Introduction to Statistical Learning

Mid-term exam

Duration : 1h30 - No documents allowed

Reminder/Notations

— Law of iterated expectation : $\mathbb{E}(U) = \mathbb{E}(\mathbb{E}(U \mid V))$ where U, V are random variables.

— McDiarmid inequality : let h be a function of n variables $x_1, \ldots, x_n$ satisfying the uniform bounded differences assumption with constant $c, \ldots, c$: for any index i,

$$\sup_{x_1,\ldots,x_n,x_i'} |h(x_1, \ldots, x_n) - h(x_1, \ldots, x_{i-1}, x_i', x_{i+1} \ldots, x_n)| \leq c \ . \tag{1}$$

Then, we have that : for any $t > 0$,

$$\mathbb{P}\{h(X_1, \ldots, X_n) - \mathbb{E}\big(h(X_1, \ldots, X_n)\big) \geq t\} \leq \exp\left(-\frac{2t^2}{nc^2}\right) \ . \tag{2}$$

and

$$\mathbb{P}\{h(X_1, \ldots, X_n) - \mathbb{E}\big(h(X_1, \ldots, X_n)\big) \leq -t\} \leq \exp\left(-\frac{2t^2}{nc^2}\right) \ . \tag{3}$$

Exercise 1 - Consider the model for binary classification data where X is a random vector on $\mathbb{R}^d$ and Y is a random variable taking values in $\{0, +1\}$.
Denote by $\eta(x) = \mathbb{P}\{Y = +1 \mid X = x\}$ the posterior probability.

1. Consider the classification error $L(g) = \mathbb{P}\{Y \neq g(X)\}$ for $g : \mathbb{R}^d \to \{0, +1\}$. Find the optimal classifier g^* with respect to L. The proof of the optimality property has to be provided.
2. Fix $\lambda \in (0, 1)$. Now assume that we aim at minimizing $\widetilde{L}(g) = L(g) + \lambda \mathbb{P}\{g(X) = +1\}$. Find the optimal classifier g_λ^* in this case and comment on the extremal cases ($\lambda \to 0$ and $\lambda \to 1$).
3. Illustrate the interest of the previous formulation of minimizing a quantity such as $\widetilde{L}(g)$: (i) What type of constraint over the class of possible classifiers can be taken into account, and (ii) What could be the practical interest when modeling real-life classification problems.

Exercise 2 - We consider the notations $S = (z_1, \dots, z_n)$ and $S_i = (z_1, \dots, z_{i-1}, z_i', z_{i+1}, z_n)$. We recall that the function h has bounded differences if it satisfies, for some constants c_i :

$$\forall i \in \{1, \dots, n\} \ , \qquad \sup_{S, z_i'} |h(S) - h(S_i)| \leq c_i \ .$$

We consider Z a random variable with distribution P and $Z_1, \dots, Z_n$ are IID copies of Z. We also introduce $\epsilon = (\epsilon_1, \dots, \epsilon_n)$ where the ϵ_i are Rademacher random variables (with uniform distribution over $\{-1, +1\}$). We denote by $\mathcal{F}$ a class of functions from $\mathbb{R}^d$ to $\mathbb{R}$ such that $\sup_{f,x} |f(x)| \leq 1$ and we set :

- $\square$ $P(f) = \mathbb{E}\big(f(Z)\big)$ et $P_n(f) = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} f(Z_i)$
- $\square$ $P_n^{\epsilon}(f) = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \epsilon_i f(Z_i)$.

1. Show that the function h defined by :

$$h(S) = \sup_{f \in \mathcal{F}} (P(f) - P_n(f)$$

   has bounded differences and provide the values of constants c_i.

2. Same question with :

$$h_R(S) = \mathbb{E}_{\epsilon} \left(\sup_{f \in \mathcal{F}} P_n^{\epsilon}(f) \right) \ .$$

3. Show that : $\mathbb{E}\big(h(S)\big) \leq c \, \mathbb{E}\big(h_R(S)\big)$ where c is a constant to be determined.
4. For any $\delta > 0$, provide a bound with probability at least $1 - (\delta/2)$ of the quantity $h(S)$ in terms of $\mathbb{E}\big(h_R(S)\big)$ up to an additive term, this corrective term depending on δ and n exclusively.
5. Show that : $\mathbb{E}\big(h_R(S)\big)$ is controlled by the quantity $h_R(S)$ (up to a corrective term) with probability at least $1 - (\delta/2)$.
6. Derive an upper bound on $h(S)$ in terms of $h_R(S)$ with probability at least $1 - \delta$.
7. What is the interest of such quantities and bounding technique for statistical learning problems ?

Exercise 3 - We consider the setup of binary classification where X is a random vector over $\mathbb{R}^d$ and Y is a random variable taking values in $\{-1, +1\}$.
We denote $\eta(x) = \mathbb{P}\{Y = 1 \mid X = x\}$.

1. Express the minimizing function of the criterion $A(f) = \mathbb{E}(\log_2(1 + e^{-Yf(X)}))$ as a function of η in the following cases :
   (i) among functions $f : \mathbb{R}^d \to \{-1, +1\}$
   (ii) among functions $f : \mathbb{R}^d \to \mathbb{R} \cup \{-\infty, +\infty\}$
2. In the case (ii), compute the minimum of $A(f)$.
3. Explain why the minimizers obtained are relevant if the criterion of interest is the classification error ?
