Statistical learning, 2023-2024
Gérard Biau

Problem 1

Let (X, Y) be a random pair taking values in $\mathbb{R} \times \{0, 1\}$, where X is uniformly distributed on $[-2, 2]$. We assume that

$$Y = \begin{cases} \mathbb{1}_{[U \leq 2]} & \text{if } X \leq 0 \\ \mathbb{1}_{[U > 1]} & \text{if } X > 0, \end{cases}$$

where U is a random variable uniformly distributed on $[0, 10]$, independent of X. Compute the Bayes rule and the Bayes error associated with (X, Y).

Problem 2

Let (X, Y) be a random pair taking values in $\mathbb{R}_+ \times \{0, 1\}$. We let $\eta(x) = \mathbb{P}(Y = 1 | X = x)$ and assume that $\eta(x) = x/(c + x)$, where c is a positive constant.

1. Show that the Bayes risk $L^\star$ associated with (X, Y) is

$$L^\star = \mathbb{E}\Big(\frac{\min(c, X)}{c + X}\Big).$$

2. Provide an expression of $L^\star$ when X is uniformly distributed on $[0, \alpha c]$, where $\alpha \geq 1$.

3. Prove that there is a value of α maximizing $L^\star$.

Problem 3

Let (X, Y) be a random pair taking values in $\mathbb{R}^3 \times \{0, 1\}$. The three components of X are denoted by T, B, and E, respectively. The variable T represents the average number of hours per week that a student spends watching TV, and the variable B the average number of hours per week he/she spends in bars. The component E is an abstract quantity measuring extra negative factors such as laziness and learning difficulties. Unfortunately, E is intangible, and not available to the observer.

Finally, the random variable Y simply models the student's results: $Y = 1$ or $Y = 0$ according to whether he/she fails or passes a course. It is assumed that

$$Y = \begin{cases} 1 & \text{if } T + B + E < 7 \\ 0 & \text{otherwise.} \end{cases}$$

It is also assumed that T, B, and E are independent with an exponential distribution (with parameter 1). The Bayes rule associated with $((T, B), Y)$ is denoted by $g^\star(T, B)$.

1. What is $L^\star$, the Bayes risk associated with $((T, B, E), Y)$?
2. Give the expression of $\mathbb{P}(Y = 1|T, B)$.
3. Deduce from the above $g^\star(T, B)$.
4. What is the probability density of the random variable $T + B$?
5. Provide the numerical expression of $\mathbb{P}(g^\star(T, B) \neq Y)$.
6. What is the error incurred by a student who decides that $Y = 1$, independently of T and B?