FTML project

TABLE DES MATIÈRES

1 BAYES ESTIMATOR AND BAYES RISK

Question 1 : 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 $l(x, y)$

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

Compute the Bayes predictor and the Bayes risk associated with this setting. (Use a setting different that the ones used in the exercises done during the course.)

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 than the settings seen during the course.

Question 2 : propose an estimator $\tilde{f}$, different than the Bayes risk :

$$\tilde{f} = \left\{ \begin{array}{l} \mathcal{X} \to \mathcal{Y} \\ x \mapsto \tilde{f}(x) \end{array} \right.$$

and run a simulation that gives a statistical approximation of its generalization error (risque réel), and compares it to the Bayes risk.

2 BAYES RISK WITH ABSOLUTE LOSS

We consider a supervised regression problem, $\mathcal{Y} = \mathbb{R}$. We have seen that when the loss used is the square loss $l_2(y, z) = (y - z)^2$, then the Bayes predictor is the conditional expectation :

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

The goal of this exercise is to determine $f^*(x)$ in a different situation where instead of using the square loss l_2, we use the absolute loss $l_1(y, z) = |y - z|$.

Question 1 : propose a setting where the Bayes predictor is different for the square loss and for the absolute loss.

Question 2 : 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 EXPECTED VALUE OF EMPIRICAL RISK

3.1 Reminder of the OLS setting

We consider the Ordinary least squares problem, and its

- $\mathcal{X} = \mathbb{R}^d$
- $\mathcal{Y} = \mathbb{R}$
- square 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 $Y = X\theta$. Hence,

$$R_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta^T x_i)^2$$
$$= \frac{1}{n}||Y - X\theta||_2^2$$

We note $||.|| = ||.||_2$. 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, is defined as :

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

3.2 Reminder on statistical analysis

We want to show the following proposition, that deals with the statistical analysis of the empirical risk, with the linear model and fixed design (same setting as the one seen during the class).

Proposition 1. *The expected value of the empirical risk of $\hat{\theta}$ writes :*

$$E[R_n(\hat{\theta})] = \frac{n-d}{n}\sigma^2 \tag{6}$$

Remark. *In this expected value, both the dataset and $\hat{\theta}$ are random variables, so the expectation is over the distribution of both variables.*

In the linear model, fixed design, we assume that

$$Y = X\theta^* + \epsilon \tag{7}$$

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

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

and ϵ_i is a centered noise (or error) ($E[\epsilon_i] = 0$) with variance σ^2. The noise is independent for all i.

As $\hat{\theta}$ depends on Y, both Y and $\hat{\theta}$ depend on ϵ.

3.3 Exercise

Step 1 : Show that :

$$E\Big[R_n(\hat{\theta})\Big] = E_\epsilon\left[\frac{1}{n}||\big(I_n - X(X^T X)^{-1} X^T\big)\epsilon||^2\right] \tag{8}$$

where E_ϵ means that the expected value is over ϵ.

Step 2 : Let $A \in \mathbb{R}^{n,n}$. Show that

$$\sum_{(i,j)\in[1,n]^2} A_{ij}^2 = tr(A^T A) \tag{9}$$

Step 3 : Show that

$$E_\epsilon\left[\frac{1}{n}||A\epsilon||^2\right] = \frac{\sigma^2}{n} tr(A^T A) \tag{10}$$

Step 4 : We note

$$A = I_n - X(X^T X)^{-1} X^T \tag{11}$$

Show that

$$A^T A = A \tag{12}$$

Step 5 : Conclude.

3.4 Simulation

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

Step 7 : 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

Perform a regression on the dataset stored in **FTML/Project/data/regression/**.

— The inputs x are stored in **inputs.npy**.

— The labels y are stored in **labels.npy**

You are free to choose the regression 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 that you tried.

— the choice of the hyperparameters and the method to choose them.

— the optimization method.

We have seen several types of regressors during the class.
You may use libraries, but if you do so, it is required that you explain their usage in your report.

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

5 CLASSIFICATION

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/**.
We have seen several types of classifiers during the class.
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

6 ORGANIZATION

Number of students per group : 4.

Submission deadline : **June 17th 2022**.

The project must be shared through a github repo, sent by email with contributions from all students. The repo should contain :

— the pdf report.

— the python files for the regression problem 4.

— the python files for the classification problem 5.

Please write "FTML 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.