Regression and Classification, EMINES. *Examination, 2024. 2 hours*

Recall the notation $[a:b] = \{a, a+1, \ldots, b\}$ for any integers $a < b$.

Exercise 1 (Weighted optimal Bayes classifier.) Consider two random variables (X, Y) taking values on $\mathbb{R}^d \times \{0,1\}$. Let $h : \mathbb{R}^d \to \{0,1\}$ be a classifier. We are first interested in minimizing (with respect to h) the weighted misclassification probability

$$M_2(h) = \alpha \mathbb{P}(Y = 0, h(X) = 1) + (1-\alpha)\mathbb{P}(Y = 1, h(X) = 0)$$

where $\alpha \in [0,1]$ is given. For example, the random variable Y may represent the illness of a patient, and in this situation, a misclassification error when the patient is ill ($Y = 1$) may be much more severe than when the patient is in good health ($Y = 0$). Hence, the coefficient α weights the importance we place on these two types of errors. Finally, we aim to find the optimal classifier in this context, that is, we aim to solve the minimization problem

$$h_2^* = \mathrm{argmin}_{h \in \mathcal{H}} M_2(h)$$

where $\mathcal{H}$ is the set of all measurable functions $h : \mathbb{R}^d \to \{0,1\}$.

1. Show that
$$M_2(h) = \mathbb{E}\left[\alpha \mathbb{P}(Y = 0|X)\mathbf{1}_{\{h(X)=1\}} + (1-\alpha)\mathbb{P}(Y = 1|X)\mathbf{1}_{\{h(X)=0\}}\right].$$

Solution.

We have

$$\begin{aligned} M_2(h) &= \alpha \mathbb{E}\left[\mathbf{1}_{\{Y=0\}}\mathbf{1}_{\{h(X)=1\}}\right] + (1-\alpha)\mathbb{E}\left[\mathbf{1}_{\{Y=1\}}\mathbf{1}_{\{h(X)=0\}}\right]. \\ &= \alpha \mathbb{E}\left[\mathbb{E}[\mathbf{1}_{\{Y=0\}}|X]\mathbf{1}_{\{h(X)=1\}}\right] + (1-\alpha)\mathbb{E}\left[\mathbb{E}[\mathbf{1}_{\{Y=1\}}|X]\mathbf{1}_{\{h(X)=0\}}\right] \end{aligned}$$

where we have used that tower property (the expectation is the expectation of the conditional expectation wrt X). Rearranging the terms and noting that $\mathbb{E}[\mathbf{1}_{\{Y=i\}}|X] = \mathbb{P}(Y = i|X)$ yields the desired formula. □

2. Deduce that $M_2(h) \geqslant \mathbb{E}\left[\min\left(\alpha \mathbb{P}(Y = 0|X), (1-\alpha)\mathbb{P}(Y = 1|X)\right)\right]$.

Solution.

Denote $a(X) = \min\left(\alpha \mathbb{P}(Y = 0|X), (1-\alpha)\mathbb{P}(Y = 1|X)\right)$. We have

$$\alpha \mathbb{P}(Y = 0|X)\mathbf{1}_{\{h(X)=1\}} + (1-\alpha)\mathbb{P}(Y = 1|X)\mathbf{1}_{\{h(X)=0\}} \geqslant a(X)\mathbf{1}_{\{h(X)=1\}} + a(X)\mathbf{1}_{\{h(X)=0\}} = a(X)$$

Taking the expectation and combining with the previous question proves the result.

□

3. Deduce that in this context, the optimal classifier h_2^* writes :

$$h_2^*(X) = \begin{cases} 1 & \text{if } \mathbb{P}(Y = 1|X) \geqslant \delta \\ 0 & \text{otherwise} \end{cases}$$

where δ should be expressed with respect to α.

Solution.

We wish to obtain equality in the inequality of the previous question for a particular classifier h_2^*. To do so, we wish to have :

$$\alpha \mathbb{P}(Y = 0|X)\mathbf{1}_{\{h_2^*(X)=1\}} + (1-\alpha)\mathbb{P}(Y = 1|X)\mathbf{1}_{\{h_2^*(X)=0\}} = a(X)$$

Hence we will choose h_2^* such that $h_2^*(X) = 1$ corresponds to the case where $\alpha \mathbb{P}(Y = 0|X) = a(X)$, which holds if and only if $\alpha \mathbb{P}(Y = 0|X) \leqslant (1-\alpha)\mathbb{P}(Y = 1|X)$. Using $\mathbb{P}(Y = 0|X) = 1 - \mathbb{P}(Y = 1|X)$, we finally get

$$\alpha(1 - \mathbb{P}(Y = 1|X)) \leqslant (1-\alpha)\mathbb{P}(Y = 1|X) \iff \alpha \leqslant \mathbb{P}(Y = 1|X)$$

Hence, $\delta = \alpha$.

□

We now turn to the more general case where Y may take d values instead of only two values. Returning to the example of a patient, the different values of Y may represent different states of illness for the patient. More precisely, we consider two random variables (X, Y) taking values on $\mathbb{R}^d \times [0:(d-1)]$. Let $h : \mathbb{R}^d \to [0:(d-1)]$ be a classifier. We are interested in minimizing the weighted misclassification probability

$$M(h) = \sum_{j=0}^{d-1} \alpha_j \mathbb{P}(Y = j, h(X) \neq j)$$

where $(\alpha_j)_{j\in[0:d-1]}$ are non-negative coefficients satisfying $\sum_{j=0}^{d-1}\alpha_j = 1$. The minimization problem hence writes

$$h^* = \mathrm{argmin}_{h\in\mathcal{H}} M(h)$$

where $\mathcal{H}$ is the set of all measurable functions $h : \mathbb{R}^d \to [0:d-1]$.

4. Show that

$$M(h) = \sum_{i=0}^{d-1}\sum_{j=0}^{d-1}\mathbb{P}(h(X)=i, Y=j)\beta_{i,j}$$

where $(\beta_{i,j})_{0\leqslant i,j\leqslant d-1}$ should be expressed in terms of $(\alpha_j)_{0\leqslant j\leqslant d-1}$.
Solution.

We have

$$M(h) = \sum_{j=0}^{d-1}\alpha_j\mathbb{P}(Y=j, h(X)\neq j) = \sum_{j=0}^{d-1}\sum_{i=0}^{d-1}\mathbf{1}_{\{i\neq j\}}\alpha_j\mathbb{P}(Y=j, h(X)=i)$$

Hence $\beta_{i,j} = \mathbf{1}_{\{i\neq j\}}\alpha_j$. □

5. Deduce that for any classifier h, $M(h) \geqslant \mathbb{E}\left[\min_{i\in[0:d-1]}\left(\sum_{j\neq i}\alpha_j\mathbb{P}(Y=j|X)\right)\right]$.
Solution.

We have, using again the tower property,

$$\begin{aligned}
M(h) &= \sum_{i=0}^{d-1}\sum_{j=0}^{d-1}\mathbb{P}(h(X)=i, Y=j)\beta_{i,j} = \sum_{i=0}^{d-1}\sum_{j=0}^{d-1}\mathbb{E}\left[\mathbf{1}_{\{h(X)=i\}}\mathbb{E}\left[\mathbf{1}_{\{Y=j\}}|X\right]\right]\beta_{i,j}\\
&= \mathbb{E}\left[\sum_{i=0}^{d-1}\mathbf{1}_{\{h(X)=i\}}\sum_{j=0}^{d-1}\underbrace{\beta_{i,j}}_{\alpha_j\mathbf{1}_{i\neq j}}\mathbb{P}(Y=j|X)\right]\\
&\geqslant \mathbb{E}\left[\underbrace{\sum_{i=0}^{d-1}\mathbf{1}_{\{h(X)=i\}}}_{=1}\left\{\min_{i\in[0:d-1]}\left(\sum_{j\neq i}\alpha_j\mathbb{P}(Y=j|X)\right)\right\}\right]
\end{aligned}$$

which concludes the proof. □

6. Deduce the expression of the optimal classifier h^*.
Solution.

From the previous question, we deduce that

$$h^*(X) = \mathrm{argmin}_{i\in[0:d-1]}\sum_{j\neq i}\alpha_j\mathbb{P}(Y=j|X)$$

□

Exercise 2 Let $Y_n = \begin{pmatrix} y_1 \\ \vdots \\ y_n\end{pmatrix} \in \mathbb{R}^n$ and $\mathbf{X}_n = \begin{bmatrix}\mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n'\end{bmatrix} \in M_{n,p}(\mathbb{R})$. Define $\hat\beta_n = \mathrm{argmin}_{\beta\in\mathbb{R}^p}\|\mathsf{Y}_n - \beta\mathbf{X}_n\|^2$.

Define $\hat y_n = \mathbf{x}_n'\hat\beta_n$ and $\hat y_n^- = \mathbf{x}_n'\hat\beta_{n-1}$. Define $H = \mathbf{X}_n(\mathbf{X}_n^T\mathbf{X}_n)^{-1}\mathbf{X}_n^T$, the orthogonal projection matrix on $\mathcal{I}(\mathbf{X}_n)$ (also called the hat matrix). In this exercise, we want to show that

$$y_n - \hat y_n^- = \frac{y_n - \hat y_n}{1-h_{n,n}}$$

where $H = (h_{k,\ell})_{1\leqslant k,\ell\leqslant n}$.

1. Recall without proof the explicit expression of $\hat\beta_n$ in terms of $\mathbf{X}_n$ and Y_n.
Solution.

From the course, we get that $\hat\beta_n = (\mathbf{X}_n^T\mathbf{X}_n)^{-1}\mathbf{X}_n^T\mathsf{Y}_n$ □

2. Define $\tilde{\mathsf{Y}}_n = (\tilde{y}_i)_{1 \leqslant i \leqslant n} \in \mathbb{R}^n$ such that $\tilde{y}_i = y_i$ for $i \in \{1, \dots, n-1\}$ and $\tilde{y}_n = \hat{y}_n^-$. Show that for any $\beta \in \mathbb{R}^p$,

$$\sum_{j=1}^{n} (\tilde{y}_j - \mathbf{x}_j'\beta)^2 \geqslant \sum_{j=1}^{n-1} (y_j - \mathbf{x}_j'\beta)^2 \geqslant \sum_{j=1}^{n-1} (y_j - \mathbf{x}_j'\hat{\beta}_{n-1})^2$$

Solution.

We have, using first that $(\tilde{y}_n - \mathbf{x}_j^T\beta)^2 \geqslant 0$ and then that $\tilde{y}_i = y_i$ for $i \in \{1, \dots, n-1\}$,

$$\sum_{j=1}^{n} (\tilde{y}_j - \mathbf{x}_j'\beta)^2 \geqslant \sum_{j=1}^{n-1} (\tilde{y}_j - \mathbf{x}_j'\beta)^2 = \sum_{j=1}^{n-1} (y_j - \mathbf{x}_j'\beta)^2$$

which proves the first inequality. The last equality follows by definition of $\hat{\beta}_{n-1} = \mathrm{argmin}_{\beta \in \mathbb{R}^p} \|\mathsf{Y}_{n-1} - \beta \mathbf{X}_{n-1}\|^2$. $\square$

3. Deduce that $\hat{\beta}_{n-1} = \mathrm{argmin}_{\beta \in \mathbb{R}^p} \|\tilde{\mathsf{Y}}_n - \mathbf{X}_n\beta\|^2$.
Solution.

If we plug $\beta = \hat{\beta}_{n-1}$ in the left hand side of the inequality (in the previous question), we get using that $\tilde{y}_i = y_i$ for $i \in \{1, \dots, n-1\}$ and $\tilde{y}_n = \hat{y}_n^-$,

$$\sum_{j=1}^{n} (\tilde{y}_j - \mathbf{x}_j'\hat{\beta}_{n-1})^2 = \sum_{j=1}^{n-1} (y_j - \mathbf{x}_j'\hat{\beta}_{n-1})^2 + (\underbrace{\hat{y}_n^- - \mathbf{x}_n'\hat{\beta}_{n-1}}_{=0})^2 = \sum_{j=1}^{n-1} (y_j - \mathbf{x}_j'\hat{\beta}_{n-1})^2$$

Hence equality holds in the inequality of the previous question for $\beta = \hat{\beta}_{n-1}$, which shows that the left hand side is minimised at $\hat{\beta}_{n-1}$. Hence, $\hat{\beta}_{n-1} = \mathrm{argmin}_{\beta \in \mathbb{R}^p} \|\tilde{\mathsf{Y}}_n - \mathbf{X}_n\beta\|^2$. $\square$

4. Deduce an expression of $\hat{\beta}_{n-1}$ in terms of $\tilde{\mathsf{Y}}_n$ and $\mathbf{X}_n$.
Solution.

From the previous question, we deduce that $\hat{\beta}_{n-1} = (\mathbf{X}_n^T\mathbf{X}_n)^{-1}\mathbf{X}_n^T\tilde{\mathsf{Y}}_n$. $\square$

5. Deduce that $\hat{y}_n^- = \sum_{j=1}^{n} h_{n,j}\tilde{y}_j$
Solution.

Since $\hat{\beta}_{n-1} = (\mathbf{X}_n^T\mathbf{X}_n)^{-1}\mathbf{X}_n^T\tilde{\mathsf{Y}}_n$, we get $\hat{y}_n^- = x_n'\hat{\beta}_{n-1} = \sum_{j=1}^{n} h_{n,j}\tilde{y}_j$ $\square$

6. Show that $\hat{y}_n^- = \hat{y}_n - h_{n,n}y_n + h_{n,n}\hat{y}_n^-$.
Solution.

$$\hat{y}_n^- = \sum_{j=1}^{n} h_{n,j}\tilde{y}_j = \sum_{j=1}^{n-1} h_{n,j}y_j + h_{n,n}\hat{y}_n^-$$

$$\hat{y}_n = \sum_{j=1}^{n} h_{n,j}\tilde{y}_j = \sum_{j=1}^{n-1} h_{n,j}y_j + h_{n,n}y_n$$

Substracting these two equations yields : $\hat{y}_n^- = \hat{y}_n - h_{n,n}y_n + h_{n,n}\hat{y}_n^-$ $\square$

7. Conclude.
Solution.

Finally, the previous question gives : $(1 - h_{n,n})\hat{y}_n^- = \hat{y}_n - h_{n,n}y_n$. Hence $\hat{y}_n^- = \frac{\hat{y}_n - h_{n,n}y_n}{1 - h_{n,n}}$. Then,

$$y_n - \hat{y}_n^- = y_n - \frac{\hat{y}_n - h_{n,n}y_n}{1 - h_{n,n}} = \frac{y_n - \hat{y}_n}{1 - h_{n,n}}$$

$\square$