Machine Learning **2021-2022**
4A-MA

Exam May, 24th, 2022

Duration : 2 hours
The following documents are authorized : course handout, course notes.
Cell phones are not permitted.
The exercises are all independent.
Approximate scale : 5-5-4-6

Exercise 1.

1. We have a learning sample $(X_i, Y_i)_{1\leq i\leq n}$, xhere $X_i \in \mathbb{R}^p$ and $Y_i \in \mathbb{R}$ and we consider le linear model
$$Y_i = X_i w + \varepsilon_i,\ 1 \leq i \leq n,$$
where $X_i \in \mathbb{R}^p$ is a row vector and $w \in \mathbb{R}^p$ is a column vector.
We introduce the following three estimators of the vector w, for a parameter $\lambda > 0$:
$$\hat{w}^1 = \text{argmin}_{w\in\mathbb{R}^p} \sum_{i=1}^{n} (Y_i - X_i w)^2 \tag{1}$$
$$\hat{w}^{2,\lambda} = \text{argmin}_{w\in\mathbb{R}^p} \sum_{i=1}^{n} (Y_i - X_i w)^2 + \lambda \sum_{j=1}^{p} w_j^2 \tag{2}$$
$$\hat{w}^{3,\lambda} = \text{argmin}_{w\in\mathbb{R}^p} \sum_{i=1}^{n} (Y_i - X_i w)^2 + \lambda \sum_{j=1}^{p} |w_j| \tag{3}$$
   (a) Specify the name of each of these estimators.
   (b) In which case(s) does one have an explicit solution ? Specify the solution in this (these) case(s) by defining all the notations you will introduce.
   (c) If we increase the value of λ in cases (2) and (3), what happens for the bias of the estimators $\hat{w}^{2,\lambda}$ and $\hat{w}^{3,\lambda}$? what happens for their variance ?
   (d) We assume that $p = 7$. We have reported in the columns A, B ou C the three estimators $\hat{w}^1$, $\hat{w}^{2,\lambda}$ and $\hat{w}^{3,\lambda}$ in a disordered manner. Specify which column corresponds to $\hat{w}^1$, to $\hat{w}^{2,\lambda}$ and to $\hat{w}^{3,\lambda}$. Explain your result.

	A	B	C
w_1	0.38	0.50	0.60
w_2	0.23	0.20	0.30
w_3	-0.02	0.00	-0.10
w_4	0.15	0.09	0.20
w_5	0.21	0.00	0.30
w_6	0.03	0.00	0.20
w_7	0.12	0.05	0.26

Exercise 2.

1. A group of project students worked on a binary classification problem. Their classification algorithm has a performance of 98% of well classified on a test sample. They claim that their result is excellent. Is this necessarily the case ? Explain why.

2. Instead of estimating the generalization error on a test sample, a Monte Carlo method is sometimes used. Explain the principle of this method and its advantages and disadvantages.
3. Specify the effect that the following operations may have on the bias and variance of the model by filling in the blanks with "increases" or "decreases" or "does not change", justifying your answers :
   (a) If we increase k in a k nearest neighbors algorithm, the bias . . . . . . and the variance . . . .
   (b) If we prune a CART tree, the bias . . . . . . and the variance . . . .
   (c) If we remove the vectors that are not support vectors in an SVM model, the bias . . . . . . and the variance . . . .
   (d) If we increase the number of hidden layers in a neural network (the size of each layer being fixed), the bias . . . . . . and the variance . . . .

Exercise 3.

The following figure gives the graph of the residuals obtained in a regression problem respectively with a linear model, with a CART algorithm and with a Support Vector Regression algorithm, in a disordered way. Specify to which algorithm the left, middle and right graphs correspond, explaining your result.

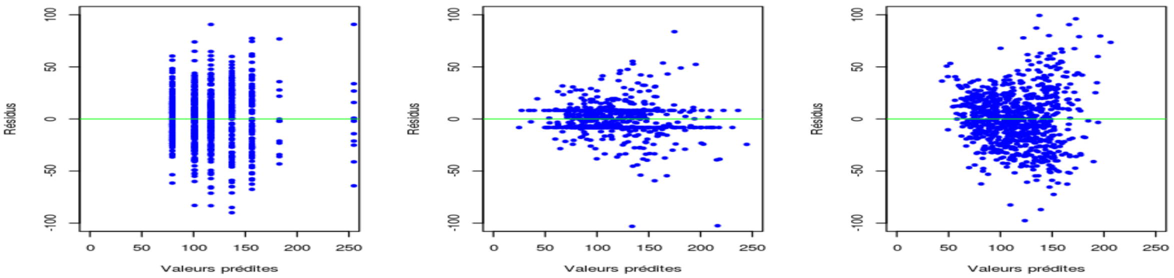

FIGURE 1 – Residuals versus predicted values for 3 models

Exercise 4. Let $\theta > 0$ be given. Consider a binary classification problem for which a random pair (X, Y) with value in $\mathbb{R}^+ \times \{0, 1\}$ satisfies the following properties :

— X ollows a uniform distribution on $[0, 2\theta]$

— For all $x \in [0, 2\theta]$, $P(Y = 1/X = x) = \frac{x}{x+\theta}$ (which means that $E(\mathbb{1}_{Y=1}/X) = \frac{X}{X+\theta}$).

1. Calculate for all $x \in [0, 2\theta]$, $P(Y = 0/X = x)$.
2. Determine the Bayes classifier g^* for this model. (You can use the expression for g^* directly without proving it).
3. Let A be some interval included in $[0, 2\theta]$, prove that
$$P(X \in A, Y = 1) = E\left(\mathbb{1}_{X\in A}\frac{X}{X+\theta}\right).$$
4. Give the expression of the density function of X, then express in the form of an integral the quantity $P(X \in A, Y = 1)$. Do the same for $P(X \in A, Y = 0)$.
5. Calculate the Bayes error $L^* = P(Y \neq g^*(X))$.