Co-funded by the Erasmus+ Programme of the European Union

Machine learning (II)

David Tomás Díaz
@d_tomas

Universidad de Alicante

Departamento de Lenguajes y Sistemas Informáticos

Contents

- Recap
- Clustering
  - K-Means
  - Agglomerative Hierarchical Clutering
  - Evaluation
- Association rules

Contents

Recap

- Main components of machine learning
  - Training instances
    - A.k.a. *corpus* or *dataset*
    - Set of examples used to train (teach) the system
  - Features
    - Attributes that represent each example
  - Algorithm
    - Algorithm that learns from the features extracted for each training instance

Recap

- Two main groups of algorithms
  - Supervised Learning
    - Training instances are labelled
    - There is a "special" attribute: the *class*
    - Approaches
      - Classification
      - Regression
  - Unsupervised Learning
    - Training instances without labels
    - Approaches
      - Clustering
      - Association rules

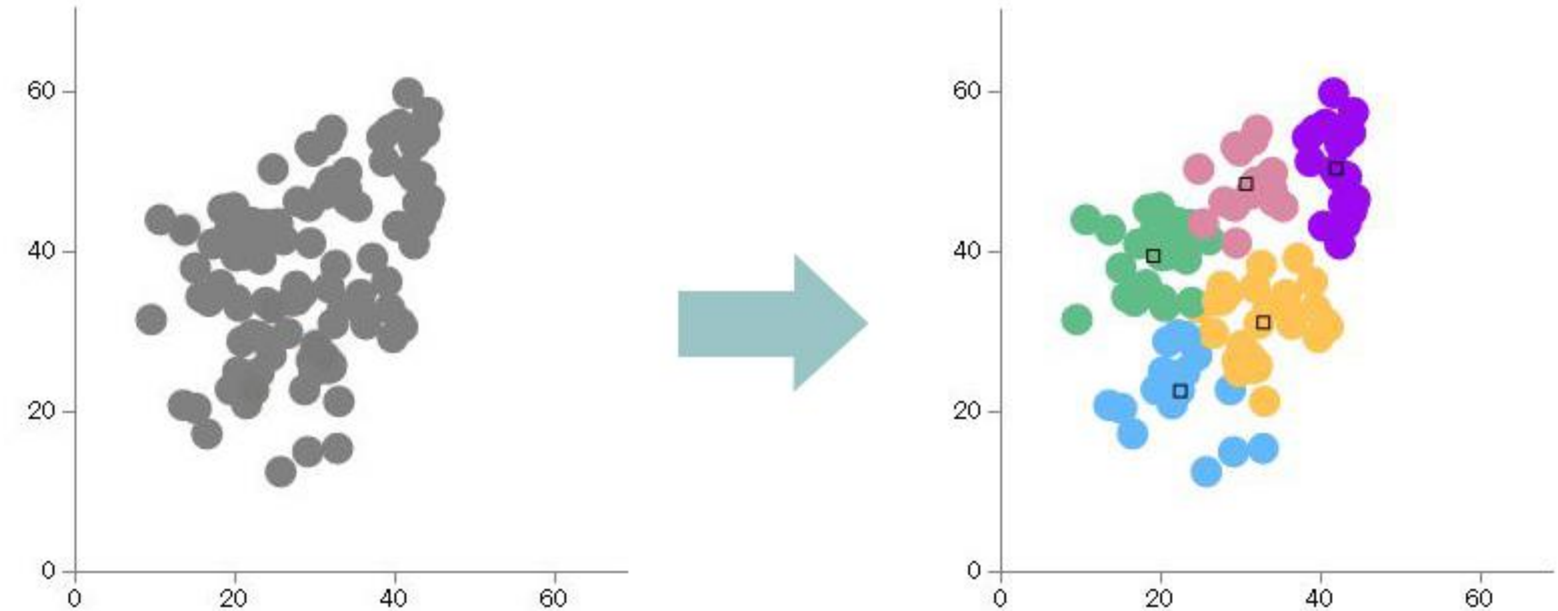
Contents

- Recap
- **Clustering**
  - K-Means
  - Agglomerative Hierarchical Clutering
  - Evaluation
- Association rules

Clustering

- *Clustering* is a data segmentation technique
- One of the most widespread descriptive methods of data analysis
- Divides a collection of disorganised objects (*instances*) into groups (*clusters*) with respect to a set of properties (*features*)
- A *cluster* is a collection of objects which are similar between them and different to objects of other clusters
- Unsupervised learning: no predefined classes!

Clustering



Clustering

- Applications
  - Marketing
    - Finding customer profiles that make a cluster
    - Business can develop a specific strategy for each cluster
  - Retail
    - Divide all stores of a particular company into groups of establishments
    - Type of customer, turnovers, etc.
  - Medical Science
    - Discover a group of patients suitable for particular treatment
    - Age, type of disease, etc.
  - Sociology
    - Divide the population into groups of individuals who are homogeneous
    - Social demographics, lifestyle, expectations, etc.

Clustering

- Distance Measures
  - It is necessary a numerical measure that indicates how different two objects are
  - The lower its value the more similar the objects are
  - Given two data objects X_1 and X_2, the distance between X_1 and X_2 is a real number denoted by $d(X_1, X_2)$

Clustering

- Distance Measures
  - Common distance measures between data objects
    - Euclidean distance

$$d(i,j)=\sqrt{(|x_{i1}-x_{j1}|^2+|x_{i2}-x_{j2}|^2+\ldots+|x_{ip}-x_{jp}|^2)}$$

    - Manhattan distance

$$d(i,j)=|x_{i1}-x_{j1}|+|x_{i2}-x_{j2}|+\ldots+|x_{ip}-x_{jp}|$$

    - Minkowski distance

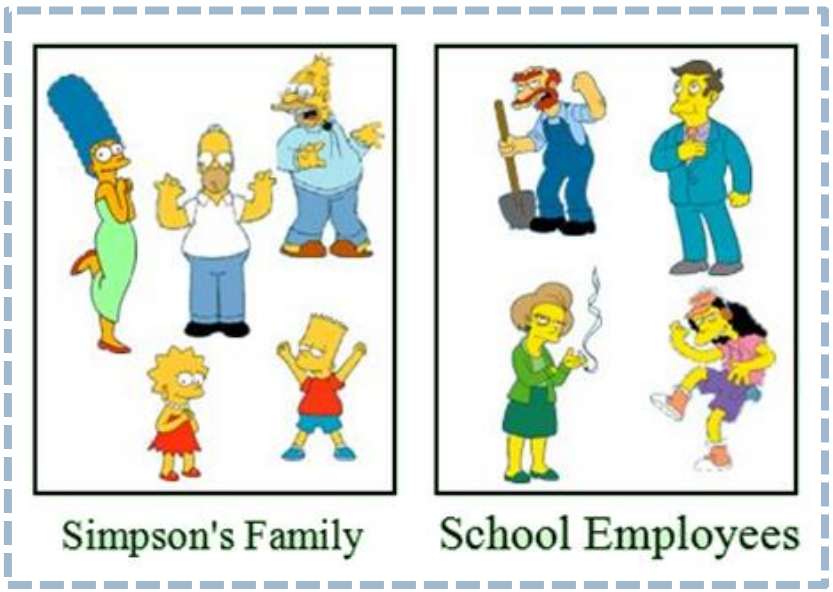
$$d(i,j)=\left((x_{i1}-x_{j1})^q+(x_{i2}-x_{j2})^q+\ldots+(x_{ip}-x_{jp})^q\right)^{1/q}$$

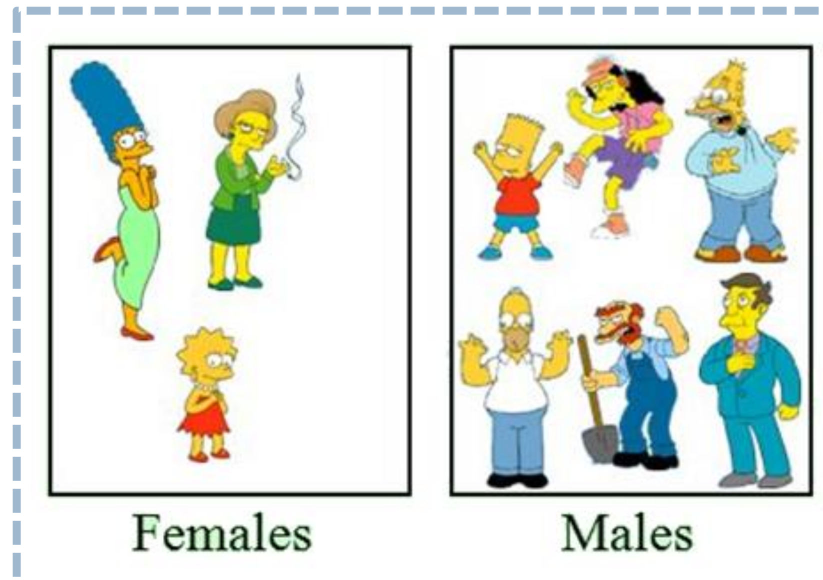
Clustering

- What is the natural grouping among these objects?

Clustering

- What is the natural grouping among these objects?
  - Clustering is a subjective task
  - Features and distance metrics are important!



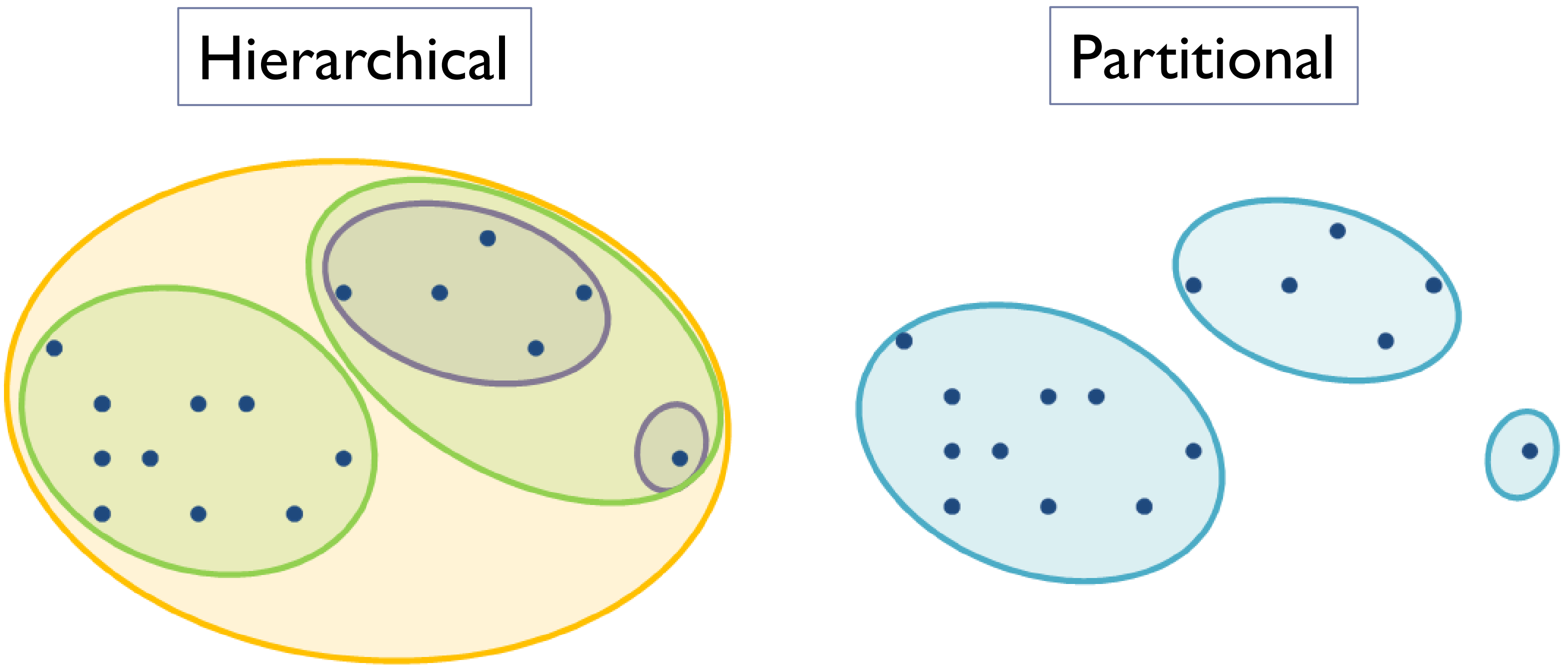


Clustering

- Types
  - Partitional
    - Objects are partitioned into non-overlapping groups
    - Each object belongs to a group only
  - Hierarchical
    - Objects are partitioned into nested groups that are organised as a hierarchical tree

Clustering

- Types



Clustering

- Algorithms
  - As in classification and regression, there are different algorithms
    - K-means
    - Expectation Maximisation (EM)
    - Cobweb
    - …
  - In some of them, it is necessary to set in advance the number of clusters

Contents

Clustering

- K-means
  - Most widely used clustering method
  - Partitions `n` units into `k` $\leq$ `n` distinct clusters
  - The number of clusters `k` must be specified
  - Each cluster is associated with a *centroid*
    - The centroid is the mean of the points in the cluster
    - Each point is assigned to the cluster with the closest centroid
    - Initial `k` centroids are chosen randomly
  - Goal: minimizing the within-cluster sum of squares (*WCSS*)

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} \left\| x - \mu_i \right\|^2$$

Clustering

- K-means
  - Two common initialization approaches
    - Randomly choose `k` units from the dataset
      - Use them as the initial cluster means
    - Randomly assign one of the `k` clusters to each unit
      - Proceed to the update step
      - Compute initial means as the centroids of the clusters' randomly assigned units
      - This partition method is preferred generally

Clustering

- K-means
  - Algorithm
    - The algorithm is implemented in four steps
      - Step 1: partition objects into `k` non-empty subsets
      - Step 2: compute seed points as the centroids of the clusters
      - Step 3: assign each object to the cluster with the nearest seed point
      - Step 4: go back to Step 2 and stop when no more new assignments
    - There is no guarantee that it converges to the global optimum
    - Final result may depend on how is the starting of the algorithm

Clustering

- K-means
  - Algorithm

https://youtu.be/5I3Ei69I40s

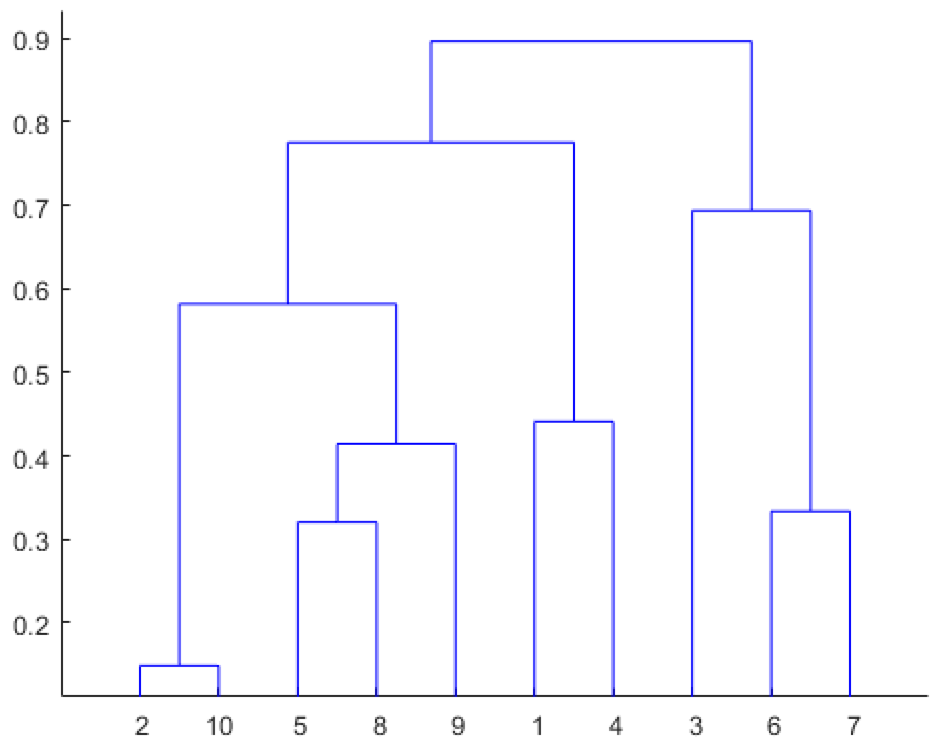
Contents

Clustering

- Agglomerative Hierarchical Clustering
  - Produces a sequence of nested partitions into n clusters
  - These nested partitions are of increasing heterogeneity
  - General form of the algorithm
    - Step 1: objects are the initial clusters
    - Step 2: calculate the distance between the clusters
    - Step 3: merge the two closest clusters together and replace with a single cluster
    - Step 4: repeat Step 2 and the complete process until a single cluster containing all the objects remains

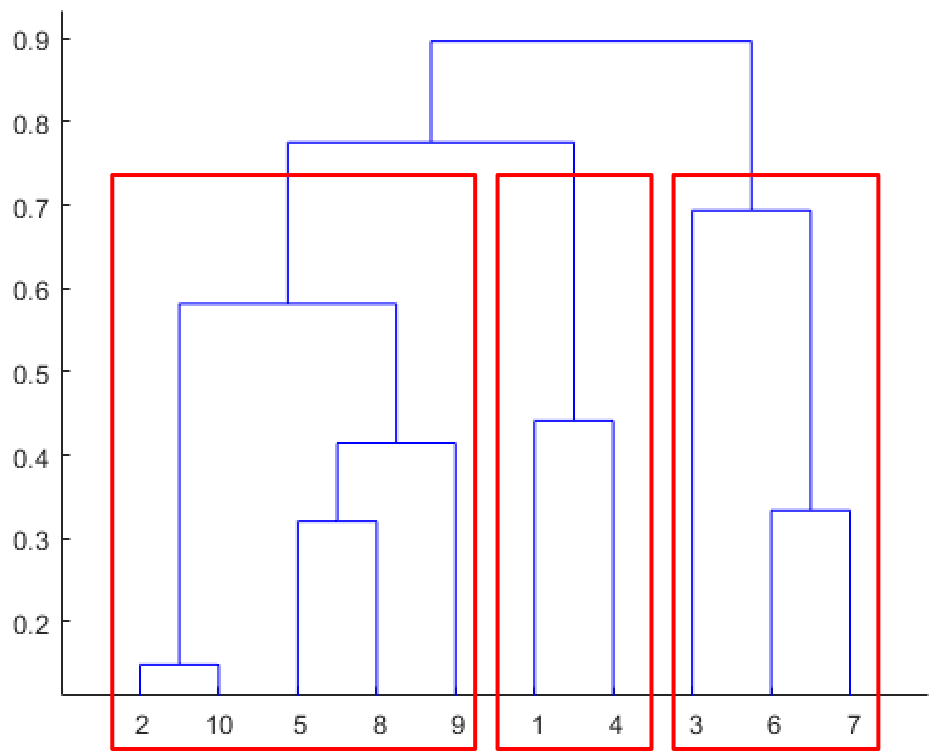
Clustering

- Agglomerative Hierarchical Clustering
  - The tree generated by AHC is also known as *dendogram*



Clustering

- Agglomerative Hierarchical Clustering
  - The tree can be cut to get clusters



Contents

- Recap
- Clustering
  - K-Means
  - Agglomerative Hierarchical Clutering
  - **Evaluation**
- Association rules

Clustering

- Evaluation
  - Not trivial: there is no truth
    - No true labels
    - No true response

Clustering

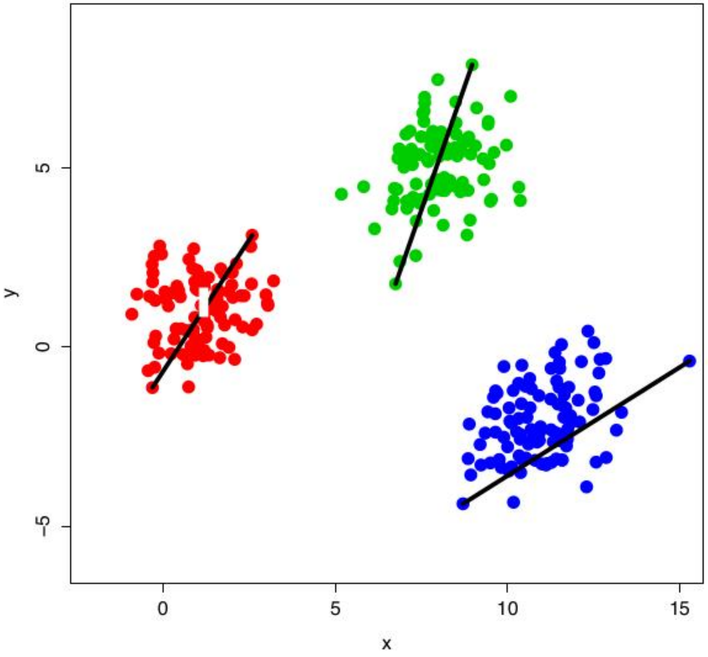
- Evaluation
  - Measure of compactness
    - Diameter

$$\mathrm{Dia}_i = \max_{x,y \in C_i} d(x, y)$$

x, y : Objects

C_i : Cluster

d : Distance (objects)



Clustering

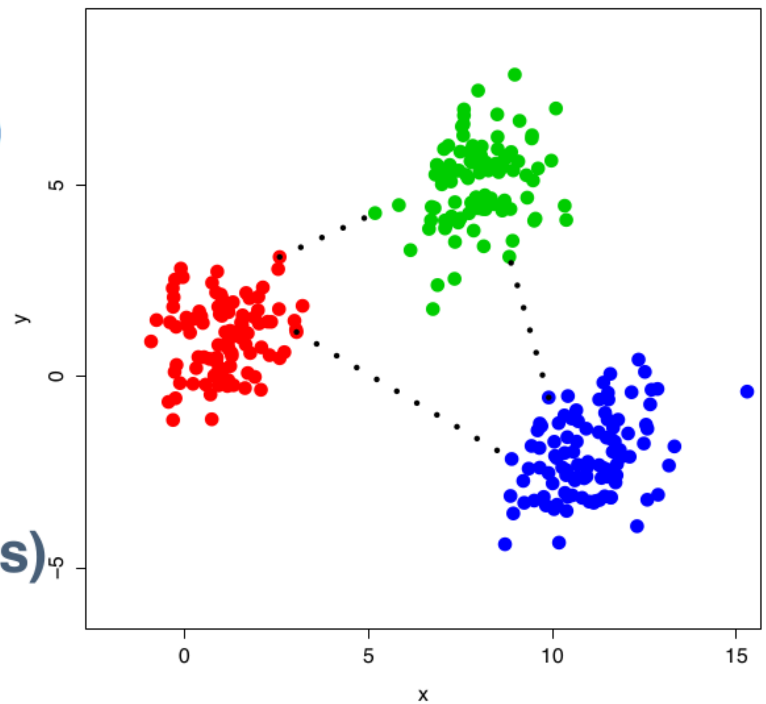
- Evaluation
  - Measure of separation
    - Intercluster Distance

$$\delta(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y)$$

x, y : Objects

C_i, C_j : Clusters

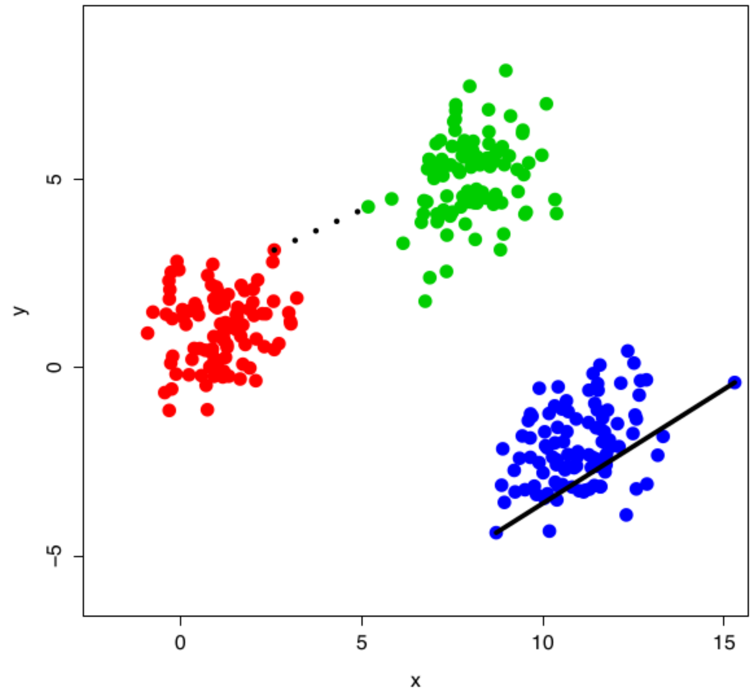
d : Distance (objects)



Clustering

- Evaluation
  - Dunn's Index

$$\frac{\displaystyle\min_{1\leq i<j\leq k} \delta(C_i, C_j)}{\displaystyle\max_{1\leq m\leq k} \mathrm{Dia}_m}$$



Clustering

Let’s practice!

https://bit.ly/3wxTdmn

Contents

- Recap
- Clustering
  - K-Means
  - Agglomerative Hierarchical Clutering
  - Evaluation
- **Association rules**

Association rules

- Works with unlabelled data (such as clustering)
- Objective: to obtain dependency rules to predict the occurrence of an item based on the occurrence of other items
- Typically used for affinity analysis (*market basket analysis*)
- If we know the purchases made by all customers during a period, we can find relationships between those products.

```
IF cheese AND milk THEN bread (probability = 0.7)
```

Association rules

- We start from a set of instances, each with a set of elements of a collection

Set of instances

Id	Items
1	bread, soda, milk
2	beer, bread
3	beer, soda, diapers, milk
4	beer, bread, diapers, milk
5	soda, diaper, milk

Rules discovered

{milk} → {soda}
{diaper, milk} → {beer}

Association rules

- Application examples
  - Management of a supermarket
    - Objective
      - Identify the items that are usually purchased together by customers
    - Approach
      - Process the point-of-sale data collected at checkouts to find dependencies between the items
    - Classic example
      - The parable of diapers and beer
      - If a customer buys diapers and milk, they are very likely to buy beer

Association rules

- Application examples
  - Inventory management
    - Objective
      - A repair company wants to anticipate the nature of repairs to its products
      - Keep service vehicles equipped with the right components to reduce the number of visits to a household
    - Approach
      - Process data on tools and components needed in previous repairs at different consumer locations
      - Discovering the co-occurrence of patterns

@d_tomas

David Tomás Díaz