ML II unsupervised learning, agents : project

NICOLAS LE HIR
nicolaslehir@gmail.com

TABLE DES MATIÈRES

INTRODUCTION

All processing should be made with python3.

A pdf report is expected in order to present your work. There is no length constraint on the report, you do not need to write more than necessary. The goal of writing a report is that you understand what you did with the project (an also that I understand more easily too and can give you some useful feedback).

The 5 parts of the project are independent.

1 PART 1 : DATA DISTRIBUTION AND THE LAW OF LARGE NUMBERS

The goal of this exercise is to manipulate a data distribution and to get familiar with the law of large numbers in an informal way.

1) Propose a 2-dimensional random variable $Z = (X, Y)$, with X and Y being two real ($\in \mathbb{R}$), discrete or continuous random variables, that are **not** independent. These two variables should represent a quantities of your choice (e.g. the age of the individuals in a population, the color of the eyes of these individuals, ...). Compute the expected value of Z, that must be finite.

2) Sample a number n (of your choice) of points from the law of Z and plot them in a 2 dimensional figure.

3) Compute the empirical average of the first n samples, as a function of the number of samples n and verify that it converges to the expected value, by plotting the euclidean distance to the expected value as a function of n.

Remark : you may use simple laws. You could for instance start with a very simple joint distribution, make everything work, and then explore more complex distributions.

2 PART 2 : METEOROLOGICAL DATA : DIMENSIONALITY REDUCTION AND VISUALIZATION

A meteorological station has gathered data that are in high dimension, thanks to a large number of sensors. The operators of the station would like to predict the risk of a tempest the next day, but first, they need to reduce the dimensionality of the data, in order to apply a supervised learning algorithm on the reduced data.

Perform a dimensionality reduction on the dataset stored in **data/dimreduction**, to a dimension of 2 or 3, and visualize the result by coloring the points as a function of the labels (without using the labels during the dimensionality reduction).

In this exercise, you must compare **two** dimensionality reduction methods and choose one that works better (allows a better separation of the classes in the visualization). You may use libraries such as scikit-learn in order to implement the methods.

You are free to choose the dimensionality reduction methods (linear or nonlinear). **However**, it is required that you explain and discuss your approach in your report. For instance, you could discuss :

— the performance of several methods and models that you tried, in terms of explained variance or of another performance measure.

— the choice of the hyperparameters and the method to tune them.

— the optimization procedures, if relevant.

3 PART 3 : COMPANY CLUSTERING CUSTOMERS

A company has gathered data about its customers and would like to identify similar clients, in order to propose relevant products to new clients, based on their features. This can be represented as a clustering problem.

Perform a clustering on the dataset stored in **data/clustering**, and propose a relevant number of clusters.

In this exercise, you must compare **two** clustering methods and choose one that works better (based on a criterion such as a very clear decrease in the inertia with a given number of clusters, or another criterion). You may use libraries such as scikit-learn in order to implement the methods.

You are free to choose the clustering methods. As usual, it is required that you explain and discuss your approach in your report. For instance, you must discuss :

— the metrics used to compare datapoints (this is very important as it determines what features are most important)

— the performance of several methods and models that you tried, in terms of inertia, knee detection, etc.

— the choice of the hyperparameters and the method to tune them.

— the optimization procedures, if relevant.

4 PART 4 : BANDITS / AGENTS

Exercise to finalize.

5 PART 5 : APPLICATION OF UNSUPERVISED LEARNING

Pick a dataset and perform an unsupervised learning on it. Ideally, your algorithm should answer an interesting question about the dataset. The supervised learning can then be either a **clustering**, a **dimensionality reduction** or a **regression**.

You are free to choose the dataset within the following constraints :

— several hundreds of lines

— at least 6 attributes (columns), the first being a unique `id`

— some features may be categorical (non quantitative).

If necessary, you can tweak an existing dataset in order to artificially make it possible to apply analysis ans visualization techniques. Example resources to find datasets :

— Link 1

— Link 2

— Link 2

— Link 4

You could start with a general analysis of the dataset, with for instance a file **analysis.py** that studies :

— histograms of quantitative variables with a comment on important statistical aspects, such as **means** , **standard deviations** , etc.

— A study of potential **outliers**

— Correlation matrices (maybe not for all variables)

— Any interesting analysis : if you have categorical data, with categories are represented most ? To what extent ?

If the dataset is very large you may also extract a random sample of the dataset to build histogram or compute correlations. You can discuss whether the randomness of the sample has an important influence on the analysis result (this will depend on the dataset).

Whether it is a clustering, a dimensionality reduction or a density estimation, you should provide an **evaluation** of your processing. This can for instance be

— for a clustering, it can be an inertia, a normalized cut...

— for a dimensionality reduction, the explained variance

— for a density estimation, the kullbach leibler divergence between the dataset and a dataset sampled from the estimated distribution

— but you are encouraged to use other evaluations if they are more relevant for your processing.

Short docstrings in the python files will be appreciated, at least at the beginning of each file.

In our report, you could include for instance :

— general informations on the dataset found in the analysis file.

— a potential comparison between several algorithm / models that you explored, if relevant

— a presentation of the method used to tune the algorithms (choice of hyperparameters, cross validation, etc).
— a short discussion of the results

Feel free to add useful visualizations for each step of your processing.

6 THIRD-PARTY LIBRARIES

You may use libraries such as networkx or graphviz, for instance for visualisations of the graph, but not for the algorithmic part that is the subject of the corresponding exercise, unless specified.

7 ORGANISATION

Number of students per group : 3.

Deadline for submitting the project :
— 1st session (December 1st, 2nd) : December 26th.
— 2nd session (February 16th, 17th) : Marth 12th.

The project should be shared through a github repo with contributions from all students. Please briefly indicate how work was divided between students (each student must have contributions to the repository).

Each exercise should be in its own folder.

If you used third-party libraries, please include a **requirements.txt** file in order to facilitate installations for my tests.

https://pip.pypa.io/en/stable/user_guide/#requirements-files

Please don't include the datasets or the project instructions in you repository, add them to your .gitignore.

You can reach me be email if you have questions.