Clustering

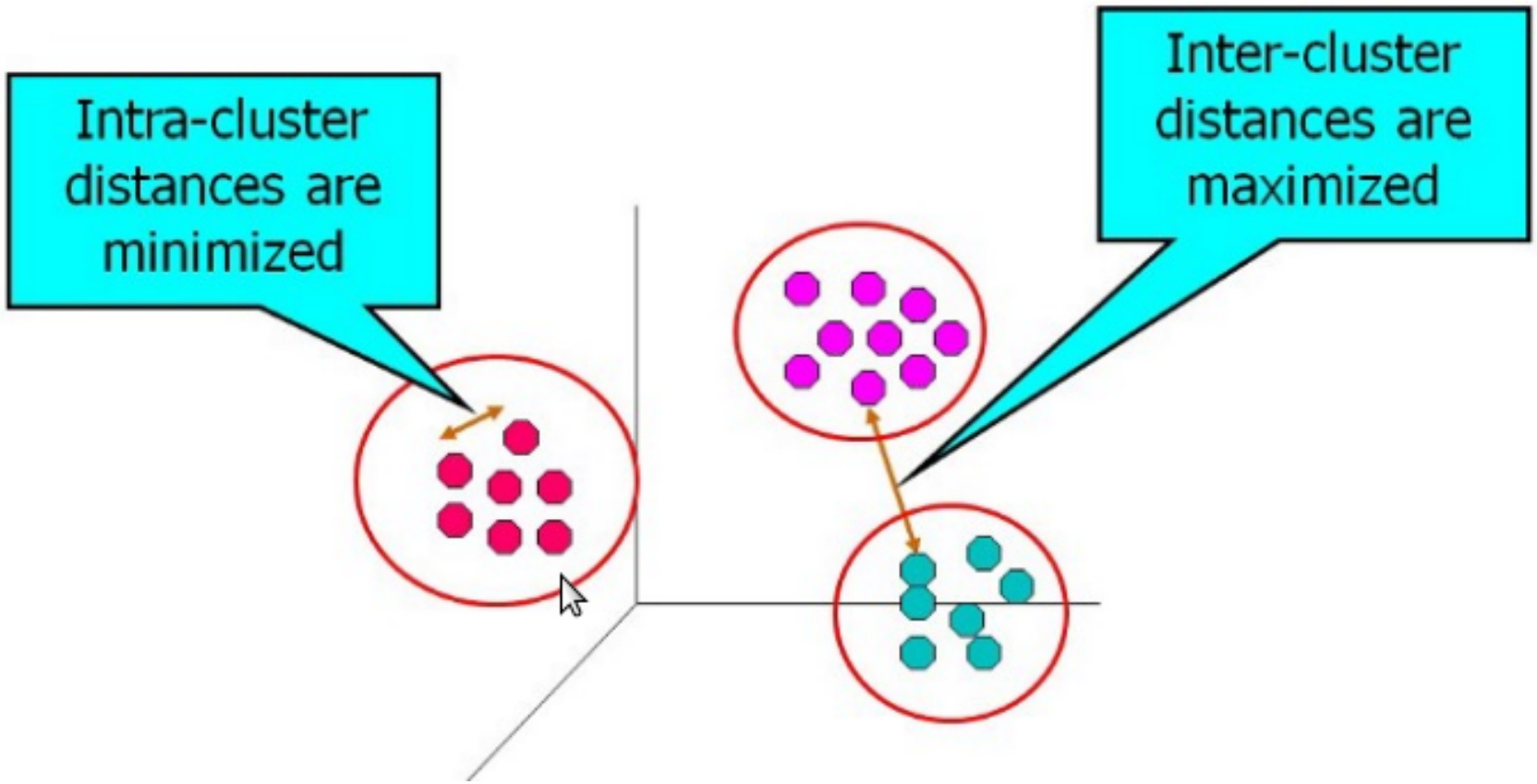
Rakotoarimalala Tsinjo Tony

2023

Définition

- The method of identifying similar groups of data in a dataset is called clustering
- Entities in each group are comparatively more similar to entities of that group than those of the other groups
- Hard Clustering: each data point either belongs to a cluster completely or not
- Soft Clustering: a probability or likelihood of that data point to be in those clusters is assigned.

Goal


Intra-cluster distances are minimized
Inter-cluster distances are maximized

Possible applications

Clustering algorithms can be applied in many fields, for instance:

- Marketing: **finding groups of customers with similar behavior given a large database of customer data containing their properties and past buying records;**
- Biology: **classification of plants and animals given their features;**
- Insurance: identifying groups of motor insurance policy holders with a high average claim cost; identifying frauds;
- City-planning: identifying groups of houses according to their house type, value and geographical location;
- Earthquake studies: clustering observed earthquake epicenters to identify dangerous zones;
- WWW: document classification; clustering weblog data to discover groups of similar access patterns.

Types of clustering algorithms

- Connectivity models: the data points closer in data space exhibit more similarity to each other than the data points lying farther away. Two approaches possible
  - They start with classifying all data points into separate clusters then aggregating them as the distance decreases
  - All data points are classified as a single cluster and then partitioned as the distance increases

  Example: hierarchical clustering algorithm
- Centroid models: These are iterative clustering algorithms in which the notion of similarity is derived by the closeness of a data point to the centroid of the clusters

  Example: K-means

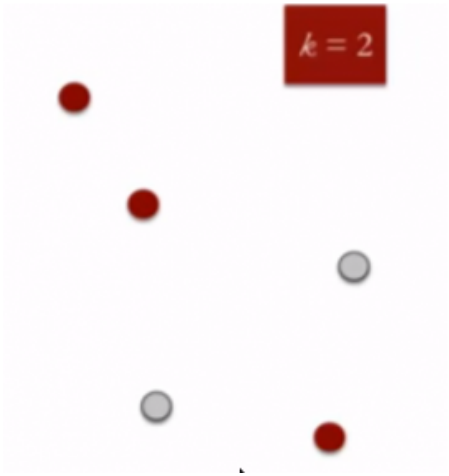
Types of clustering algorithms

- Distribution models: how probable is it that all data points in the cluster belong to the same distribution Example: Expectation-maximization algorithm
- Density Models: search the data space for areas of varied density of data points in the data space. It isolates various different density regions and assign the data points within these regions in the same cluster.

K Means Clustering

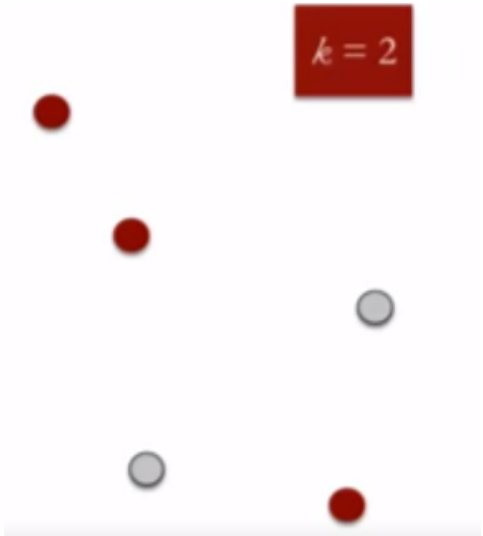
K means is an iterative clustering algorithm that aims to find local maxima in each iteration. This algorithm works in these 5 steps

1) Specify the desired number of clusters K : Let us choose $k = 2$ for these 5 data points in 2-D space.



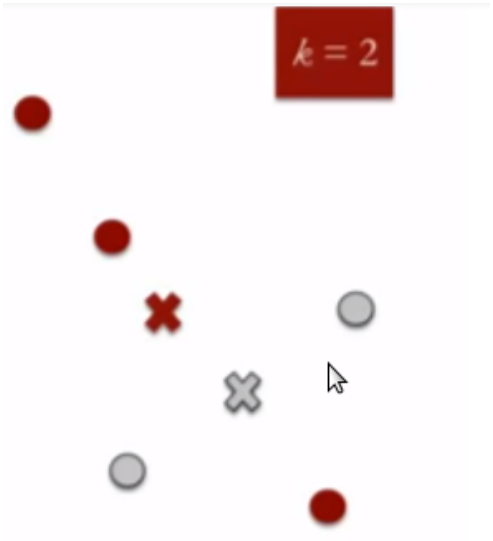
K Means Clustering

2) Specify the desired number of clusters K : Let us choose $k = 2$ for these 5 data points in 2-D space.



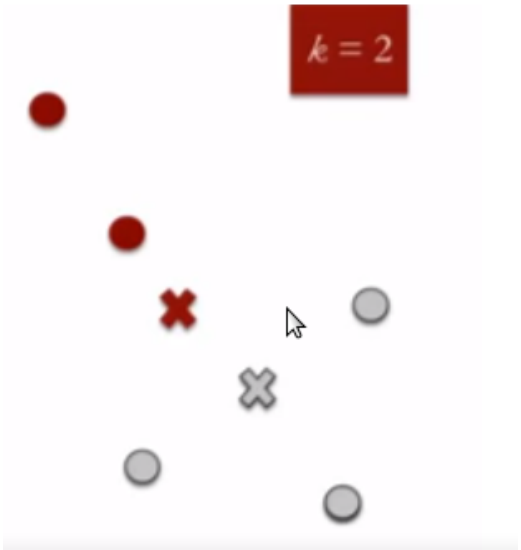
K Means Clustering

3) Compute cluster centroids : The centroid of data points in the red cluster is shown using red cross and those in grey cluster using grey cross.



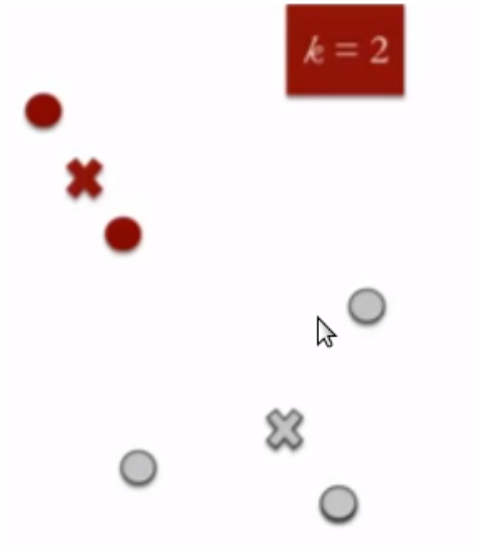
K Means Clustering

4) Re-assign each point to the closest cluster centroid.



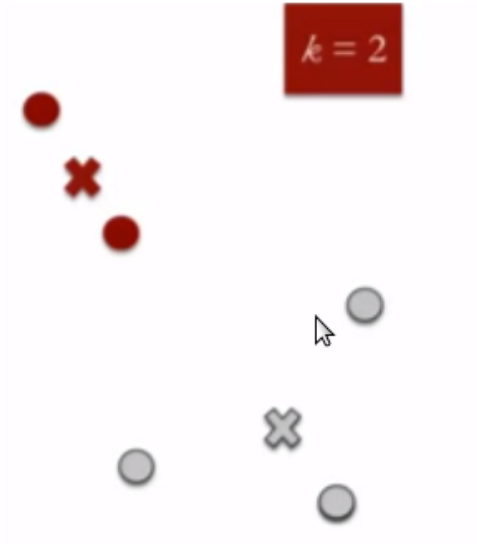
K Means Clustering

5) `Re-compute cluster centroids.`



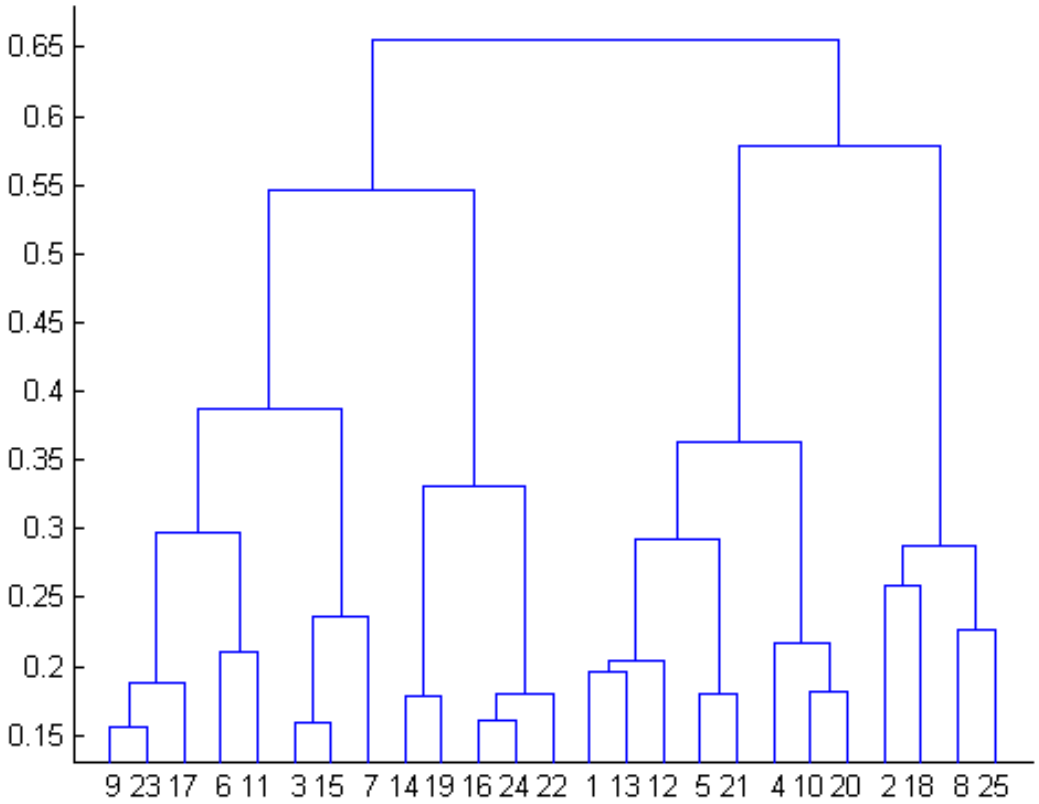
K Means Clustering

6) Repeat steps 4 and 5 until no improvements are possible: When there will be no further switching of data points between two clusters for two successive repeats. It will mark the termination of the algorithm if not explicitly mentioned.



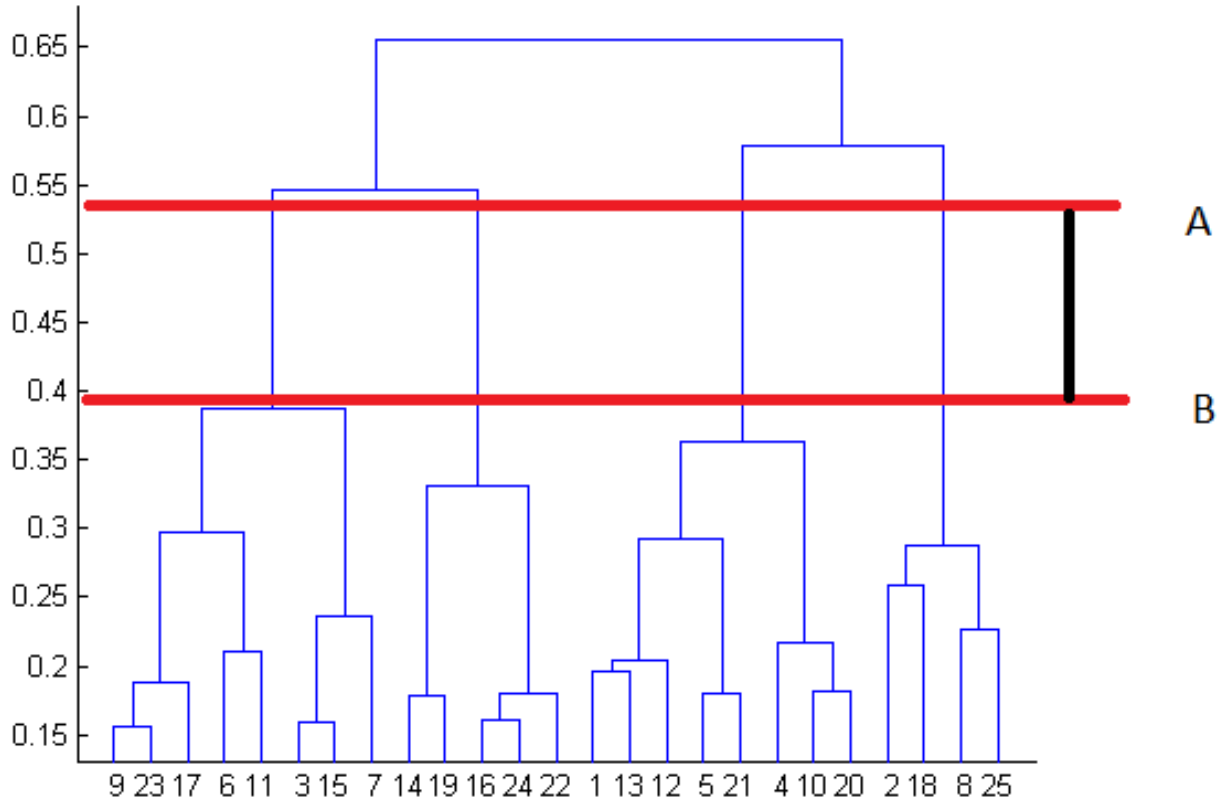
Hierarchical Clustering

This algorithm starts with all the data points assigned to a cluster of their own. Then two nearest clusters are merged into the same cluster. (see dendrogram below)



Hierarchical Clustering

The decision of the no. of clusters that can best depict different groups can be chosen by observing the dendrogram. (figure below shows for 4 clusters).



Hierarchical Clustering

The decision of merging two clusters is taken on the basis of closeness of these clusters. There are multiple metrics for deciding the closeness of two clusters :

- **Minimum step**

$$dissim(C_1, C_2) = \min_{x \in C_1, y \in C_2} dissim_1(x, y)$$

- **Euclidean distance** $||a - b||_2 = \sqrt{\sum (a_i - b_i)^2}$
- **Ward distance**

$$dissim(C_1, C_2) = \frac{n_1 * n_2}{n_1 + n_2} dissim_1(G_1, G_2)$$

where $|C_1| = n_1, |C_2| = n_2$, G_1 centroids of C_1 and G_2 centroids of C_2