FTML 2025 project

TABLE DES MATIÈRES

INTRODUCTION

The project contains several independent exercises. Some of them are more code oriented (tagged as "C"), some of them more math oriented (tagged as "M"), some are a mix of the two. Code exercises must be made in python 3. You may use libraries. For "M" exercises, formal mathematical proofs are expected.

Important : Please send me an email before may 18th (although the earlier, the better) with your choice of dataset and processing objectives for exercises 6 and 7. The objective is that we validate these objectives and processing together.

0.1 Report

Some form of reporting must accompany your code, in order to explain and comment it. 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 and I have a clear understanding of what you did, so I can more easily give you a useful feedback. General explanations and **conclusions** on your global approach are expected, rather than low level explanations on elementary functions.

— Example of explanations to include in your report : discussion over the choice of models, hyperparameters, and **conclusion / discussion** on which model(s) worked best, or over the choice of preprocessing methods on the datasets (see for instance parts 4 and 5).

— Example of explanations **not** to include in your report : presentation of elementary python function, or library functions from scikit-learn, matplotlib, numpy, or the libraries themselves. You also should not present the algorithms that we studied during the class : for instance, you should not present the concepts of Ridge regression and regularization if you use them.

For exercises 6 and 7, **Please** do not directly copy a text or a direct translation from a tutorial or a kaggle dataset description. Either write it in your own words or point me to the description. For instance, if you write in french (you can write in french or in english, I only write in english in order to avoid having to translate it one day),

please try to avoid rough translation of english idiomatic expressions, leading to sentences like "vous pourriez vouloir utiliser le modèle ... " or "basiquement".

General, important guidelines :

— when you mention the score of an estimator, **always** explicitely mention whether it is a train or test score, or even a cross-validation score.

— never forget to label the axes of plots

— never forget the units of the quantities, when relevant (exercices 6 and 7).

The preferred format is using one notebook per exercise, with markdown comments explaining your approach. If you use python scripts, you may also write a pdf report. Short docstring at the top of files and functions will be appreciated, if relevant.

0.2 Organization

Number of students per group : 3 or 4.

It is possible to mix NLP and CV within the groups.

Submission deadline : Sunday **June 29th 2024**.

The project must be shared through a git repo, sent by email (Please write "FTML project" in the subject of your email) to nicolaslehir@gmail.com. Each exercise should be in its own folder. Please include a **requirements.txt** for the necessary libraries, **if these libraries are not already included in the requirements of the course**. You can reach me by email (at the same address) if you have questions about the project.

1 BAYES ESTIMATOR AND BAYES RISK

Question 1 (M) : Propose a supervised learning setting :

— input space $\mathcal{X}$, representing a physical quantity of your choice, with units.

— output space $\mathcal{Y}$, representing a physical quantity of your choice, with units.

— a random variable (X, Y) with a joint distribution.

— a loss function $l(x, y)$

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

Compute the Bayes predictor $f^* : \mathcal{X} \to \mathcal{Y}$ and the Bayes risk associated with this setting. **Remark :** you have to use a setting different than the settings seen during the course, in terms of input space $\mathcal{X}$ and output space $\mathcal{Y}$. However, you can use any classical loss function l (square loss, "0-1" loss, etc).

Question 2 (C) : propose an estimator $\tilde{f} : \mathcal{X} \to \mathcal{Y}$, different than the Bayes estimator and run a simulation that gives a statistical approximation of its generalization error (risque réel) by computing its empirical risk on a test set. Perform the same simulation also for f^*, and verify that the generalization error is smaller for f^* than for $\tilde{f}$, and that your the test error for f^* is close to the Bayes risk (this should be the case if the number of samples is sufficiently large).

2 BAYES RISK WITH ABSOLUTE LOSS

The questions of this exercise can be done in whatever order.

We consider a regression problem with output space $\mathcal{Y} = \mathbb{R}$ and some input space $\mathcal{X}$. We have seen that when the loss function used is the squared loss $l_{\text{squared}}(y, z) = (y - z)^2$, then the Bayes predictor $f^*_{l_{\text{squared}}}$ is the conditional expectation of Y given x, for each value of $x \in \mathcal{X}$.

$$f^*_{l_{\text{squared}}}(x) = \mathbb{E}\left[Y|X = x\right] \tag{1}$$

The goal of this exercise is to determine the Bayes predictor for the absolute loss, noted $f^*_{l_{\text{absolute}}}(x)$, which means that instead of using the squared loss l_{squared}, we use the absolute loss $l_{\text{absolute}}(y, z) = |y - z|$. We note $R_{l_{\text{squared}}}(g)$ and $R_{l_{\text{absolute}}}(g)$ the risk (generalization error, risque réel) of an estimator g for the l_{squared} and l_{absolute} losses respectively.

Question 0 (M, to avoid making a mistake later) : propose a function from $\mathbb{R}$ to $\mathbb{R}$ that has a zero derivative a some real value x_0, but $f(x_0)$ is not a local extremum of the function.

Question 1 (M + C) : propose a setting where $f^*_{l_{\text{absolute}}} \neq f^*_{l_{\text{squared}}}$. To show this, several options are possible. For instance, you might find an estimator h such that $R_{l_{\text{absolute}}}(h) < R_{l_{\text{absolute}}}(f^*_{l_{\text{squared}}})$, or the opposite.
Run simulations that verify your result by computing empirical test errors, as in exercice 1.

Question 2 (M) : 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 for the absolute loss, which means for a fixed x, determine

$$\begin{aligned} f^*_{l_{\text{absolute}}}(x) &= \arg\min_{z \in \mathbb{R}} \mathbb{E}\left[|y - z||X = x\right] \\ &= \arg\min_{z \in \mathbb{R}} (g(z)) \end{aligned} \tag{2}$$

with

$$g(z) = \int_{y \in \mathbb{R}} |y - z| p_{Y|X=x}(y) dy \tag{3}$$

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

3 EXPECTED VALUE OF EMPIRICAL RISK FOR OLS

3.1 Reminders of the OLS setting

We recall the Ordinary least squares (OLS) problem and notations :

— $\mathcal{X} = \mathbb{R}^d$ (input space)

— $\mathcal{Y} = \mathbb{R}$ (output space)

— squared loss :

$$l(y, y') = (y - y')^2$$

— hypothesis space :

$$F = \{x \mapsto x^T\theta, \theta \in \mathbb{R}^d\}$$

θ^T is the transposition of θ.

As usual, the dataset is stored in the **design matrix** $X \in \mathbb{R}^{n\times d}$.

$$X = \begin{pmatrix} x_1^T \\ \dots \\ x_i^T \\ \dots \\ x_n^T \end{pmatrix} = \begin{pmatrix} x_{11}, \dots, x_{1j}, \dots x_{1d} \\ \dots \\ x_{i1}, \dots, x_{ij}, \dots x_{id} \\ \dots \\ x_{n1}, \dots, x_{nj}, \dots x_{nd} \end{pmatrix}$$

The vector of predictions of the estimator writes $X\theta$. Hence the empirical risk $R_n(\theta)$ writes

$$\begin{aligned} R_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta^T x_i)^2 \\ &= \frac{1}{n}\|y - X\theta\|_2^2 \end{aligned} \tag{4}$$

With $y = (y_1, \dots, y_n)^T$ being the vector containing the labels. $R_n(\theta)$ is a random variable that depends on y, X, θ.

We assume that X is **injective.** Necessary, $d \leqslant n$. As we have seen in the class, the ordinary least squares estimator, that minimizes the empirical risk, given X and y, is defined as :

$$\hat{\theta} = (X^TX)^{-1}X^Ty \tag{5}$$

3.2 Statistical setting

3.2.1 *Linear model*

In the **linear model**, we assume that

$$y = X\theta^* + \epsilon \tag{6}$$

where ϵ is a vector of centered Gaussian noise with variance matrix $\sigma^2 I_n$. Equivalently this can be written in the following formulation :

$$y_i = \theta^{*T}x_i + \epsilon_i, \forall i \in [1, n]$$

and ϵ_i is a centered noise (or error) ($E[\epsilon_i] = 0$) with variance σ^2. The noise is independent for all i. Hence, both y and $\hat{\theta}$ are random variables and depend on ϵ.

3.2.2 *Fixed design*

In the **fixed design** setting, X is **deterministic and fixed.**. Although not necessary totally intuitive, this setting is convient for this kind of theoretical studies, because it simplifies some computations. We note that the same study could be done for random designs, meaning X is also random, but this leads to more involved computations. In this setting, the only source of randomness is ϵ, and all expectated values are computed with respect to ϵ (or equivalently, to y) and / or to θ. In this setting, given θ, we define the **fixed design risk** of θ.

$$\begin{aligned} R_X(\theta) &= E_y[\frac{1}{n}\sum_{i=1}^{n}(y_i - \theta^T x_i)^2] \\ &= E_y[\frac{1}{n}\|y - X\theta\|_2^2] \\ &= E_y[R_n(\theta)] \end{aligned} \tag{7}$$

In 7, the expectation is with respect to $y = (y_1, \dots, y_n)^T$. If θ is a random variable, then this quantity is itself a random variable as well.

3.3 Objective

We want to show that in the linear model, fixed design we have

Proposition 1.

$$E[R_X(\hat{\theta})] = \frac{n-d}{n}\sigma^2 \tag{8}$$

In this expression, both y and $\hat{\theta}$ are random variables, that are not independent : since $\hat{\theta}$ is the OLS estimator, $\hat{\theta}$ depends on y. The expectation is over the distribution of both variables.

3.4 Exercise

We note $\|.\| = \|.\|_2$.

Question 1 (M) : Compare 1 to the Bayes risk. Which value is smaller ? Interpret this result and discuss it as a function of n and d.

Question 2 (M) : Show that :

$$E\left[R_n(\hat{\theta})\right] = E_\epsilon\left[\frac{1}{n}\|(I_n - X(X^TX)^{-1}X^T)\epsilon\|^2\right] \tag{9}$$

where E_ϵ means that the expected value is over ϵ.

Question 3 (M) : Let $A \in \mathbb{R}^{n,n}$. Show that

$$\sum_{(i,j)\in[1,n]^2} A_{ij}^2 = tr(A^TA) \tag{10}$$

Question 4 (M) : Show that

$$E_\epsilon\left[\frac{1}{n}\|A\epsilon\|^2\right] = \frac{\sigma^2}{n}tr(A^TA) \tag{11}$$

Question 5 (M) : We note

$$A = I_n - X(X^TX)^{-1}X^T \tag{12}$$

Show that

$$A^TA = A \tag{13}$$

Question 6 (M) : Conclude.

3.5 Simulation

Question 7 (M) : Still in the same setting, what is the expected value of $\frac{\|y-X\hat{\theta}\|_2^2}{n-d}$?

Question 8 (C) : Produce a numerical simulation that estimates σ^2 thanks to the result of step 6. Check that the result is consistent with the theoretical value you have chosen.

4 REGRESSION ON A GIVEN DATASET

("C" exercise) Perform a regression on the dataset stored in **FTML/Project/data/regression/**. You are free to choose the regression methods, but you must compare at least two methods. You can do more than 2 but this is not mandatory for this exercise. Discuss the choice of the optimization procedures, solvers, hyperparameters, cross-validation, etc. The Bayes estimator for this dataset and the squared loss reaches a R2 score of approximately 0.92, for at least 1 of the 2 estimators (1 estimator is enough).

Your objective is be to obtain a R2 score superior than 0.88 on the test set, that must **not** be used during training. Remember that training is the complete model optimisation procedure, including model selection and hyperparameters testing, not only when you call a **.fit()** method ! This is the topic that we discussed during the practical sessions on train / validation / test and cross-validation. However, since you have the test set, all you can do is "pretend" not to use it during training, since you can always compute the score test several times without putting it in your solution.

5 CLASSIFICATION ON A GIVEN DATASET

("C" exercise) Same instructions as in 4, except that this time a classification has to be performed and the data and the dataset is stored in **FTML/Project/data/classification/**. Your objective should be to obtain a mean accuracy superior to 0.85 on the test set (same remark about the test set).

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.

6 APPLICATION OF SUPERVISED LEARNING

("C" exercise) Pick a dataset of your choice and perform a classification or a regression on it, in order to study or solve a problem (of your choice as well).

Mandatory (very important) : before any processing, you must fulfil these two 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. On the one hand, it is very important to present the features of this dataset that are not obvious for someone (me) that is not necessary familiar with the dataset. On the other hand, please do not present a feature called "weight in kilos".
- **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**. If relevant, discuss why solving this problem would be interesting or have a value for an industry.

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

this course is dedicated to discovering and exploring some of the many principles of machine learning, rather than being a production-oriented course. Hence, you are encouraged to explore original and personal approaches in this exercise (and the next one 7). It is not a huge deal is the final scores are not outstanding, if you took the chance to explore a custom approach and learned a new possible method.

Suggestion of steps :

— 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 optimization details : cross validation, hyperparameters, etc

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

— (**mandatory**) 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 3. 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.

The following datasets are **not** accepted :

— red wine quality : `https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009`

— mall customers : `https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutoria`

— country data : `https://www.kaggle.com/datasets/rohan0301/unsupervised-learning-on-country-d`

7 APPLICATION OF UNSUPERVISED LEARNING

("C" exercise) Same instructions, but with an unsupervised learning, which will most likely mean a clustering, a dimensionality reduction or a density estimation.

As in 6, it is very important to explicitely present the dataset and the problem you work on, and ideally to discuss why solving this problem could be useful. Discuss the important algorithmic details : for instance, if you perform a clustering, you must discuss the **metric** used. Also, it is important to discuss the results obtained (for instance by interpreting the clusters obtained in the case of a clustering). An **evaluation** of the algorithm is also expected. You must choose a scoring adapted to the problem. To summarize, the processing should be **meaningful**.