Introduction to Statistical Learning

Mid-term exam

Duration : 2h - Lecture notes not allowed

Exercise 1 - Consider a random pair (X, Y) over $\mathbb{R}^d \times \{-1, +1\}$ and decision rules denoted by g. Find the minimizer g^* of the following error measures, as well as their minimum value $L(g^*)$:

1. Consider classifiers $g : \mathbb{R}^d \to \{-1, +1\}$ and the error measure L defined for any $\alpha, \beta > 0$ as follows :

$$L(g) = \mathbb{E}\big(\alpha \mathbb{I}\{Y = +1\}\mathbb{I}\{g(X) = -1\} + \beta \mathbb{I}\{Y = -1\}\mathbb{I}\{g(X) = +1\}\big)$$

2. Consider decision rules $g : \mathbb{R}^d \to \{-1, \Omega, +1\}$ and the error measure L defined for any $\gamma \in (0, 1/2)$ as follows :

$$L_d^R(g) = \mathbb{P}(Y \neq g(X)\ ,\ g(X) \neq \Omega) + \gamma \mathbb{P}(g(X) = \Omega)$$

Exercise 2 - Consider IID random pairs (X, Y) and (X', Y') over $\mathbb{R}^d \times \mathcal{Y}$. Set the following posterior probabilities :

$$\forall x, x' \in \mathbb{R}^d, \quad \begin{aligned} \rho_+(x, x') &= \mathbb{P}\{Y - Y' > 0 \mid X = x, X' = x'\} \\ \rho_-(x, x') &= \mathbb{P}\{Y - Y' < 0 \mid X = x, X' = x'\} \end{aligned}$$

and for any preference rule $\pi : \mathbb{R}^d \times \mathbb{R}^d \to \{-1, 0, 1\}$, consider the pairwise error measure

$$L(\pi) = \mathbb{P}\left\{(Y - Y') \cdot \pi(X, X') \leq 0\right\} \ .$$

1. Find the minimizer π^* and minimum $L^* = L(\pi^*)$ for this problem, as well as the excess of risk $L(\pi) - L^*$ for any preference rule π (will involve ρ_+ and ρ_-).
2. Assume $\mathcal{Y} = \{-1, +1\}$ and denote by $\eta(x) = \mathbb{P}\{Y = +1 \mid X = x\}$. Provide the expressions for $\rho_+(x, x')$ and $\rho_-(x, x')$ and discuss how the behavior of η could lead to difficult situations for the learning process to be efficient.
3. Assume now that $\mathcal{Y} = \mathbb{R}$ and that $Y = m(X) + \sigma(X) \cdot N$ where m and σ are P_X-measurable functions, N is a random noise variable with normal distribution $\mathcal{N}(0, 1)$, while N and X are independent random variables. Provide the expressions for $\rho_+(x, x')$ and $\rho_-(x, x')$ in this case and discuss the relation between properties of the model and the learning process.

Exercise 3

1. Consider Q a real-valued random variable such that : $\mathbb{E}(Q) = 0$ and $\mathbb{P}(Q \in [a,b]) = 1$. Prove the following upper bound : for any $s > 0$,

$$\mathbb{E}\big(e^{sQ}\big) \leq \exp\left(\frac{s^2(b-a)^2}{8}\right)$$

2. Consider $V = (V_1, \ldots, V_n, ...)$ and $Z = (Z_1, \ldots, Z_n, ...)$ two sequences of real-valued random variables. We assume the following : for any $n \geq 1$,

   - $V_n = \psi(Z_1, \ldots, Z_n)$ for some mesurable function ψ
   - $\mathbb{E}(V_{n+1} \mid Z_1, \ldots, Z_n) = 0$
   - there exists a sequence T_n which is measurable wrt the σ-algebra generated by $(Z_1, \ldots, Z_{n-1})$ and $c \geq 0$ such that : for any n, $T_n \leq V_n \leq T_n + c$

   Find the expression of $\kappa(t,n,c)$ such that, for any $t > 0$

$$\mathbb{P}\left(\sum_{i=1}^n V_i > t\right) \leq \kappa(t,n,c), \text{ and } \mathbb{P}\left(\sum_{i=1}^n V_i < -t\right) \leq \kappa(t,n,c) \ .$$

3. Consider a real-valued and measurable function h of n variables such that there exist $c > 0$ such that : for any $i \in \{1, \ldots, n\}$

$$\sup_{z_1,\ldots,z_n,z_i'} |h(z_1, \ldots, z_n) - h(z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n)| \leq c$$

   Assume that $Z_1, \ldots, Z_n$ are IID random variables. Prove that, for any $t > 0$

$$\mathbb{P}\left(h(Z_1, \ldots, Z_n) - \mathbb{E}\big(h(Z_1, \ldots, Z_n)\big) > t\right) \leq \kappa(t,n,c)$$

   and

$$\mathbb{P}\left(h(Z_1, \ldots, Z_n) - \mathbb{E}\big(h(Z_1, \ldots, Z_n)\big) < -t\right) \leq \kappa(t,n,c)$$

   where $\kappa(t,n,c)$ is as before.

4. Provide two examples of applications of the previous inequality that relevant to statistical learning theory.

Exercise 4 - Assume we have access to IID classification data $(X_1, Y_1) \ldots (X_n, Y_n)$ and consider a set of functions h over $\mathbb{R}^d$ parameterized by $\theta \in \mathbb{R}^d$ and defined as $h(x, \theta) = \sin(\theta^T x)$.

1. Propose a strategy to infer θ from the data.
2. How to assess the performance of the selected decision rule ?
3. What guarantees can be given on future performance of the selected decion rule ?