FTML 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.

2023-05-21 : the datasets for exercises 4 and 5 will be pushed shortly.

0.1 Report

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). However, docstrings are not considered to be a report (although short, useful docstrings at the top of files or functions will be appreciated). 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.

Please do not directly copy a text from a tutorial or a kaggle dataset description. Either write it in your own words or point me to the description.

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 25th 2023**.

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. 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 $\mathcal{X}$
- output space $\mathcal{Y}$
- 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 computation in question 1 was correct (the test error for f^* should be close to the Bayes risk if there is a sufficiently large number of samples).

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 $\mathcal{Y} = \mathbb{R}$. We have seen that when the loss function used is the squared loss $l_2(y, z) = (y - z)^2$, then the Bayes predictor is the conditional expectation of Y given x, for each value of $x \in \mathcal{X}$ ($\mathcal{X}$ is the input space).

$$f^*_{l_2}(x) = E\big[Y|X = x\big] \tag{1}$$

The goal of this exercise is to determine $f^*_{l_1}(x)$ when instead of using the squares loss l_2, we use the absolute loss $l_1(y, z) = |y - z|$. We note $R_{l_2}(g)$ and $R_{l_1}(g)$ the risk (generalization error, risque réel) of an estimator g for the l_2 and l_1 losses respectively.

Question 1 (M + C) : propose a setting where $f^*_{l_1} \neq f^*_{l_2}$. To show this, several options are possible. For instance, you might find an estimator h such that $R_{l_1}(h) < R_{l_1}(f^*_{l_2})$, or the opposite.

Run simulations that verify your results by computing empirical test errors, as in section 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, which means for a fixed x, determine

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

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.

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 :

$$\hat{\theta} = (X^T X)^{-1} X^T y \tag{5}$$

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.** Hence, now all expectations are with respect to ϵ (or equivalently, to y) and to θ. In this setting, given θ, we define the **fixed design risk**.

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

In 7, the expectation is with respect to $y = (y_1, \dots, y_n)^T$. This quantity is itself a random variable that depends on θ.

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, as $\hat{\theta}$ is the OLS estimator. The expectation is over the distribution of both variables.

3.4 Exercise

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

Question 1 (M) : Show that :

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

where E_ϵ means that the expected value is over ϵ.

Question 2 (M) : Let $A \in \mathbb{R}^{n,n}$. Show that

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

Question 3 (M) : Show that

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

Question 4 (M) : We note

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

Show that

$$A^TA = A \quad (13)$$

Question 5 (M) : Conclude.

3.5 Simulation

Question 6 (M) : Still in the same setting, what is the expected value of $\frac{||y-X\hat{\theta}||_2^2}{n-d}$?

Question 7 (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. Discuss the choice of the optimization procedures, solvers, hyperparameters, cross-validation, etc. The Bayes estimator for this dataset and the square loss reaches a R2 score of approximately 0.88. Your objective should be to obtain a R2 score superior than 0.84 on the test set (that must **not** be used during training).

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 than 0.85 on the test set (that must **not** be used during training).

6 APPLICATION OF SUPERVISED LEARNING

("C" exercise) Pick a dataset and perform a classification or a regression on it.
Important : please try to justify your processing. Why is it be interesting to do a regression or a classification on this dataset ?

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 (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.

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.

7 APPLICATION OF UNSUPERVISED LEARNING

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

Important : the same remarks apply, you want to justify your processing. Why would it be interesting and **meaningful** to do a clustering / dimensionality reduction / density estimation ? Also, it is important to discuss the results obtained. For instance : can we interpret the clusters obtained ? An **evaluation** of the algorithm is also expected. You must choose a scoring adapted to the problem.

8 BONUS : OPTIMIZATION LOWER BOUND

(This exercise is not mandatory but might give extra points). Complete the different steps in part 3 of the practical session 7 "lower bound on the performance of gradient-based algorithms".

("M" exercise)