ML I supervised learning : 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 : BAYES ESTIMATOR AND BAYES RISK

Step 1 : Propose an applied supervised learning setting, which means :

— an input space $\mathcal{X}$ representing a quantity of your choice. Examples : the age of a person ($\mathcal{X} = \mathbb{R}_+$) , the number of children in a family ($\mathcal{X} = \mathbb{N}$) , the amount of rain in a given time period ($\mathcal{X} = \mathbb{R}_+$). $\mathcal{X}$ might be of dimension > 1.

— an output space $\mathcal{Y}$ representing a quantity of your choice. Examples : the number of children of a person ($\mathcal{Y} = \mathbb{N}$), the fact that an animal is a cheetah ($\mathcal{Y} = \{0, 1\}$), the Body Mass Index ($\mathcal{Y} = \mathbb{R}_+$). $\mathcal{Y}$ must be of dimension 1.

— a random variable (X, Y) with a joint probability distribution, ($X \in \mathcal{X}$ and $Y \in \mathcal{Y}$), representing the relationship between the two quantities. As we have seen during the class, a convenient and natural way to do this is to first set a probabilty law on X and then set a conditional law on Y, conditioned by X.

— a loss function $l(x, y)$ (e.g. squared loss, "0-1" loss, absolute loss, custom loss, ...)

$$l = \begin{cases} \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+ \\ (x, y) \mapsto l(x, y) \end{cases}$$

Compute the Bayes predictor f^* and the Bayes risk R^* associated with this setting.

We recall the definition of the Bayes predictor

$$f^*(x) = \arg\min_{y \in \mathcal{Y}} E[l(Y, y)|X = x] \tag{1}$$

Remark : you have to use a setting different from the settings seen during the course, in terms of input space $\mathcal{X}$ and output space $\mathcal{Y}$. This does not mean that you have to do a very complicated setting.

Step 2 : Run a simulation that computes the empirical risk of the bayes estimator on a dataset sampled according to your setting and verify that when the number of samples is large, the empirical risk is close to the Bayes risk (which is the generalization error ("risque réel")).

Step 3 : Propose an estimator $\tilde{f}$, different from the Bayes estimator :

$$\tilde{f} = \begin{cases} \mathcal{X} \to \mathcal{Y} \\ x \mapsto \tilde{f}(x) \end{cases}$$

and run a simulation that computes its empirical risk on a dataset sampled according to your setting, which gives a statistical approximation of its generalization error (risque réel), and compares it to the Bayes risk.

Remark : the estimated generalization error $\tilde{f}$ must be larger than that of f^*. If not, there is a mistake somewhere.

2 PART 2 : NUMBER OF STREAMS, BAYES RISK AND ABSOLUTE LOSS

A music label is interested in predicting the number of streams of an artist, as a function of the investment. We will consider that the investment is represented by the number of persons who work with the artist, which will allow the label to have an estimation of the number of streams as a function of the budget spent by the label.

More formally, we predict the number of streams of a song on a given streaming platform, during the first month after release, as a function of the number of persons involved in its production.

- The number of streams will be represented by 5 categories : $\mathcal{Y} = \{1, 2, 3, 4, 5\}$. Let S be the exact number of streams. The categories are :
  - $Y = 1$ if $S < 100$
  - $Y = 2$ if $100 \leqslant S < 1000$
  - $Y = 3$ if $1000 \leqslant S < 10000$
  - $Y = 4$ if $10000 \leqslant S < 100000$
  - $Y = 5$ if $100000 \leqslant S$

  We predict the category Y, not the actual number of streams S.
- The number persons in the production team is also represented by a discrete random variable $X \in \mathbb{N}$

Hence, this is a **classification** problem.

The goal of this exercise is to study the impact of the loss function and to verify that different loss functions might lead to different Bayes predictors $f^*(x)$. We will compare the square loss l_2 and the absolute loss $l_1(y, z) = |y - z|$.

We have seen that when the loss used is the squared loss $l_2(y, z) = (y - z)^2$, then the Bayes predictor is the conditional expectation :

$$f^*(x) = E\left[Y|X = x\right] \tag{2}$$

Step 1 : propose a setting (a joint probability distribution on (X, Y) where the Bayes predictor is different for the square loss and for the absolute loss (with a proof). You need to propose a distribution that makes sense with regards to the problem. For instance, the probability that $X = 0$ should be 0.

Step 2 : run a simulation that computes the empirical risks for both losses and corresponding bayes estimators in order to verify the result of the previous question.

Step 3 (optional) : General case : we consider a setting where for each value $x \in \mathcal{X}$, the conditional probability $P(Y|X = x)$ has a continuous density, noted $p_{Y|X=x}$, and that the conditional variable $Y|X = x$ has a moment of order 1. We note that for all $z \in \mathbb{R}$, this implies that $Y - z|X = x$ also has a moment of order 1 .

Determine the Bayes predictor, which means for a fixed x, determine

$$\begin{aligned} f^*(x) &= \arg\min_{z \in \mathbb{R}} E\left[|y - z||X = x\right] \\ &= \arg\min_{z \in \mathbb{R}}(g(z)) \end{aligned} \tag{3}$$

with

$$g(z) = \int_{y \in \mathbb{R}} |y - z| p_{Y|X=x}(y) dy \tag{4}$$

where $g(z)$ is correctly defined, according to the previous assumptions.

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 **data/classification_NBA** :

- The inputs x representing the features are stored in **inputs.csv**.
- The labels y are stored in **labels.csv**

You are free to choose the classification method. **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.
- the choice of the hyperparameters and the method to tune them.
- the optimization procedure.

Your objective should be to obtain a mean accuracy superior than 0.85 on a test set or as a cross validation score.

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

https://scikit-learn.org/stable/modules/cross_validation.html

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 **data/regression_windfarm** :

- The inputs x are stored in **inputs.csv**.
- The labels y are stored in **labels.csv**

The instructions are the same as in 3.

Your objective should be to obtain a R2 score superior to 0.85 on a test set or as a cross validation score.

https://fr.wikipedia.org/wiki/Coefficient_de_d%C3%A9termination

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

5 PART 5 : APPLICATION OF SUPERVISED LEARNING

Pick a dataset and perform a supervised learning on it. Ideally, your algorithm should answer an interesting question about the dataset. The supervised learning can then be either a **classification** 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 classification or a regression, you must provide an **evaluation** of your processing. 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.

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 is the subject of the corresponding exercise, unless specified.

7 ORGANISATION

Number of students per group : 3.

Deadline for submitting the project :

- 1st session (October 13th, 14th) : November 6th.
- 2nd session (January 5th, 6th) : January 29th.

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.