M.Sc. in Artificial Intelligence for Science and Technology (Unimib-Unimi-Unipv)

Lab session #3:
Clustering: basics

Giulia Cisotto

Department of Informatics, Systems and Communication

University of Milan-Bicocca

giulia.cisotto@unimib.it

Academic year 2023/2024 – II semester

Motivation

Steps:

1. Randomly select features and visualize the reduced dataset **[Part I - TASK 1,2]**
2. Compute proximity matrix, centroids and intra-/inter-cluster distances for the reduced dataset **[Part I - TASK 3,4,5,6]**
3. Properly select features and repeat step 1. and 2. **[Part II - TASK 7, 8]**
4. Scaling and its effect on clustering **[Part III - TASK 9, 10*]**

*optional tasks (if you have time)

Motivation

This third lab session aims to explore the importance of **input preparation** (e.g., normalization/scaling) and **feature selection** to obtain a "good" clustering solution. The **effectiveness of the clustering** is also evaluated. This lab session refers to Prof. Stella's lecture no.3 "Introduction to clustering".

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**

Make your own copy and work on it!

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

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

Motivation

Steps:

1. Randomly select features and visualize the reduced dataset **[Part I - TASK 1,2]**
2. Compute proximity matrix, centroids and intra-/inter-cluster distances for the reduced dataset **[Part I - TASK 3,4,5,6]**
3. Properly select features and repeat step 1. and 2. **[Part II - TASK 7, 8]**
4. Scaling and its effect on clustering **[Part III - TASK 9, 10*]**

*optional tasks (if you have time)

At the lab beginning, you will be given with the data matrix X.

LOAD, REDUCE AND VISUALIZE THE DATASET [TASK 1-2]

The dataset (X) is generated synthetically and is characterized by

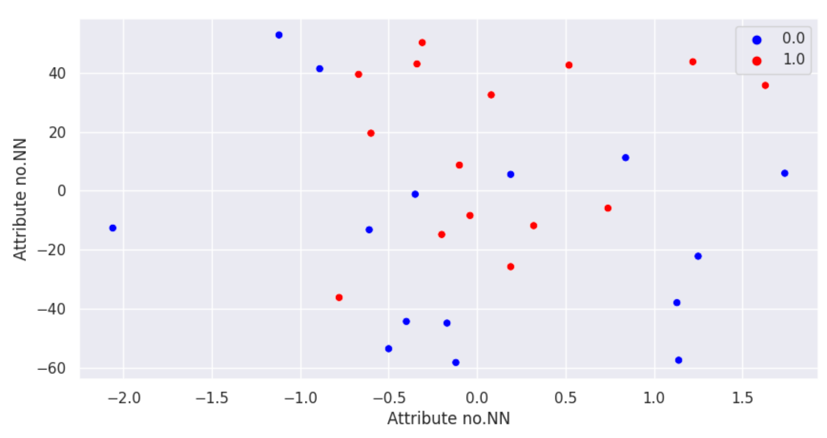
- *N* objects
- *M* attributes
- *K=2* groups (or classes, or categories)

```
The matrix has shape = ...
It has ... objects and ... attributes.
```

Make your choice of the attributes:

```
feat1 =
feat2 =
```

Plot the objects in a scatterplot in 2D using the attributes selected. Hint: use `sns.scatterplot`



Motivation

Steps:

1. Randomly select features and visualize the reduced dataset **[Part I - TASK 1,2]**
2. Compute proximity matrix, centroids and intra-/inter-cluster distances for the reduced dataset **[Part I - TASK 3,4,5,6]**
3. Properly select features and repeat step 1. and 2. **[Part II - TASK 7, 8]**
4. Scaling and its effect on clustering **[Part III - TASK 9, 10*]**

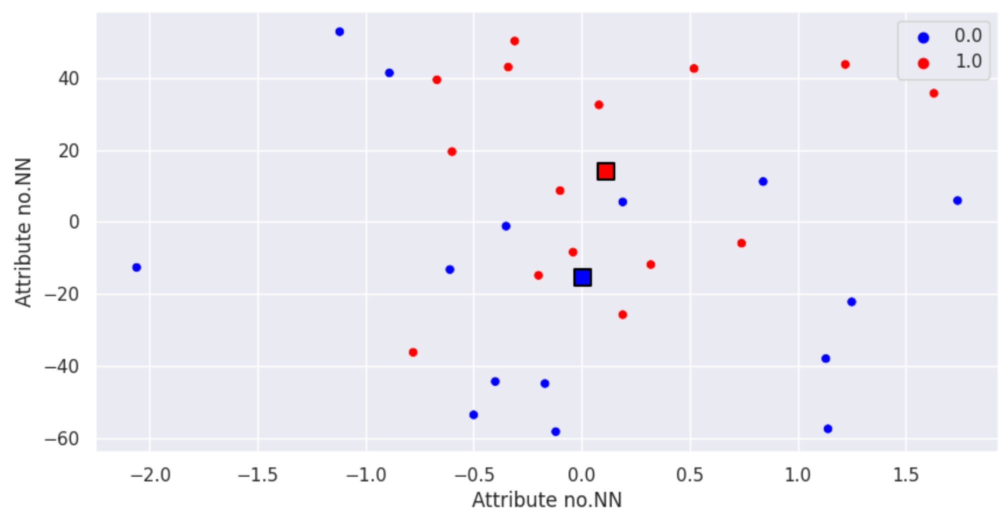
   *optional tasks (if you have time)

COMPUTE PROXIMITY MATRIX AND CENTROIDS [TASK 3-4]

Consider the reduced dataset (`X_red`).

Compute and visualize the proximity matrix, as we did in Lab02, for the reduced dataset (using only 2 attributes). Hint: choose a metric, then use `pdist()` and the `scipy` package.

Compute the centroids, as the mean point of each group, as we did in Lab02. Then, add them to the previous scatterplot.



Note. Pay attention to color/marker style and use always the same (i.e., blue/red for the groups, dots for objects, squares with black borders for centroids).

COMPUTE INTRA- AND INTER-CLUSTER DISTANCES [TASK 5-6]

To compute the intra- and the inter-cluster distances, you may use a number of possible definitions.

In this lab, you will implement the following ones:

- **intra-cluster distance**: *average* distance between the pairwise distances of all objects in the cluster (i.e., group)

- **inter-cluster distance**: distance between centroids

Print the values of both distances.

Finally, answer the question: *"Based on inter-/intra-cluster distances, do you think the found one is a good clustering solution?"*

Motivation

Steps:

1. Randomly select features and visualize the reduced dataset **[Part I - TASK 1,2]**
2. Compute proximity matrix, centroids and intra-/inter-cluster distances for the reduced dataset **[Part I - TASK 3,4,5,6]**
3. Properly select features and repeat step 1. and 2. **[Part II - TASK 7, 8]**
4. Scaling and its effect on clustering **[Part III - TASK 9, 10*]**

   *optional tasks (if you have time)

PROPERLY SELECT FEATURES AND REPEAT TASKS 3-4 [TASK 7-8]

The new reduced dataset (`X_red2`) is composed by all objects with `feat3` and `feat4`.

Compute the centroids, proximity matrix [OPTIONAL], intra- and inter-cluster distances. Then, visualize the new reduced dataset, with colors identifying groups, and the centroids (using a scatterplot with the usual visualization *conventions*).

Print the following information:

- your new choice of the features (`feat3` and `feat4`, this time) and have a motivation
- the coordinates of the new centroids
- the intra- and inter-cluster distances

Motivation

Steps:

1. Randomly select features and visualize the reduced dataset **[Part I - TASK 1,2]**
2. Compute proximity matrix, centroids and intra-/inter-cluster distances for the reduced dataset **[Part I - TASK 3,4,5,6]**
3. Properly select features and repeat step 1. and 2. **[Part II - TASK 7, 8]**
4. Scaling and its effect on clustering **[Part III - TASK 9, 10*]**

   *optional tasks (if you have time)

SCALING AND EFFECT ON CLUSTERING [TASK 9-10*]

Very often, before going to clustering or any other ML-based modelling, input scaling or normalization can be considered (see also Lab02). Here, you will explore the effect of the four main transformations (using sklearn) on this specific input dataset:

sklearn.preprocessing.StandardScaler

sklearn.preprocessing.normalize

sklearn.preprocessing.RobustScaler

sklearn.preprocessing.MinMaxScaler

Note: to use them, use .fit_transform() that computes parameters (e.g., mean, std) during the "fit" phase, and then applies the transformation on the data during the "transform" phase.

SCALING AND EFFECT ON CLUSTERING [TASK 9-10*]

Starting from the reduced dataset obtained in Task 7 (`X_red2`), apply one of the transformations, and obtain the new reduced and *transformed* dataset as `X_red2_transformed`.

Compute the centroids, proximity matrix [OPTIONAL], intra- and inter-cluster distances. Then, visualize the new reduced transformed dataset with centroids, and compare it with the previous one, without transformation (use a scatterplot with the usual visualization *conventions*).

Print/Produce the following information:

- the coordinates of the new centroids
- the two plots to compare the results with/without transformation
- the new proximity matrix
- the new intra- and inter-cluster distances
- repeat the analysis above with another transformation choice

Next Lab on *k-Means*

April, 9th – 2.30 p.m.

Meanwhile, if you have any questions…

use the Lab Forum @eLearning!