Regression and Classification, EMINES. *Examination, 2024. 3 hours*

Exercise 1 (A classical optimization property.) Let $h(\beta) = (\beta - u)^2 + c|\beta|$ where $c \geqslant 0$, $u \in \mathbb{R}^*$ and $\beta \in \mathbb{R}$.

1. Show that there exists a unique minimum for h that is attained on some $\beta^\star \in \mathbb{R}$.
2. Deduce that
$$\beta^\star = u\left(1 - \frac{c}{2|u|}\right)^+$$

Solution.

1. The function h is strictly convex and $\lim_{\beta\to\pm\infty} |h(x)| = \infty$. This implies that h admits a unique minimizer $\beta^\star$.
2. **Case 1** $\beta^\star \neq 0$, in which case $h'(\beta^\star) = 0$. This implies $2(\beta^\star - u) + c\operatorname{sgn}(\beta^\star) = 0$. Therefore $2u = \operatorname{sgn}(\beta^\star)\,(2|\beta^\star| + c)$, which implies $\operatorname{sgn}(u) = \operatorname{sgn}(\beta^\star)$. Therefore $2(\beta^\star - u) + c\operatorname{sgn}(u) = 0$ from which we deduce $\beta^\star = u\left(1 - \frac{c}{2|u|}\right)$. Using again $\operatorname{sgn}(u) = sgn(\beta^\star)$, we deduce $1 - \frac{c}{2|u|} \geqslant 0$ and finally,
$$\beta^\star = u\left(1 - \frac{c}{2|u|}\right)^+$$
Case 2 $\beta^\star = 0$. In this case, for all $\beta \neq 0$, $h(\beta) \geqslant h(0) = u^2$, which is equivalent to $\beta^2 - 2\beta u + c|\beta| \geqslant 0$. Dividing by $|\beta|$ and letting $\beta \to 0$, we get $-2u\operatorname{sgn}(\beta) + c \geqslant 0$ which in turn implies $-2|u| + c \geqslant 0$. This shows $1 - \frac{c}{2|u|} \leqslant 0$ and we therefore have again :
$$\beta^\star = 0 = u\left(1 - \frac{c}{2|u|}\right)^+$$

□

Exercise 2 (Elastic-Net) Let $Y \in \mathbb{R}^n$ and $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_p] \in \mathbb{R}^{n\times p}$. The Elastic-Net estimator involves both a ℓ^2 and a ℓ^1 penalty. It is defined for $\lambda \geqslant 0$ and $\mu \geqslant 0$ by

$$\widehat{\beta}_{\lambda,\mu} \in \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \mathcal{L}(\beta) \quad \text{with} \quad \mathcal{L}(\beta) = \|Y - \mathbf{X}\beta\|^2 + \lambda\|\beta\|^2 + \mu|\beta|_{\ell^1}.$$

In the equation above, we have used the notation : $\|\beta\|^2 = \sum_{i=1}^p \beta_i^2$ and $|\beta|_{\ell^1} = \sum_{i=1}^p |\beta_i|$. In the following, we assume that the columns of $\mathbf{X}$ have norm 1, that is, $\boxed{\mathbf{X}_i'\mathbf{X}_i = 1}$ for any $i \in [1:p]$.

1. Let $j \in [1:p]$. Define $R_j = \mathbf{X}_j'\left(Y - \sum_{k:\,k\neq j} \beta_k \mathbf{X}_k\right)$. Writing $\mathbf{X}\beta = \sum_{i=1}^p \beta_i \mathbf{X}_i$, show that
$$\mathcal{L}(\beta) = \beta_j^2(1+\lambda) - 2\beta_j R_j + \mu|\beta_j| + H_j((\beta_k)_{k\in[1:p]\setminus\{j\}})$$
where $H_j((\beta_k)_{k\in[1:p]\setminus\{j\}})$ does not depend on β_j.
Solution.

We have
$$\begin{aligned}\mathcal{L}(\beta) &= \left\|Y - \sum_{i=1}^p \beta_i \mathbf{X}_i\right\|^2 + \lambda\sum_{i=1}^p \beta_i^2 + \mu\sum_{i=1}^n |\beta_i| \\ &= \beta_j^2 \underbrace{\mathbf{X}_j'\mathbf{X}_j}_{1} - 2\beta_j \underbrace{\mathbf{X}_j^\top\left(Y - \sum_{k:\,k\neq j} \beta_k \mathbf{X}_k\right)}_{R_j} + \lambda\beta_j^2 + \mu|\beta_j| + H_j((\beta_k)_{k\in[1:p]\setminus\{j\}})\end{aligned}$$
which completes the proof. □

2. Using Exercise 1, prove that the minimum of $\beta_j \to \mathcal{L}(\beta_1, \ldots, \beta_j, \ldots, \beta_p)$ is reached at some β_j^* and give the expression of β_j^* with respect to R_j.
Solution.

Write

$$\begin{aligned}\mathcal{L}(\beta) &= \beta_j^2(1+\lambda) - 2\beta_j R_j + \mu|\beta_j| + H_j((\beta_k)_{k\in[1:p]\setminus\{j\}})\\
&= (1+\lambda)\left(\beta_j^2 - 2\beta_j\frac{R_j}{1+\lambda} + \frac{\mu}{1+\lambda}|\beta_j|\right) + +H_j((\beta_k)_{k\in[1:p]\setminus\{j\}})\\
&= (1+\lambda)\left(\left(\beta_j - \frac{R_j}{1+\lambda}\right)^2 + \frac{\mu}{1+\lambda}|\beta_j|\right) + \tilde{H}_j((\beta_k)_{k\in[1:p]\setminus\{j\}})\end{aligned}$$

Applying the first exercise with $u = \frac{R_j}{1+\lambda}$ and $c = \frac{\mu}{1+\lambda}$, we get that $\mathcal{L}(\beta)$ is minimized at

$$\beta_j^* = \frac{R_j}{1+\lambda}\left(1 - \frac{\mu}{2|R_j|}\right)_+$$

□

3. What algorithm seems reasonable to you in order to approximate the Elastic-Net estimator ?
Solution.

You can for example choose an index j uniformly in $\{1,\ldots,p\}$ and then use the update formula at Question 2 to get the new value for β_j. □

Exercise 3 Let $Y \in \mathbb{R}^n$ and $\mathbf{X} = [\mathbf{X}_1,\ldots,\mathbf{X}_p] \in \mathbb{R}^{n\times p}$. Define $\mathbf{X}_{(0)} = [\mathbf{X}_1,\ldots,\mathbf{X}_{p-1}] \in \mathbb{R}^{n\times(p-1)}$. We assume that $\mathrm{rank}(\mathbf{X}) = p$, and $Y = \mathbf{X}\beta + \epsilon$ where $\epsilon \sim \mathcal{N}(0,\sigma^2 I_n)$.

Consider the following Hypothesis :

$$H_0 : \beta_p = 0 \quad \text{versus} \quad H_1 : \beta_p \neq 0$$

The aim to this exercise is to show that the Fisher Test (or F-test) associated with the statistic $F = \frac{\|\hat{Y}-\hat{Y}_{(0)}\|^2}{\|Y-\hat{Y}\|^2/(n-p)}$ where $\hat{Y}_{(0)} = P_{\mathbf{X}_{(0)}}Y$ is the orthogonal projection of Y on $\mathcal{I}(\mathbf{X}_{(0)})$ is equivalent to the Student test associated to the test statistic :

$$T = \frac{\hat{\beta}_p}{\hat{\sigma}\times\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{pp}}} \quad \text{where} \quad \hat{\sigma}^2 = \|Y-\hat{Y}\|^2/(n-p)$$

Define .

1. According to the course, what is the distribution of T under H_0 ? (No proof is needed).
Solution.

According to the course, $T \sim t(n-p)$. □

2. According to the course, what is the distribution of F under H_0 ? (No proof is needed).
Solution.

According to the course, $F \sim \mathcal{F}(1, n-p)$. □

3. Show that $P_{\mathbf{X}_{(0)}}(Y-\hat{Y}) = 0$.
Solution.

$Y-\hat{Y}$ is orthogonal to $\mathcal{I}(\mathbf{X})$. But $\mathcal{I}(\mathbf{X}_{(0)}) \subset \mathcal{I}(\mathbf{X})$. Hence, $Y-\hat{Y} \in \mathcal{I}(\mathbf{X}_{(0)})^\perp$, which implies that $P_{\mathbf{X}_{(0)}}(Y-\hat{Y}) = 0$. □

4. Deduce that $\hat{Y}_{(0)} = P_{\mathbf{X}_{(0)}}(\mathbf{X}\hat{\beta})$.
Solution.

Hence,

$$\hat{Y}_{(0)} = P_{\mathbf{X}_{(0)}}(Y) = P_{\mathbf{X}_{(0)}}(\hat{Y}) = P_{\mathbf{X}_{(0)}}(\mathbf{X}\hat{\beta})$$

□

5. Using that $\mathbf{X}\hat{\beta} = \sum_{i=1}^p \hat{\beta}_i\mathbf{X}_i$, deduce that $\hat{Y} - \hat{Y}_{(0)} = \hat{\beta}_p(\mathbf{X}_p - P_{\mathbf{X}_{(0)}}(\mathbf{X}_p))$.
Solution.

Hence,

$$\hat{Y} - \hat{Y}_{(0)} = \mathbf{X}\hat{\beta} - P_{\mathbf{X}_{(0)}}(\mathbf{X}\hat{\beta}) = \sum_{i=1}^p \hat{\beta}_i\mathbf{X}_i - \sum_{i=1}^p \hat{\beta}_i P_{\mathbf{X}_{(0)}}(\mathbf{X}_i) = \sum_{i=1}^p \hat{\beta}_i\left[\mathbf{X}_i - P_{\mathbf{X}_{(0)}}(\mathbf{X}_i)\right]$$

but $P_{\mathbf{X}_{(0)}}(\mathbf{X}_i) = \mathbf{X}_i$ for any $i \in [1:p-1]$. After cancelling the terms for $i \in [1:p-1]$, it remains $\hat{Y} - \hat{Y}_{(0)} = \hat{\beta}_p\mathbf{X}_p - \hat{\beta}_p P_{\mathbf{X}_{(0)}}(\mathbf{X}_p)$. □

6. Deduce that $F = \frac{\hat\beta_p^2}{\hat\sigma^2}\alpha$ where α is a real number that you will express.
Solution.

Using the previous question,

$$F = \frac{\|\hat Y - \hat Y_{(0)}\|^2}{\hat\sigma^2} = \frac{\hat\beta_p^2 \mathbf{X}_p'(Id - P_{\mathbf{X}_{(0)}})^2(\mathbf{X}_p)}{\hat\sigma^2} = \frac{\hat\beta_p^2}{\hat\sigma^2}\mathbf{X}_p'(Id - P_{\mathbf{X}_{(0)}})\mathbf{X}_p$$

since $P_{\mathbf{X}_{(0)}} = P'_{\mathbf{X}_{(0)}} = P^2_{\mathbf{X}_{(0)}}$. Hence, $\alpha = \mathbf{X}_p'(Id - P_{\mathbf{X}_{(0)}})\mathbf{X}_p$. □

7. Writing $\mathbf{X}'\mathbf{X} = \begin{pmatrix} \mathbf{X}_{(0)}'\mathbf{X}_{(0)} & \mathbf{X}_{(0)}'\mathbf{X}_p \\ \mathbf{X}_p'\mathbf{X}_{(0)} & \mathbf{X}_p'\mathbf{X}_p \end{pmatrix}$, we admit (without proof) that

$$(\mathbf{X}'\mathbf{X})^{-1}_{pp} = (\mathbf{X}_p'\mathbf{X}_p - \mathbf{X}_p'\mathbf{X}_{(0)}(\mathbf{X}_{(0)}'\mathbf{X}_{(0)})^{-1}\mathbf{X}_{(0)}'\mathbf{X}_p)^{-1}$$

Deduce that $F = h(T)$ where $h : \mathbb{R} \to \mathbb{R}$ is a function that you will express explicitely.
Solution.

Note that

$$(\mathbf{X}'\mathbf{X})^{-1}_{pp} = (\mathbf{X}_p'\mathbf{X}_p - \mathbf{X}_p'\mathbf{X}_{(0)}(\mathbf{X}_{(0)}'\mathbf{X}_{(0)})^{-1}\mathbf{X}_{(0)}'\mathbf{X}_p)^{-1} = (\mathbf{X}_p'(Id - P_{\mathbf{X}_{(0)}})\mathbf{X}_p)^{-1} = \alpha^{-1}$$

Combining with the previous question, $F = T^2 = h(T)$ □

8. Prove the following assertion : if T follows a Student distribution with $n-p$ degrees of freedom, show that T^2 follows a Fisher distribution $\mathcal{F}_{i,j}$ and express i and j in terms of $n-p$.
Solution.

If $U \sim N(0,1)$ and $V \sim \chi_2(n-p)$ such that U and V are independent then, T has the same law as

$$S = U/\sqrt{V/(n-p)} \sim t(n-p)$$

Then, $S^2 = \frac{U^2}{V/(n-p)}$ has the same law as T^2 and $U^2 \sim \chi_2(1)$, $V \sim \chi_2(n-p)$ anc U^2 and V are independent. Hence, S^2 and hence T^2 follows the distribution $\mathcal{F}_{1,n-p}$. □

Exercise 4 In this exercise, we consider iid observations $(X_i, Y_i)_{1\leqslant i\leqslant n}$ where Y_i takes values in $\{0,1\}$ and X_i takes values in $\mathbb{R}^d$. For a given $x \in \mathbb{R}^d$, we will consider the following classifier : $\hat h_n(x) = Y_{\phi_n(x)}$ where

$$\phi_n(x) = \mathrm{argmin}_{i\in[1:n]}\|x - X_i\|$$

In words, $\phi_n(x)$ is the index of the nearest neighbor of x among the data set $\{X_i;\ i \in [1:n]\}$. In all the exercise, we consider a couple of random variable (X,Y) with the same law as (X_i,Y_i) for any $i \geqslant 1$. And to avoid any ambiguity in the definition of the index $\phi_n(X)$ we assume that with probability 1, all the $\|X - X_i\|$ for any $i \in \mathbb{N}$ are strictly different.

Recall that the Bayes optimal classifier is defined by

$$h^\star(X) = \begin{cases} 1 & \text{if } \mathbb{P}(Y=1|X) > \mathbb{P}(Y=0|X) \\ 0 & \text{otherwise} \end{cases}$$

and recall that defining $\eta(X) = \mathbb{P}(Y=1|X)$ and $r^\star(X) = \min(\eta(X), 1-\eta(X))$, we have

$$R^\star = \mathbb{P}(Y \neq h^\star(X)) = \mathbb{E}[r^\star(X)] \leqslant \mathbb{P}(Y \neq h(X))$$

for any other classifier $h : \mathbb{R}^d \to \{0,1\}$. The aim of this exercise is to show the bound

$$R^\star \leqslant \lim_{n\to\infty} \mathbb{P}(Y \neq \hat h_n(X)) \leqslant 2R^\star(1-R^\star)$$

Define $X_{(n)} = X_{\phi_n(X)}$ the nearest neighbor of X among the set $\{(X_i);\ i \in [1:n]\}$. We admit that, almost-surely,

$$\lim_{n\to\infty} X_{(n)} = X$$

and we assume that η is continuous, so that $\lim_{n\to\infty}\eta(X_{(n)}) = \eta(X)$.

1. Show that $r^\star(X)(1-r^\star(X)) = \eta(X)(1-\eta(X))$.
Solution.

If $\eta(X) \leqslant 1/2$, then $r^\star(X) = \eta(X)$ and $r^\star(X)(1 - r^\star(X)) = \eta(X)(1 - \eta(X))$.
If $\eta(X) > 1/2$, then $r^\star(X) = 1 - \eta(X)$ and $r^\star(X)(1 - r^\star(X)) = \eta(X)(1 - \eta(X))$. □

2. Show that for any $i \in [1 : n]$, we have almost surely,

$$\mathbb{P}(Y = 0, Y_i = 1, \phi_n(X) = i | X, X_{1:n}) = (1 - \eta(X))\eta(X_i)\mathbb{1}_{\{\phi_n(X)=i\}}$$

Solution.

We have

$$\begin{aligned}\mathbb{P}(Y = 0, Y_i = 1, \phi_n(X) = i | X, X_{1:n}) &= \mathbb{E}\left(\mathbb{1}_{Y=0}\mathbb{1}_{Y_i=1}\mathbb{1}_{\phi_n(X)=i} | X, X_{1:n}\right) \\ &= \mathbb{E}(\mathbb{1}_{Y=0}|X)\mathbb{E}(\mathbb{1}_{Y_i=1}|X_i)\mathbb{1}_{\{\phi_n(X)=i\}} \\ &= (1 - \eta(X))\eta(X_i)\mathbb{1}_{\{\phi_n(X)=i\}}\end{aligned}$$

□

3. Noting that almost-surely, $\eta(X_{(n)}) = \sum_{i=1}^n \eta(X_i)\mathbb{1}_{\{\phi_n(X)=i\}}$, deduce that, almost-surely,

$$\mathbb{P}(Y = 0, \hat{h}_n(X) = 1 | X, X_{1:n}) = (1 - \eta(X))\eta(X_{(n)}),$$

Solution.

$$\begin{aligned}\mathbb{P}(Y = 0, \hat{h}_n(X) = 1 | X, X_{1:n}) &= \sum_{i=1}^n \mathbb{P}(Y = 0, Y_i = 1, \phi_n(X) = i | X, X_{1:n}) \\ &= \sum_{i=1}^n (1 - \eta(X))\eta(X_i)\mathbb{1}_{\{\phi_n(X)=i\}} \\ &= (1 - \eta(X))\eta(X_{(n)})\end{aligned}$$

□

4. In the same way, show that

$$\mathbb{P}(Y = 1, \hat{h}_n(X) = 0 | X, X_{1:n}) = \eta(X)(1 - \eta(X_{(n)})),$$

Solution.

This is a simple adapatation from the two previous questions. □

5. Using that if $(Z_n)_{n\in\mathbb{N}}$ is a family of bounded random variables, converging almost surely to Z, then $\lim_{n\to\infty} \mathbb{E}[Z_n] = \mathbb{E}[Z]$, show that

$$\lim_{n\to\infty} \mathbb{P}(Y \neq \hat{h}_n(X)) = 2\mathbb{E}[\eta(X)(1 - \eta(X))]$$

Solution.

We have

$$\begin{aligned}\mathbb{P}(Y \neq \hat{h}_n(X)) &= \mathbb{E}[\mathbb{P}(Y = 0, \hat{h}_n(X) = 1 | X, X_{1:n})] + \mathbb{E}[\mathbb{P}(Y = 1, \hat{h}_n(X) = 0 | X, X_{1:n})] \\ &= \mathbb{E}[\eta(X_{(n)})(1 - \eta(X))] + \mathbb{E}[\eta(X)(1 - \eta(X_{(n)}))] \to 2\mathbb{E}[\eta(X)(1 - \eta(X))]\end{aligned}$$

where we have used that $0 \leqslant \eta(X_{(n)})(1 - \eta(X)) \leqslant 1$ and $\lim_{n\to\infty} \eta(X_{(n)})(1 - \eta(X)) = \eta(X)(1 - \eta(X))$ almost surely. And the same reasoning holds for the second expectation $\mathbb{E}[\eta(X)(1 - \eta(X_{(n)}))]$. □

6. Deduce that

$$\lim_{n\to\infty} \mathbb{P}(Y \neq \hat{h}_n(X)) = 2\mathbb{E}[r^\star(X)(1 - r^\star(X))]$$

Solution.

The results holds immediately from the previous question and the first question. □

7. Deduce that

$$\lim_{n\to\infty} \mathbb{P}(Y \neq \hat{h}_n(X)) \leqslant 2R^\star(1 - R^\star)$$

Solution.

From the previous question,

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}(Y \neq \hat{h}_n(X)) &= 2\mathbb{E}[r^\star(X)(1-r^\star(X))] \\ &= 2\left[\mathbb{E}[r^\star(X)] - \mathbb{E}[r^\star(X)^2]\right] \\ &\leqslant 2\left[\mathbb{E}[r^\star(X)] - \mathbb{E}^2[r^\star(X)]\right] = 2R^\star(1-R^\star)\end{aligned}$$

□

8. Conclude.
Solution.

Since $\hat{h}_n$ is a classifier, we also have

$$R^\star \leqslant \mathbb{P}(Y \neq \hat{h}_n(X))$$

Combining with the previous question, we finally get

$$R^\star \leqslant \lim_{n\to\infty} \mathbb{P}(Y \neq \hat{h}_n(X)) \leqslant 2R^\star(1-R^\star)$$

□