Introduction to Statistical Learning

Mid-term exam

Duration : 1h30 - No documents allowed

Reminder/Notations

- IID means Independent and Identically Distributed.
- If U and V are independent random variables, then $\mathbb{E}(UV) = \mathbb{E}(U)\mathbb{E}(V)$.
- Law of iterated expectation : $\mathbb{E}(U) = \mathbb{E}(\mathbb{E}(U \mid V))$ where U, V are random variables.
- Jensen's inequality : let ψ convex function, then we have $\psi(\mathbb{E}(U)) \leq \mathbb{E}\big(\psi(U)\big)$.
- Cauchy-Schwarz inequality : for any vectors u, v, we have $u^T v \leq \|u\| \cdot \|v\|$ where u^T is the transpose vector of vector u.
- Subadditivity of supremum operator : $\sup(f + g) \leq \sup(f) + \sup(g)$ and $\sup(f) - \sup(g) \leq \sup(f - g)$.
- McDiarmid inequality : let h be a function of n variables $x_1, \dots, x_n$ satisfying the bounded differences assumption with constants $c_1, \dots, c_n$: for any index i,

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

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

$$\mathbb{P}\{|h(X_1, \dots, X_n) - \mathbb{E}\big(h(X_1, \dots, X_n)\big)| \geq t\} \leq 2 \exp\left(-\frac{2t^2}{c_1^2 + \dots + c_n^2}\right) \ .$$

Exercise 1 - Let $\mathcal{G}$ a class of functions from $\mathbb{R}^d$ to $[-B, B]$, with $B > 0$. Consider random sign variables $\varepsilon_1, \dots, \varepsilon_n$ IID such that $\mathbb{P}\{\varepsilon_1 = -1\} = \mathbb{P}\{\varepsilon_1 = +1\} = 1/2$. Consider the *empirical Rademacher complexity* defined as

$$\widehat{R}_n(\mathcal{G}) = \frac{1}{n}\mathbb{E}\left(\sup_{g\in\mathcal{G}} \sum_{i=1}^n \varepsilon_i g(X_i) \,\middle|\, X_1, \dots, X_n\right)$$

and the *average Rademacher complexity* as :

$$\overline{R}_n(\mathcal{G}) = \frac{1}{n}\mathbb{E}\left(\sup_{g\in\mathcal{G}} \sum_{i=1}^n \varepsilon_i g(X_i)\right)$$

1. Show that for fixed g, the empirical Rademacher complexity seen as a function of $X_1, \dots, X_n$ satisfies the bounded differences condition.
2. Provide an upper bound on the average Rademacher complexity in terms of the empirical Rademacher complexity that holds with high probability.

3. Consider a sample $\{X_1, \ldots, X_n\}$ of points included in the closed ball $\{x \in \mathbb{R}^d \; : \; \|x\| \leq M\}$ (with respect to Euclidean distance $\| \cdot \|$). Let $\mathcal{G}$ be the class of linear rules defined as $\{x \mapsto w^T x \; : \; w \in \mathbb{R}^d, \; \|w\| \leq L\}$. Derive an upper bound on the empirical Rademacher complexity that involves M, L and n.

Exercice 2 - Consider (X, Y) a random pair that models classification data with labels in $\{0, 1\}$.

1. For a classifier $g \; : \; \mathbb{R}^d \to \{0, 1\}$, define $L(g) = \mathbb{P}\{Y \neq g(X)\}$. What is the minimizing argument of $L(g)$ over all possible classifiers g ?
2. Now define $L_c(g) = c_1 \mathbb{P}\{Y \neq g(X), Y = 1\} + c_0 \mathbb{P}\{Y \neq g(X), Y = 0\}$. What is the minimizing argument of $L_c(g)$ over all possible classifiers g ?
3. We now consider classifiers with reject option $g \; : \; \mathbb{R}^d \to \{R, 0, 1\}$, and $L_R(g) = \mathbb{P}\{Y \neq g(X), g(X) \neq R\} + c\mathbb{P}\{g(X) = R\}$. What is the minimizing argument of $L_R(g)$ over all possible classifiers g with reject option ? Give a practical interpretation of the result.

Exercice 3 - We consider the model for classification data where X is a random vector on $\mathbb{R}^d$ and Y is a random variable taking values in $\{-1, +1\}$. We denote by $\eta(x) = \mathbb{P}\{Y = +1 \mid X = x\}$ the posterior probability.

1. We consider the following problems for which the question is to compute the optimal decision rule g^* or f^*.
   (a) $R(g) = \mathbb{E}\big((Y - g(X))^2\big)$ where $g \; : \; \mathbb{R}^d \to \{-1, +1\}$
   (b) $R(f) = \mathbb{E}\big((Y - f(X))^2\big)$ where $f \; : \; \mathbb{R}^d \to \mathbb{R}$
   (c) $A(f) = \mathbb{E}\big(\log_2(1 + e^{-Yf(X)})\big)$ where $f : \mathbb{R}^d \to \mathbb{R} \cup \{-\infty, +\infty\}$ and $\log_2(2) = 1$.
2. Explain why such criteria are relevant for the binary classification problem.
3. In the case of 1.(c), compute the function H such that : $A(f^*) = \mathbb{E}(H(\eta(X)))$.
4. Plot H and state its main properties. Compare $(u - 1/2)^2$ and $(1 - H(u))$.
5. Consider $L(g) = \mathbb{P}\{Y \neq g(X)\}$ and $L* = \inf_g L(g)$. What upper bound can be given on the quantity $L(\text{sgn}(f)) - L^*$?
