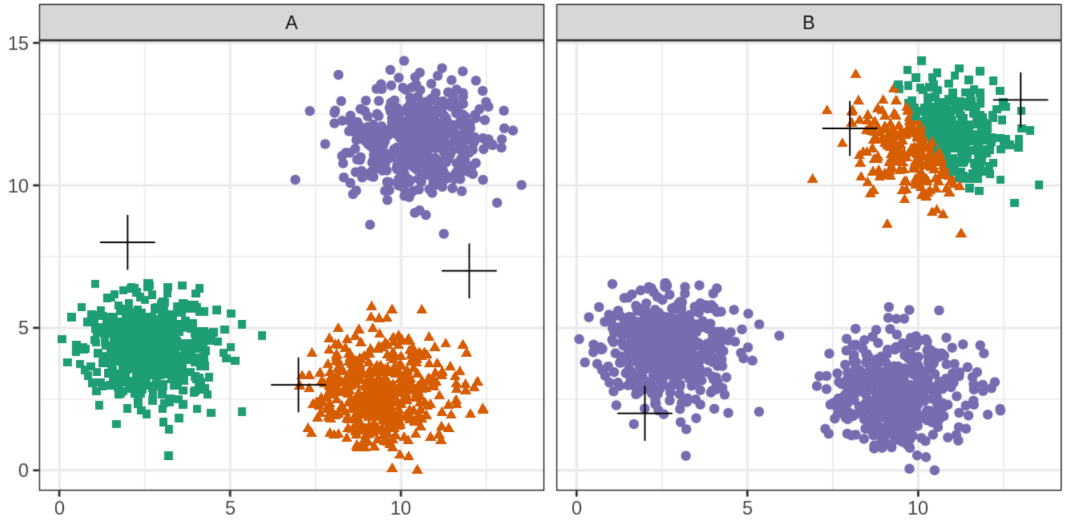
Clustering. K means clustering



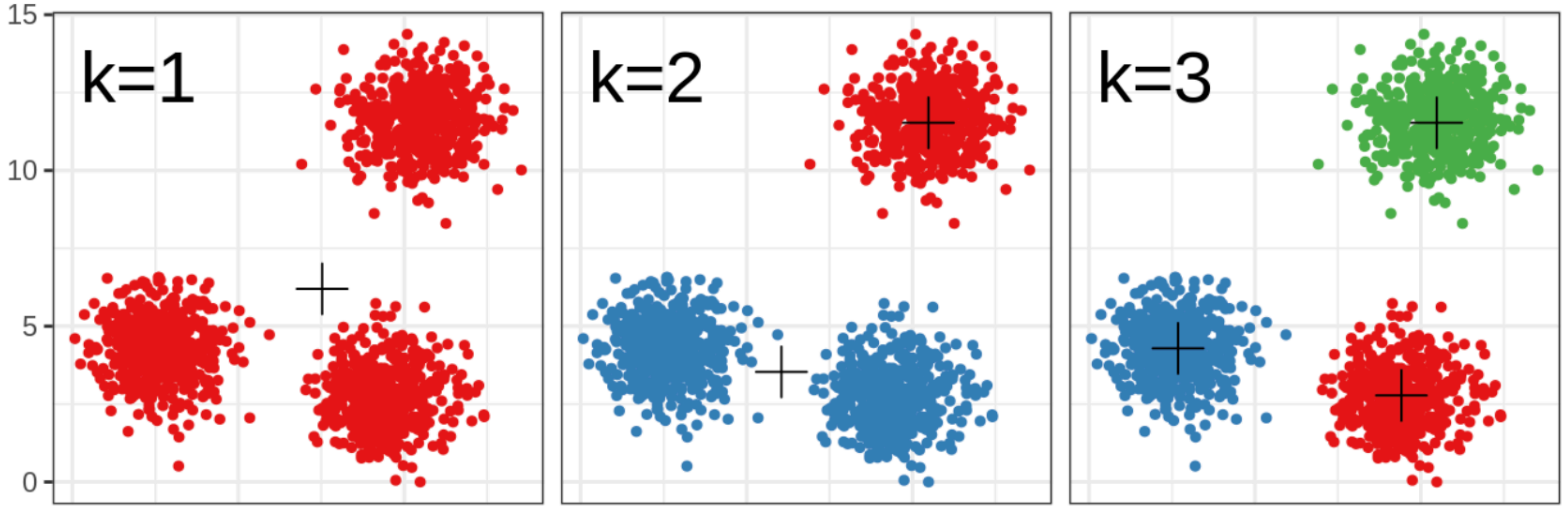
Clustering. K means clustering

Advantages:
[a] fast convergence
[b] local optima can be avoided

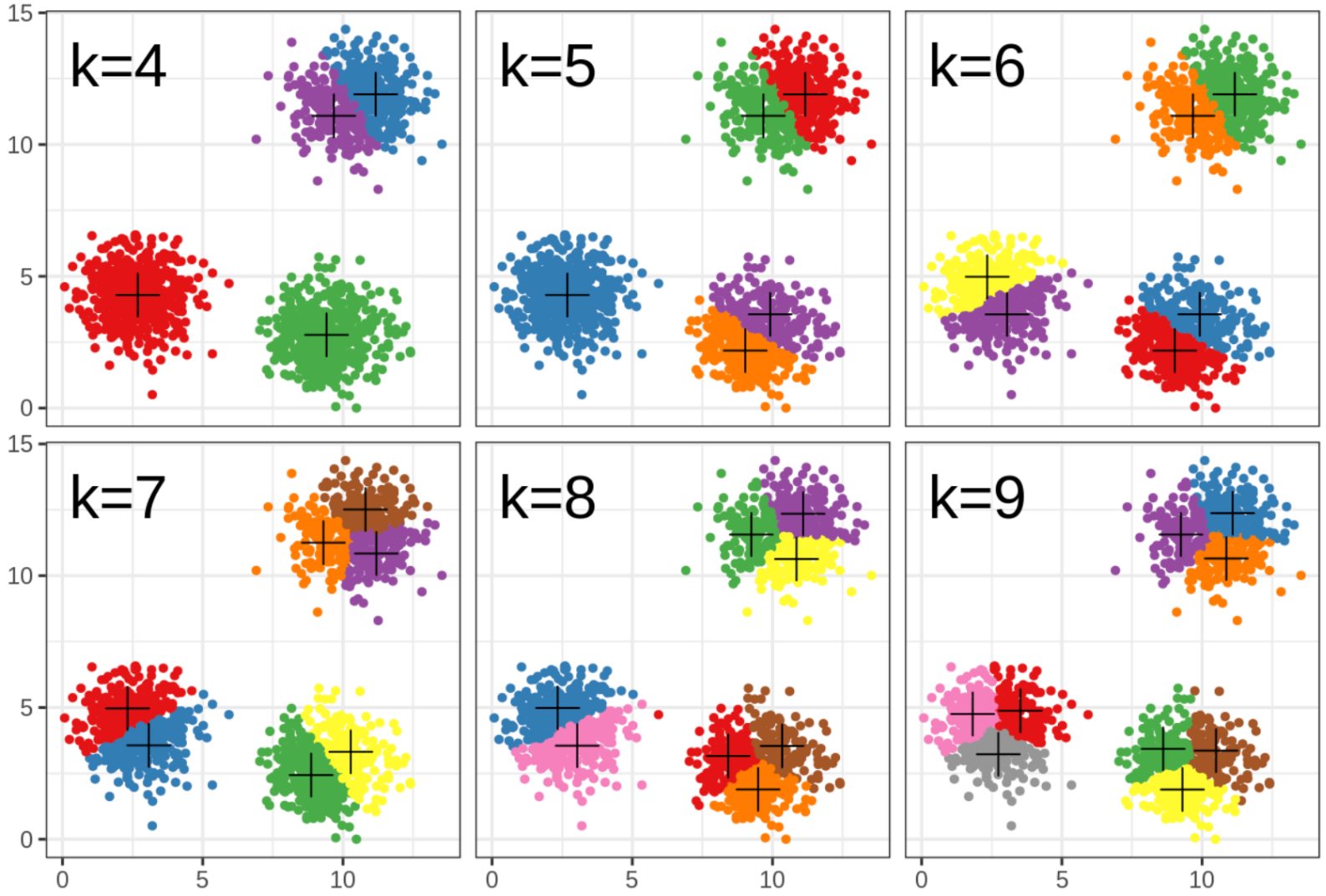
Disadvantages:
[a] starting point
[b] choosing k – the number of clusters



Initial centres determine clusters. The starting centres are shown as crosses. **A**, real clusters found; **B**, convergence to a local minimum.



Clustering. K means clustering



The variance within a cluster can be estimated.

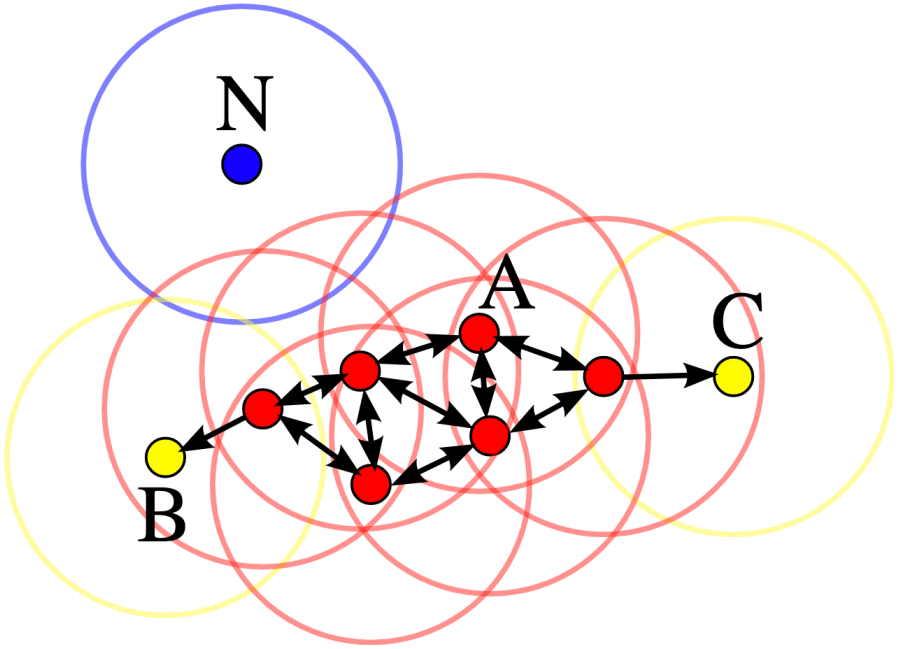
Approaches for determining the optimal number of clusters:
ARI – Adjusted Rand Index
PAC – proportion of ambiguous clustering

Approaches focused on consensus clustering

Clustering. DBSCAN

Abstract DBSCAN algorithm in pseudocode (Schubert et al. 2017)

```
1 Compute neighbours of each point and identify core points   // Identify core points
2 Join neighbouring core points into clusters                 // Assign core points
3 foreach non-core point do
      Add to a neighbouring core point if possible            // Assign border points
      Otherwise, add to noise                                 // Assign noise points
```

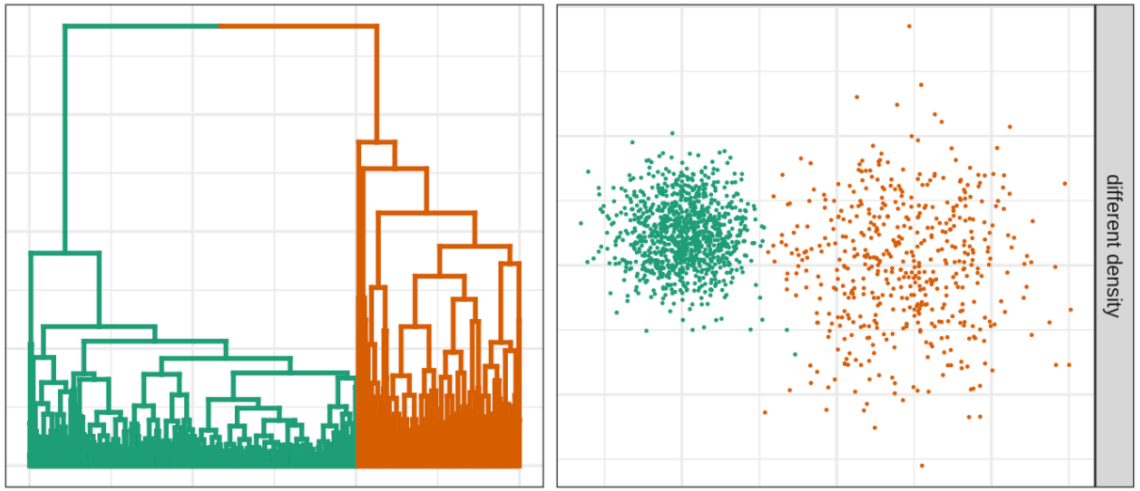


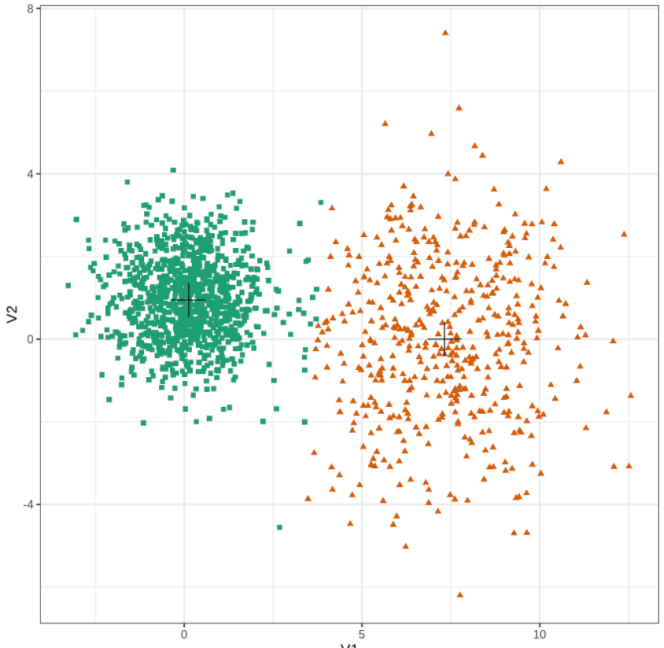
The two main parameters are:

e (eps): the radius of neighbourhoods around a data point p.

minPts: the minimum number of data points we want in a neighbourhood to define a cluster.

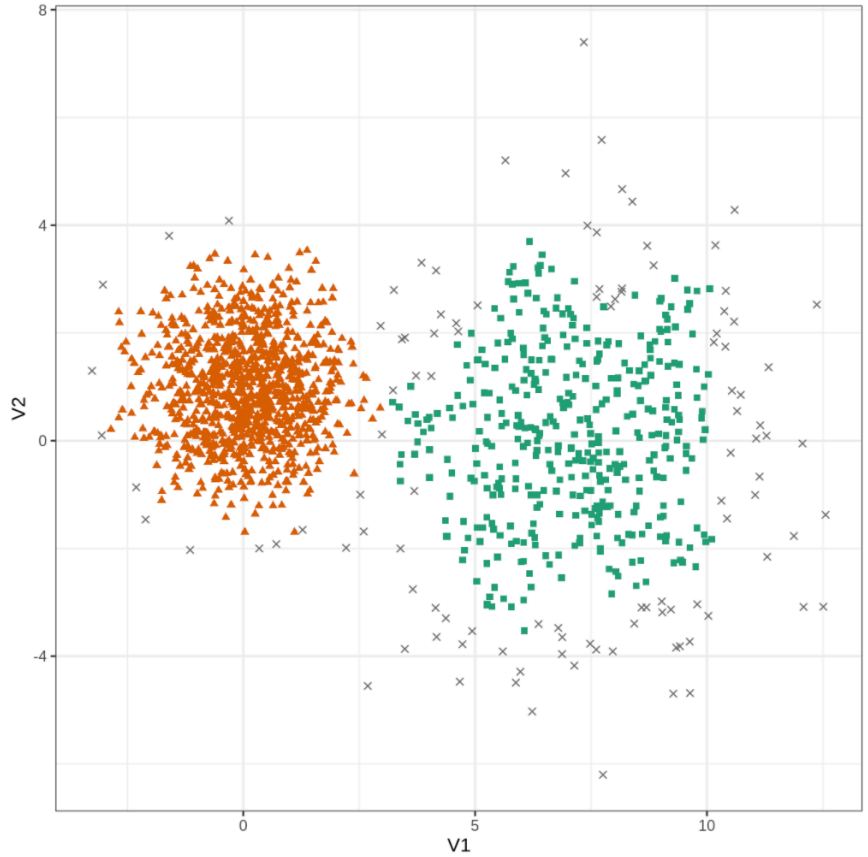
Clustering. Comparison





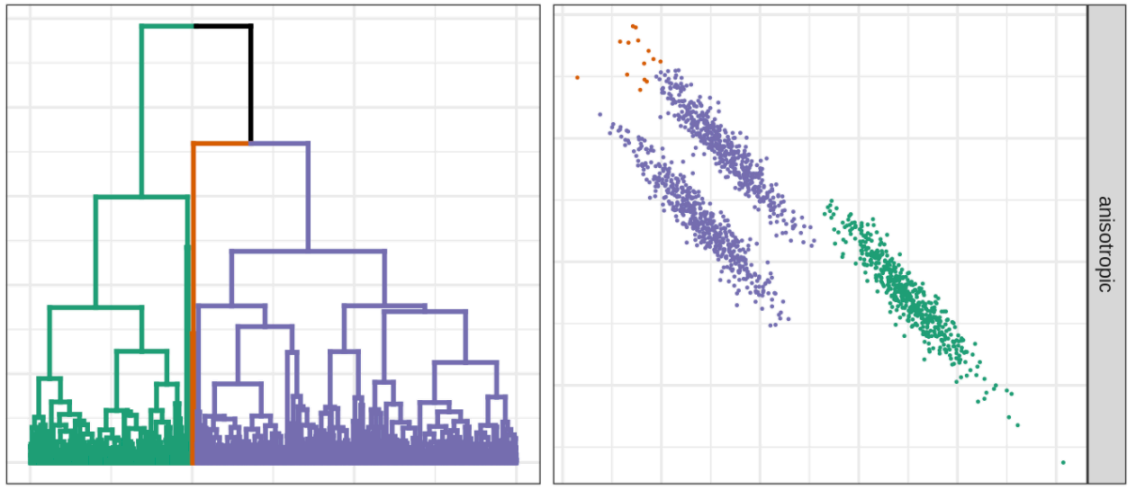
k-means, k=2

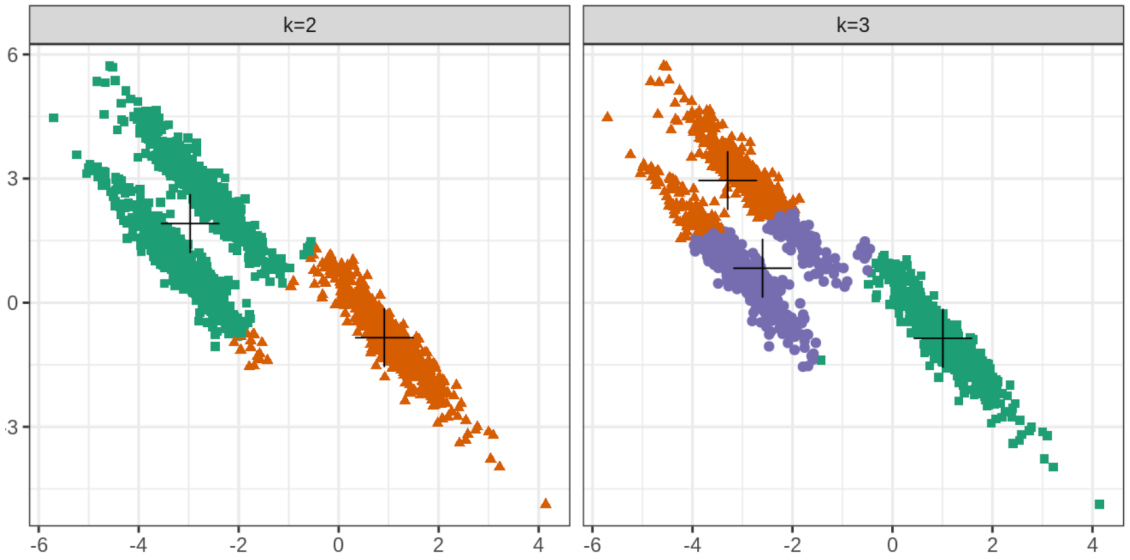
Hierarchical clustering



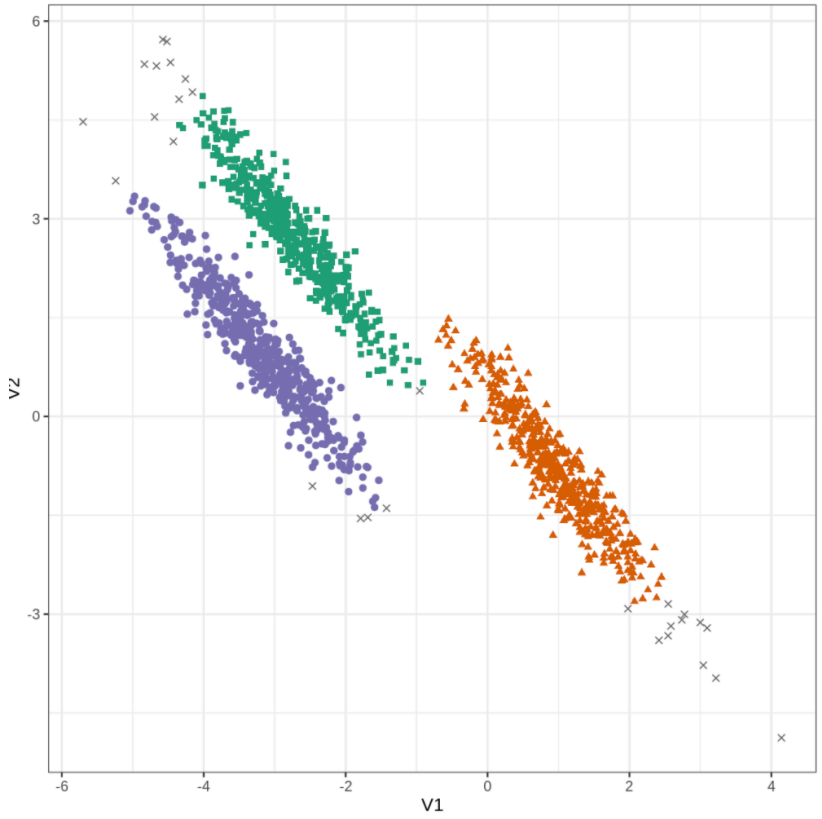
DBSCAN, eps = 0.6, minPt = 10

Clustering. Comparison





Hierarchical clustering



K means
K=2 and K = 3

DBSCAN, eps = 0.3, minPt = 10