Unsupervised Learning
K-Means

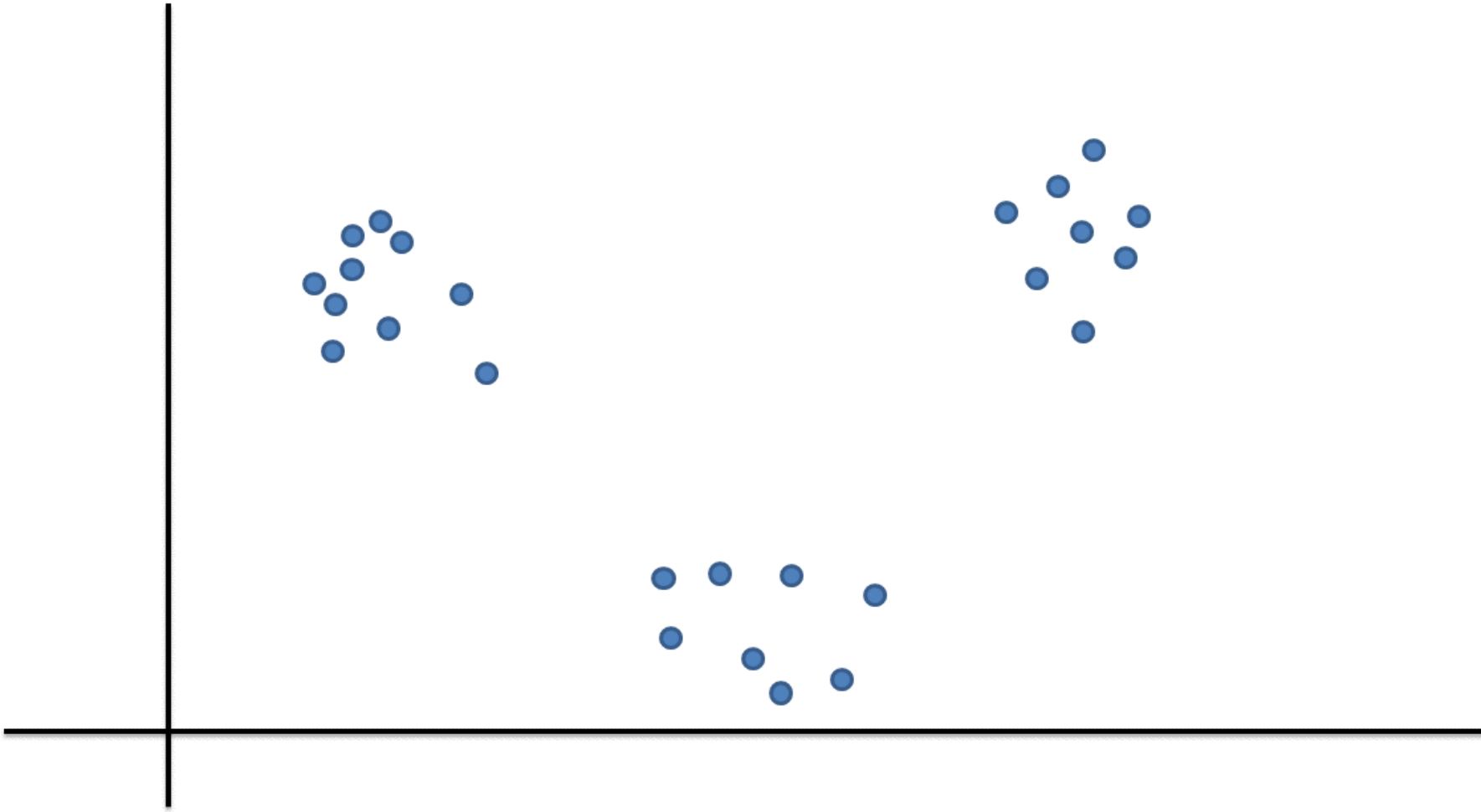
Houcine Senoussi

March 10, 2016

What is it about ?

- Unsupervised learning also called **clustering**.
  - Data have no class attributes.
  - Objective : to find some intrinsic structure in the data.
  - Clustering : organizes data into **similarity groups** called **clusters** :
    - Data instances in the same cluster are similar to each other.
    - Data instances in different clusters are very different from each other.
- Clustering can be partitionnal or hierarchical.
- Similarity is measured by similarity/distance function.

Example



Applications

- Marketing : Segmentation. Partition customers into a small number of groups according to their similarities and design some marketing materials for each group.

K-Means Algorithm

- The best known partitionnal clustering algorithm.
- Simple and efficient.
- Input : Data records D and the number of clusters k.

K-Means Algorithm

- $D = \{x_1, x_2, ..., x_n\}$.
- $x_i = (x_{i1}, ..., x_{ip})$.
- In other words $x_i \in \mathbb{R}^p$.
- $K - Means$ partitions D into k clusters. Each cluster has a **center** (also called **centroid**).
- Centroid is the mean of all the data points in the cluster.

K-Means Algorithm

- **Algorithm** *K-Means(k, D)*.
  1. Choose k data points as the initial centroids (cluster centers).
  2. **repeat**
     1. **for** each data point $x \in D$.
        Compute the distance from x to each centroid.
        Assign x to the closest centroid.
     2. **endfor**.
     3. Re-compute the centroid using the current cluster memberships.
  3. **until** the stopping criterion is met.

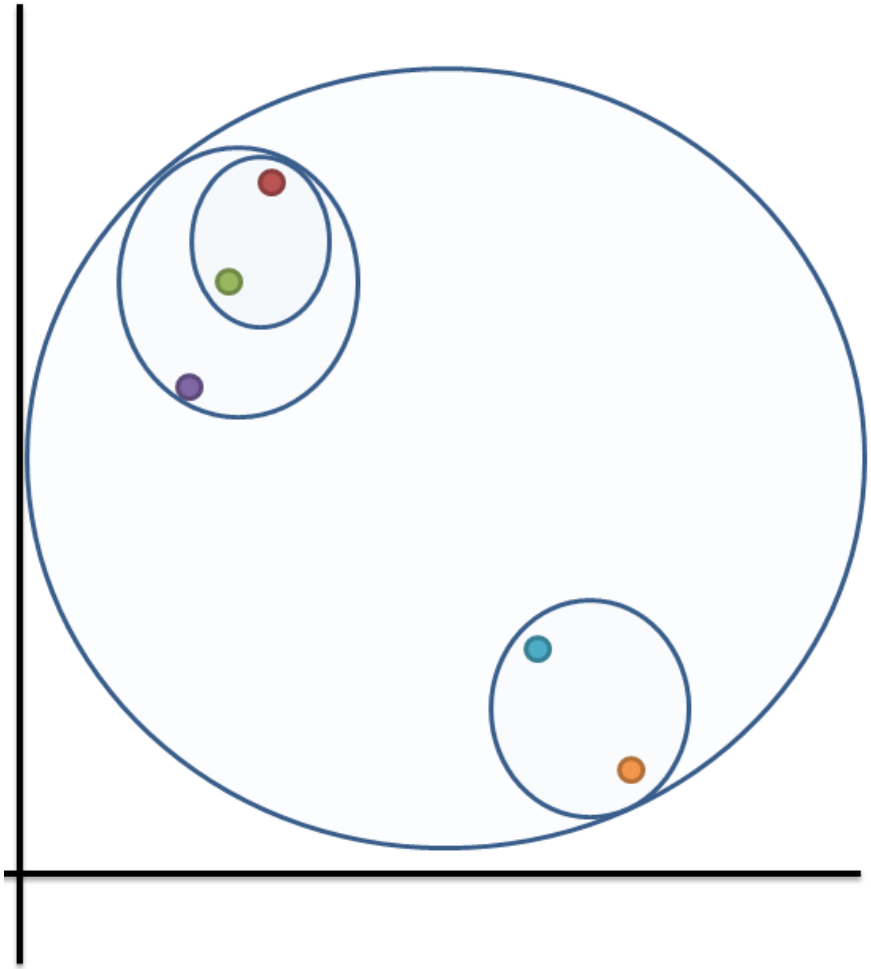
Principle

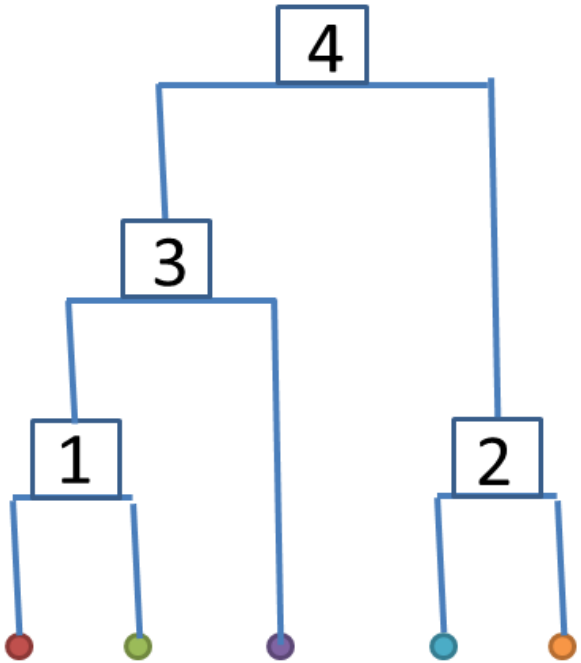
- Produces a nested sequence of clusters.
- This sequence can be represented by a tree :
  - Root : one cluster covering all the data.
  - Leaves : singleton clusters (data points).
- There are two main types of hierarchical clustering methods :
  - Bottom up (agglomerative) clustering (see below) : builds the tree from the bottom level and merges the most similar clusters. The process continues until all data points are in a single cluster.
  - Top down (Divisive) clustering : starts with all the data points in one cluster (the root) and recursively splits clusters until singleton clusters are obtained.

Algorithm

- **Algorithm** *Agglomerative(D)*.
  1. Make each data point in the data set D a cluster.
  2. Compute all pair-wise distances of $x_1, x_2, ..., x_n \in D$.
  3. **Repeat**
     1. Find two clusters that are nearest to each other.
     2. Merge the two clusters form a new cluster c.
     3. Compte the distance from c to all other clusters.
  4. **Until** there is only one cluster left.

Example





References

- Liu B. Web Data Mining. Springer. 2007, 532 pages.
- Witten I. H., Frank E., Hall M. A. Data Mining. Morgan Kaufmann Publishers. 2011, 628 pages.