PTML project

NICOLAS LE HIR
nicolaslehir@gmail.com

TABLE DES MATIÈRES

1 OBJECTIVE

The goal of the project is to apply machine learning methods of your choice to real datasets.

2 PART 1 : SUPERVISED LEARNING

Pick a dataset and perform a supervised learning on it. Ideally, your algorithm should answer an interesting question about the dataset.

2.1 Dataset constraints

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`, separated by commas

— you may use some categorical (non quantitative) features.

— some fields should be correlated

If necessary, you can tweak a 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

2.2 Processing

The processing must be made with **python 3**.

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).

The supervised learning part can then be either a **classification** or a **regression**.

In either case, you must provide an **evaluation** of your algorithm. For supervised learning, this could be an average squared error, coefficient of determination (R2 score), etc (`https://scikit-learn.org/stable/modules/model_evaluation.html`).

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

2.3 Report

You must write a pdf report that explains you work. In your report, you should discuss important aspects of your results, such as :

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

— a potential comparison between several algorithm types, if relevant

— the tuning of the algorithms (choice of hyperparameters, cross validation, etc).

— the results

There is no length constraint on the report, it does not need to be very long. The goal of writing a report is that you understand what you did with the project (an also that I understand more easily too).

3 PART 2 : UNSUPERVISED LEARNING

Same instructions, but with an unsupervised learning algorithm. Ideally, use a different dataset than in Part 1.

For unsupervised learning the evaluation of the quality of the result will be different than for supervised learning. It could be the inertia, distorsion, KL divergence, etc.

4 LIBRARIES

For both parts, you may use third-party libraries, but have to briefly explain their usage in your report (choice of the functions, of the hyperparameters, etc).

5 ORGANIZATION

Number of students per group : 4.

Submission deadline : **July 4th 2022**.

The project can be shared through a github repo or a compressed folder, sent by email. Include :

— the pdf report.

— the python files.

Please write "PTML project" in the subject of your email.

You can reach me by email, I will answer faster if you use the gmail address rather than the Epita address.