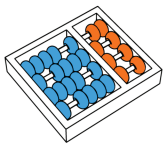
MC886A/MO444A - Machine Learninng
Institute of Computing - Unicamp
First Semester 2021



Prof. Esther Colombini
esther@ic.unicamp.br
PEDs
Alana Correia (alana.correia@ic.unicamp.br
Patrick Ferreira (patrickctrf@gmail.com)

Project 1 - Deadline: 14/04/2021

1 Goal

This assignment aims to apply unsupervised learning methods to solve clustering and dimensionality reduction in two distinct tasks:

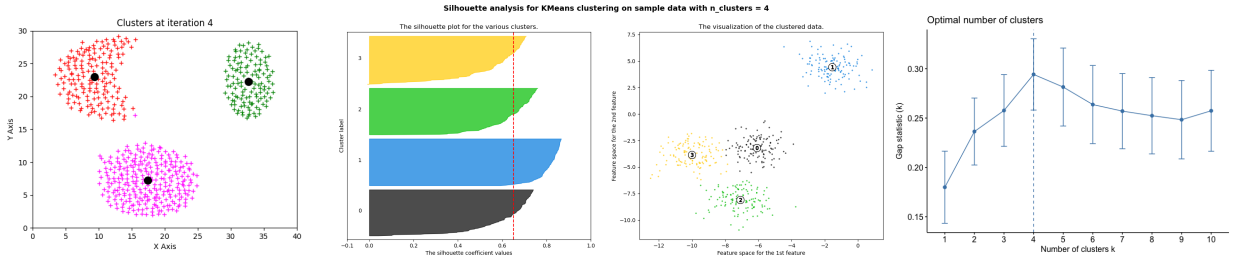
- a standard 2d task
- a task specified by the group with more dimensions. For the task that you select, you are required to:
  - describe the problem addressed, the features used (type/values/description), and the dataset size

2 PART I - Clustering Methods

The work consists of implementing **k-means** and **another unsupervised learning method** of your group's choice. You are not allowed to use a library (like sklearn) that implements the methods at this point. Hence, you are supposed to offer an implementation for a k-means function and a myOtherULMethod function. After implementing your methods, you should:

- Split your data into training/test sets (90/10)
- Apply pre-processing steps over your data that you think necessary (feature scaling, normalization)
- Train the methods in the 2d clustering task with the data provided (cluster.dat)
- Train the methods in the task selected by your group
- Evaluate a different number of clusters and their effect in the tasks to select the best configuration (silhouette coefficient, elbow, among others)
- Use appropriate metrics to evaluate all the experiments (cluster distance, density, etc.)
- Choose your best models and test them with the test set. Show to which cluster the new data has been assigned

You are expected to use graphs and tables to show your results. You can use libraries for this purpose.



3 PART II - Dimensionality reduction

In very high-dimensional spaces, Euclidean distances used by k-means tend to become inflated. To minimize this problem, it is common to use PCA prior to k-means. To assess this method in the task selected by your group, do:

- Run PCA over your data. Consider 3 different energies (variance) for reducing the data dimensionality.
- Run k-means with a different number of clusters
- Compare the results with those achieved in PART I

You do not have to implement PCA. You can use libraries to do it.

4 Programming language

You should use Pyhton as programming language.

5 Evaluation and Discussion

The system should be evaluated according to the quality of the solutions found and a critical evaluation is expected on the relationship between adopted parameters x solution quality. Graphs, tables and images representing the results are expected. Further comparisons with the literature are welcome, although not mandatory.

Please, discuss in the report:

- How/if normalization affected your results
- If the number of clusters achieved is good representatives of your data
- How/if the initialization of cluster centers affected the solution
- The advantages and disadvantages of each method
- How dimensionality reduction affected your results

6 Groups

The groups must be composed of 2 members.

7 Report

The definition of the problem, the solution, and the results obtained must be presented in a report created as a Jupyter notebook. Please, make sure you put the graphs, tables, comparisons, and critical analysis in the notebook. The report should clearly indicate what the contribution of each team member was.

8 Grading

This work will be evaluated according to the following criteria:

- Submission within deadline
- Quality of the solution employed
- Report and discussions
- Code analysis
- Individual student participation in the project

8.1 Penalty policy for late submission

You are not encouraged to submit your assignment after due date. However, in case you did, your grade will be penalized as follows:

- late submission one day after the deadline: grade * 0.75
- late submission two days after the deadline: grade * 0.5
- late submission three days after the deadline: grade * 0.25