ML I supervised learning : project

NICOLAS LE HIR
nicolaslehir@gmail.com

TABLE DES MATIÈRES

INTRODUCTION

All processing should be made with python3.

A report must accompany your code, in order to explain and comment it. If you use python scripts, you may write a pdf report. If you use notebooks, you may use markdown inside the notebooks as a report (this is actually preferred over pdf). Short docstring at the top of files and functions will be appreciated, if relevant. There is no length constraint on the report, you do not need to write more than necessary. The goal of writing a report is to help me give you useful feedback.

The 5 parts of the project are independent.

1 PART 1 : ARTIFICIAL DATASET GENERATION

The goal of this exercise is to work with statistical notions such as mean, standard deviation, and correlation.

Write a file or a notebook that generates a numerical dataset with 300 data-points (i.e. lines) and at least 6 columns and saves it to a csv file named **artificial_dataset.csv**.

The columns must satisfy the following requirements :
— they must all have a different mean
— they must all have a different standard deviation (English for "écart type")
— at least one column should contain integers.
— at least one column should contain floats.
— one column must have a mean close to 2.5.
— some columns must be positively correlated.

— some columns must be negatively correlated.

— some columns must have a correlation close to 0.

2 PART 2 : DEFINITION OF A METRIC

A dataset representing a population is stored in **dataset.csv** inside the **project/ex_2_metric/** folder.

Define a **metric** in this dataset, which means define a **dissimilarity** between the samples, by taking into account all their features (columns of the dataset). Remember that you **choose** what is represented by your metric. There is no bad or good metric *per se*.

Some features are numerical and others are categorical, hence you can not use a standard euclidean metric, and you need to define a custom metric, like we did in the **code/metrics/hybrid_data/** exercise during the course. Compute the mean dissimilarity and the standard deviation of the dissimilarity distribution that you obtain, and save the dissimilarity matrix to a file (e.g. a npy file).

Importantly, you must define and explain which features are more important with this metric, since you have to balance the contribution of all the features. Your metric should be meaningful in the sense that not all feature values should induce the same contribution to the dissimilarity : for instance the music style "technical death metal" is closer to "metal" than it is to "classical".

3 PART 3 : PREDICTION OF THE WINNER OF A NBA GAME (CLASSIFICATION)

We would like to predict the winner of a Basketball game, as a function of the data gathered at half-time.

The dataset is stored in **project/ex_3_classification_NBA/** :

— The train and test inputs representing the features are stored in **X_train.npy** and **X_test.npy** respectively.

— If the home team wins, the label is 1, −1 otherwise. The train and test labels are stored in **y_train.npy** and **y_test.npy** respectively.

Your objective is to obtain a mean accuracy superior than 0.84 on the test set, that must not be used for training.

https://scikit-learn.org/stable/modules/model_evaluation.html#accuracy-score

You are free to choose the classification methods, but you must compare at least two methods. Discuss the choice of the optimization procedures, solvers, hyperparameters, cross-validation, etc.

Several methods might work, including some methods that we have not explicitely studied in the class, do not hesitate to try them.

Indication : a solution, with the correct hyperparameters, exists in scikit among the following scikit classes : linear_model.LogisticRegression, svm.SVC, neighbors.KNeighborsClassifier, neural_network.MLPClassifier, ensemble.AdaBoostClassifier.

4 PART 4 : PREDICTION OF THE AMOUNT OF ELECTRICITY PRODUCED (REGRESSION)

We would like to predict the amount of electricity produced by a windfarm, as a function of the information gathered in a number of physical sensors (e.g. speed of the wind, temperature, ...).

The dataset is stored in **project/ex_4 _regression_windfarm/**, similarly to 3 and the instructions are the same (compare at least two estimators and discuss the various settings). Your objective is to obtain a R2 score superior to 0.85 on the test set (that is not used for training).

https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html

Several methods might work, including some methods that we have not explicitely studied in the class. Do not hesitate to try such methods.

Indication : a solution, with the correct hyperparameters, exists in scikit among the following scikit classes : linear_model.Ridge, linear_model.Lasso, neural_network.MLPRegressor, svm.SVR, ensemble.AdaBoostRegressor.

5 PART 5 : APPLICATION OF SUPERVISED LEARNING

Pick a dataset of your choice and perform a classification or a regression on it, in order to study or solve a problem.

Mandatory : Before any processing, **explain very explicitely what problem you are trying to solve, and in particular what quantity you are trying to predict, as a function of which features**.

You are encouraged to compare several estimators / optimization procedures, from different points of view (scoring, computation time, etc).

Suggestion of steps :

— present the dataset shortly in your own words (please do not copy a description from another resource) and link to the url where you downloaded it from.

— provide general analysis of the dataset, that studies its statistical properties, outliers, correlation matrices, or any other interesting analysis. You may produce visualizations.

— if relevant or necessary, preprocess the data, and to justify this preprocessing. You could compare the estimator(s) obtained with and without preprocessing.

— discuss the relevant optimization details : cross validation, hyperparameters, etc

— (mandatory) provide an **evaluation** or multiple evaluations of the obtained estimator(s), thanks to scorings of your choice.

— discuss the results obtained. Have we solved a problem with this processing ?

Some resources to find datasets (but you probably know other good resources already) : Link 1, Link 2, Link 4. If necessary, you can tweak a dataset in order to artificially make it possible to apply analysis ans visualization techniques that you like, or downsample it. This is not a production project and you are encouraged to experiment.

6 THIRD-PARTY LIBRARIES

You may use third-party libraries, but need to slightly explain their usage in your report (choice of hyperparameters, etc.) Please do **not** present elementary python functions like pyplot.plot, etc, or basic library functions like np.random.normal().

7 ORGANISATION

Number of students per group : 3.

Deadline for submitting the project :
— 1st session (November 9, 10th 2023) : December 10th.
— 2nd session (February 15th, 16th 2024) : March 17th 2024.

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

You can reach me be email if you have questions.