Exercices 3 solutions

TABLE DES MATIÈRES

1 BAYES ESTIMATOR AND BAYES RISK

Consider the following joint random variable (X, Y).

— $\mathcal{X} = \{0, 1, 2\}$

— $\mathcal{Y} = \{0, 1\}$.

— X follows a uniform law on $\mathcal{X}$.

—
$$Y = \begin{cases} B(1/5) \text{ if } X = 0 \\ B(3/4) \text{ if } X = 1 \\ B(2/3) \text{ if } X = 2 \end{cases}$$

With $B(p)$ a Bernoulli law with parameter p.

Compute the Bayes estimator and the bayes risk.

1.1 Solution

1.1.1 *Bayes predictor, general case*

We prove again the general result on the Bayes predictor in the case of binary classification. We have seen that the Bayes predictor is defined by

$$f^*(x) = \arg\min_{z \in \mathcal{Y}} E\left[l(y, z)|X = x\right] \quad (1)$$

Hence

$$\begin{aligned} f^*(x) &= \arg\min_{z \in \mathcal{Y}} E\Big[l(y,z)|X=x\Big] \\ &= \arg\min_{z \in \mathcal{Y}} P(Y \neq z|X=x) \\ &= \arg\min_{z \in \mathcal{Y}} 1 - P(Y=z|X=x) \\ &= \arg\max_{z \in \mathcal{Y}} P(Y=z|X=x) \end{aligned} \tag{2}$$

The optimal classifier selects the most probable output given $X = x$.

1.1.2 *Application*

In this case :

— $f^*(0) = 0$

— $f^*(1) = 1$

— $f^*(2) = 1$

1.1.3 *Bayes risk, general case*

We have also seen that using the law of total expectation, with the "0-1" loss,

$$\begin{aligned} R^* &= E\Big[l(Y, f^*(X))\Big] \\ &= E_X\Big[E_Y\Big(l(Y \neq f^*(X)|X\Big)\Big] \\ &= E_X\Big[P(Y \neq f^*(X)|X)\Big] \end{aligned} \tag{3}$$

But we have

$$P\big(Y \neq f^*(X)|X = x\big) = P\big(Y \neq f^*(x)\big) \tag{4}$$

We note $\eta(x) = P(Y = 1|X = x)$. Then,

— If $\eta(x) > \frac{1}{2}$, then $f^*(x) = 1$, and $P\big(Y \neq f^*(x)\big) = P\big(Y = 0\big) = 1 - \eta(x)$

— If $\eta(x) < \frac{1}{2}$, then $f^*(x) = 0$, and $P\big(Y \neq f^*(x)\big) = P\big(Y = 1\big) = \eta(x)$

In both cases, $P\big(Y \neq f^*(x)\big) = \min(\eta(x), 1 - \eta(x))$.
We conclude that

$$R^* = E_X\Big[\min(\eta(X), 1 - \eta(X))\Big] \tag{5}$$

1.1.4 *Application*

In this setting :

$$\begin{aligned} R^* &= \frac{1}{3}\frac{1}{5} + \frac{1}{3}\frac{1}{4} + \frac{1}{3}\frac{1}{3} \\ &= \frac{1}{3}(\frac{1}{5} + \frac{1}{4} + \frac{1}{3}) \\ &= \frac{1}{3}(\frac{12}{60} + \frac{15}{60} + \frac{20}{60}) \\ &= \frac{1}{3}(\frac{47}{60}) \\ &= \frac{47}{180} \end{aligned} \tag{6}$$

2 LOGISTIC REGRESSION

Summary of the setting : in the context of binary classification, we consider the following setting.

— $\mathcal{X} = \mathbb{R}^d$

— $\mathcal{Y} = \{0, 1\}$ (sometimes $\mathcal{Y} = \{-1, 1\}$)

— $l_{0-1}(y, z) = 1_{y \neq z}$ ("0-1" loss)

Note that we can extend these definitions to non-binary classification. We would like a predictor that minimizes the binary loss.

Definition 1. Binary loss function

$$\hat{R}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{y_i \neq \hat{f}(x_i)}$$

However, as we have seen in the class, it is hard to minimize the binary loss as it is neither differentiable nor convex in θ. We can replace it by a **convex, differentiable surrogate loss** (**substitut convexe**). Several possibilities exist instead of using $\mathbb{1}_{y_i \neq \hat{f}(x_i)}$ as l (binary loss). The **logistic loss** is one of them

Definition 2. Logistic loss

$$l(\hat{y}, y) = y \log(1 + e^{-\hat{y}}) + (1 - y) \log(1 + e^{\hat{y}}) \tag{7}$$

We can define the corresponding empirical risk.

$$R_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} l(x_i^T \theta, y_i) \tag{8}$$

Definition 3. Logistic regression estimator
If l is the logistic loss, it is defined as

$$\hat{\theta}_{logit} = \arg\min_{\theta \in \mathbb{R}^d} R_n(\theta)$$

Compute the gradient of $R_n(\theta)$, $\nabla_\theta R_n$.

2.1 Solution

We introduce the following functions :

$$g_i = \begin{cases} \mathbb{R}^d \to \mathbb{R} \\ \theta \mapsto l(x_i^T \theta, y_i) \end{cases}$$

$$u = \begin{cases} \mathbb{R} \to \mathbb{R} \\ \hat{y} \mapsto l(\hat{y}, y_i) \end{cases}$$

$$v_i = \begin{cases} \mathbb{R}^d \to \mathbb{R} \\ \theta \mapsto x_i^T \theta \end{cases}$$

Then, $\forall i$

$$l(x_i^T \theta, y_i) = g_i(\theta) = (u \circ v_i)(\theta) \tag{9}$$

Hence, by composition of the jacobian matrices,

$$L_\theta^{g_i} = L_{v_i(\theta)}^{u} L_\theta^{v_i} = u'(v_i(\theta)) L_\theta^{v_i} \tag{10}$$

We have :

— $L_\theta^{v_i} = x_i^T$
— We have seen that $\forall y, \hat{y}$,

$$\frac{\partial l}{\partial \hat{y}}(\hat{y}, y) = \sigma(\hat{y}) - y \tag{11}$$

Hence, $u'(v_i(\theta)) = \sigma(v_i(\theta)) - y_i$

Finally,

$$L_\theta^{g_i} = (\sigma(x_i^T\theta) - y_i)x_i^T \tag{12}$$

And

$$\nabla_\theta g_i = (\sigma(x_i^T\theta) - y_i)x_i \tag{13}$$

We can now compute $\nabla_\theta R_n$.

$$\begin{aligned} \nabla_\theta R_n &= \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta g_i \\ &= \frac{1}{n}\sum_{i=1}^{n} (\sigma(x_i^T\theta) - y_i)x_i \end{aligned} \tag{14}$$

3 OLS RISK DECOMPOSITION

Show the first part of proposition 15 in FTML.pdf (Risk decomposition for OLS, linear model, fixed design).

$$R_X(\theta) - R_X(\theta^*) = ||\theta - \theta^*||_{\hat{\Sigma}}^2$$

where $R_X(\theta)$ is the fixed design risk, defined by

$$R_X(\theta) = E_Y\Big[\frac{1}{n}||Y - X\theta||^2\Big] \tag{15}$$

with Σ the non-centered empirical covariance matrix :

$$\hat{\Sigma} = \frac{1}{n}X^TX \in \mathbb{R}^{d,d} \tag{16}$$

and the Mahalanobis distance norm.

$$||\theta||_{\Sigma}^2 = \theta^T\hat{\Sigma}\theta \tag{17}$$

3.0.1 *Solution*

We note that

$$\begin{aligned} R_X(\theta^*) &= E_Y\Big[\frac{1}{n}||Y - X\theta^*||^2\Big] \\ &= \frac{1}{n}E_\epsilon\Big[||\epsilon||^2\Big] \\ &= \frac{1}{n}E_\epsilon\Big[\sum_{i=1}^{n}\epsilon_i^2\Big] \\ &= \sigma^2 \end{aligned} \tag{18}$$

We now decompose $R_X(\theta)$:

$$\begin{aligned} R_X(\theta) &= E_Y\Big[\frac{1}{n}||Y - X\theta||^2\Big] \\ &= \frac{1}{n}E_Y\Big[||Y - X\theta^* + X\theta^* - X\theta||^2\Big] \end{aligned} \tag{19}$$

For for any vectors z and $z' \in \mathbb{R}^n$, we have

$$\begin{aligned} \|z+z'\|^2 &= \langle z+z', z+z' \rangle \\ &= \langle z,z\rangle + 2\langle z,z'\rangle + \langle z',z'\rangle \\ &= \|z\|^2 + 2\langle z,z'\rangle + \|z'\|^2 \end{aligned} \tag{20}$$

Hence,

$$\begin{aligned} R_X(\theta) &= E_Y\Big[\frac{1}{n}\|Y - X\theta^*\|^2 + \frac{2}{n}\langle Y - X\theta^*, X\theta^* - X\theta\rangle + \frac{1}{n}\|X\theta^* - X\theta\|^2\Big] \\ &= R_X(\theta^*) + \frac{2}{n}E_Y\Big[\langle Y - X\theta^*, X\theta^* - X\theta\rangle\Big] + E_Y\Big[\frac{1}{n}\|X(\theta^* - \theta)\|^2\Big] \\ &= R_X(\theta^*) + \frac{2}{n}E_Y\Big[\langle \epsilon, X(\theta^* - \theta)\rangle\Big] + \frac{1}{n}\|X(\theta^* - \theta)\|^2 \end{aligned} \tag{21}$$

But using that for all $z \in \mathbb{R}^n$, $\|z\|^2 = \langle z,z\rangle = z^T z$, the last term writes :

$$\begin{aligned} \frac{1}{n}\|X(\theta^* - \theta)\|^2 &= \frac{1}{n}\big(X(\theta^* - \theta)\big)^T\big(X(\theta^* - \theta)\big) \\ &= \big(\theta^* - \theta\big)^T \frac{1}{n}X^T X\big(\theta^* - \theta\big) \\ &= \|\theta - \theta^*\|^2_{\hat{\Sigma}} \end{aligned} \tag{22}$$

We now focus on the second term of the sum :

$$\begin{aligned} E_Y\Big[\langle \epsilon, X(\theta^* - \theta)\rangle\Big] &= E_Y\Big[\epsilon^T, X(\theta^* - \theta)\Big] \\ &= (E_Y\big[\epsilon\big])^T X(\theta^* - \theta) \\ &= 0 \end{aligned} \tag{23}$$

This concludes the proof.