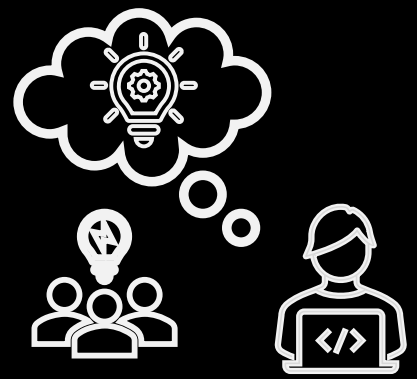
Lab session #4:
Clustering: k-means

Giulia Cisotto

Department of Informatics, Systems and Communication

University of Milan-Bicocca

giulia.cisotto@unimib.it



Motivation

Steps:

1. Run the basic version of k-means with ONE random seed initialization **[TASK 1]**
2. Run multiple runs of the basic k-means **[TASK 2]**
3. Apply k-means++ **[TASK 3]**
4. Pre-process and then cluster **[TASK 4-5]**
5. Clustering for feature selection **[TASK 6]**

MOTIVATION

This fourth lab session aims **to apply k-means algorithm and its variants** to cluster an unknown matrix of data (with low dimensionality and continuous attributes). This lab session refers to Prof. Stella's lecture no.5 "Cluster Analysis: k-means clustering".

You are going to re-use already known packages (matplotlib, seaborn, sklearn.preprocessing…). Check the three previous lab solutions. Moreover, the **sklearn.cluster.Kmeans** package will be introduced to cluster data (see documentation here).

Read the step-by-step instructions below carefully and write your own code to fill the missing steps in the Colab notebook (instructions are also reported in the notebook).

Here is the link to **the Python code @Colab for today**

The **data** to work on will be **available on Moodle** at the beginning to the lab session.

Useful **packages**: numpy, pandas, scipy, matplotlib, seaborn, sklearn, **sklearn.cluster (NEW!)**

Check (and eventually re-use) those ***functions defined in previous solutions*** to compute centroids and find inter-/intra-cluster distances.

Motivation

Steps:

1. Run the basic version of k-means with ONE random seed initialization **[TASK 1]**
2. Run multiple runs of the basic k-means **[TASK 2]**
3. Apply k-means++ **[TASK 3]**
4. Pre-process and then cluster **[TASK 4-5]**
5. Clustering for feature selection **[TASK 6]**