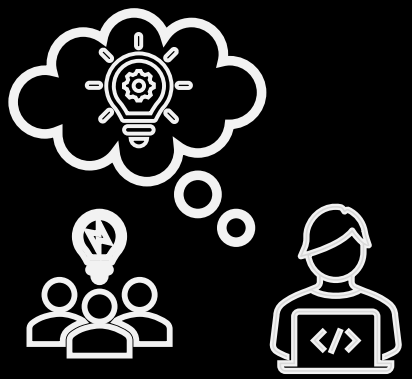
Lab session #5:
Hierarchical Clustering

Giulia Cisotto

Department of Informatics, Systems and Communication

University of Milan-Bicocca

giulia.cisotto@unimib.it



Academic year 2023/2024 – II semester

MOTIVATION

This fifth lab session aims **to apply hierarchical clustering 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.6 "Cluster Analysis: hierarchical clustering".

You are going to (re-)use already known packages (matplotlib, scipy, numpy, seaborn, scikit-learn.preprocessing). Check the previous lab solutions.
Moreover, the **scipy.cluster.hierarchy** package will be introduced to easily cluster data using hierarch ical clustering (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, **scipy.cluster (NEW!)**

Useful Python **data structures**: 2D matrix, list, ndarray

Motivation

Steps:

Load the input data and import useful packages **[TASK 1]**

Prepare the dataset **[TASK 2]**

Design the clustering algorithm **[TASK 3]**

Apply the clustering algorithm to the dataset **[TASK 4]**

Use the clustering solution to form clusters **[TASK 5]**

Compute and visualize the cluster centers **[TASK 6]**

Validation **[TASK 7]**

- compute the inter-cluster distances and the intra-cluster distances
- compute the silhouette score

Motivation

Steps:

Load the input data and import useful packages **[TASK 1]**

Prepare the dataset **[TASK 2]**

Design the clustering algorithm **[TASK 3]**

Apply the clustering algorithm to the dataset **[TASK 4]**

Use the clustering solution to form clusters **[TASK 5]**

Compute and visualize the cluster centers **[TASK 6]**

Validation **[TASK 7]**

- compute the inter-cluster distances and the intra-cluster distances
- compute the silhouette score

This time you will *not* be given with the parameter K, then you need to guess an appropriate no. clusters by analysis.