

Exercices 3 solutions

TABLE DES MATIÈRES

1	Bayes estimator and Bayes risk	1
1.1	Solution	1
1.1.1	Bayes predictor, general case	1
1.1.2	Application	2
1.1.3	Bayes risk, general case	2
1.1.4	Application	2
2	Logistic regression	3
2.1	Solution	3
3	OLS risk decomposition	4
3.0.1	Solution	4

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

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

Hence

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

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,

$$\begin{aligned}
 R^* &= \mathbb{E} \left[l(Y, f^*(X)) \right] \\
 &= \mathbb{E} \left[\mathbb{P}(Y \neq f^*(X)) \right] \\
 &= \mathbb{E}_X \left[\mathbb{E}_Y \left(\mathbb{P}(Y \neq f^*(X) | X) \right) \right] \\
 &= \mathbb{E}_X \left[\mathbb{P}(Y \neq f^*(X) | X) \right]
 \end{aligned} \tag{3}$$

But we have

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

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

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

In both cases, $\mathbb{P}(Y \neq f^*(x)) = \min(\eta(x), 1 - \eta(x))$.

We conclude that

$$R^* = \mathbb{E}_X \left[\min(\eta(X), 1 - \eta(X)) \right] \tag{5}$$

1.1.4 Application

In this setting :

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

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

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

We can define the corresponding empirical risk.

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

Definition 3. Logistic regression estimator

If l is the logistic loss, it is defined as

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

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

2.1 Solution

We introduce the following functions :

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

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

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

Then, $\forall i$

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

Hence, by composition of the jacobian matrices,

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

We have :

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

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

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

Finally,

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

And

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

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

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

3 OLS RISK DECOMPOSITION

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

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

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

$$R_X(\theta) = E_Y \left[\|Y - X^T \theta\|^2 \right] \quad (15)$$

3.0.1 Solution

We note that

$$\begin{aligned} R_X(\theta^*) &= E_Y \left[\|Y - X\theta^*\|^2 \right] \\ &= E_{\epsilon} \left[\|\epsilon\|^2 \right] \\ &= \sigma^2 \end{aligned} \quad (16)$$

We now decompose $R_X(\theta)$:

$$\begin{aligned} R_X(\theta) &= E_Y \left[\|Y - X\theta\|^2 \right] \\ &= E_Y \left[\|Y - X\theta^* + X\theta^* - X\theta\|^2 \right] \end{aligned} \quad (17)$$

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

$$\begin{aligned} \|z + z'\|^2 &= \langle z + z', z + z' \rangle \\ &= \langle z, z \rangle + 2\langle z, z' \rangle + \langle z', z' \rangle \\ &= \|z\|^2 + 2\langle z, z' \rangle + \|z'\|^2 \end{aligned} \quad (18)$$

Hence,

$$\begin{aligned}
 R_X(\theta) &= E_Y \left[\|Y - X\theta^*\|^2 + 2\langle Y - X\theta^*, X\theta^* - X\theta \rangle + \|X\theta^* - X\theta\|^2 \right] \\
 &= R_X(\theta^*) + 2E_Y \left[\langle Y - X\theta^*, X\theta^* - X\theta \rangle \right] + E_Y \left[\|X(\theta^* - \theta)\|^2 \right] \\
 &= R_X(\theta^*) + 2E_Y \left[\langle \epsilon, X(\theta^* - \theta) \rangle \right] + \|X(\theta^* - \theta)\|^2
 \end{aligned} \tag{19}$$

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

$$\begin{aligned}
 \|X(\theta^* - \theta)\|^2 &= (X(\theta^* - \theta))^T (X(\theta^* - \theta)) \\
 &= (\theta^* - \theta)^T X^T X (\theta^* - \theta) \\
 &= \|\theta - \theta^*\|_{\hat{\Sigma}}^2
 \end{aligned} \tag{20}$$

We now focus on the second term of the sum :

$$\begin{aligned}
 E_Y \left[\langle \epsilon, X(\theta^* - \theta) \rangle \right] &= E_Y \left[\epsilon^T, X(\theta^* - \theta) \right] \\
 &= (E_Y[\epsilon])^T X(\theta^* - \theta) \\
 &= 0
 \end{aligned} \tag{21}$$

This concludes the proof.