Introduction to Statistical Learning

Final exam

Duration : 2h - Lecture notes allowed

Exercise 1 - Rademacher complexity and multiclass classification

A. We recall that given $X_1, \dots, X_n$ random vectors on $\mathbb{R}^d$ and $\mathcal{F}$ being a class of bounded real-valued functions, the empirical Rademacher average is defined as the random quantity :

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

where $\varepsilon_1, \dots, \varepsilon_n$ are IID random sign variables such that $\mathbb{P}\{\varepsilon_1 = -1\} = \mathbb{P}\{\varepsilon_1 = +1\} = 1/2$. We also admit the following result : let $\Psi : \mathbb{R} \to \mathbb{R}$ be a Lipschitz function (*i.e.* $\exists L : \forall u, v \in \mathbb{R},\ |\Psi(u) - \Psi(v)| \leq L|u - v|$), then we have :

$$\widehat{R}_n(\Psi \circ \mathcal{F}) \leq L\widehat{R}_n(\mathcal{F})$$

(a) What is the Lipschitz constant L for the function $u \mapsto |u|$?

(b) Consider two classes of bounded real-valued functions $\mathcal{F}_1$, $\mathcal{F}_2$. Find a simple upper bound of the following quantity :

$$\frac{1}{n}\mathbb{E}\left(\sup_{f_1\in\mathcal{F}_1, f_2\in\mathcal{F}_2} \sum_{i=1}^{n} \varepsilon_i |f_1(X_i) - f_2(X_i)| \middle| X_1, \dots, X_n \right)$$

depending on $\widehat{R}_n(\mathcal{F}_1)$, $\widehat{R}_n(\mathcal{F}_2)$.

(c) Express $\max\{f_1, f_2\}$ as a linear relation involving $|f_1 - f_2|$.

(d) Consider the class $\mathcal{F} = \{\max\{f_1, f_2\} : f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}$ and provide a simple upper bound of $\widehat{R}_n(\mathcal{F})$ depending on $\widehat{R}_n(\mathcal{F}_1)$, $\widehat{R}_n(\mathcal{F}_2)$.

B. Consider a multiclass classification problem with observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ IID copies of the random pair (X, Y) where the output variable Y takes values in $\{1, \ldots, K\}$. The decision rules are functions g_h of the form :

$$g_h \ : \ x \mapsto \underset{y \in \{1,\ldots,K\}}{\arg\max} \ h(x, y)$$

where h is a real-valued function in a class $\mathcal{H}$ of functions over the set $\mathbb{R}^d \times \{1, \ldots, K\}$. The complexity of learning in the multiclass classification setup relies on the complexity of the class $\mathcal{H}$ that will be considered here under the margin approach. We thus define the margin ρ_h of function h as :

$$(x, y) \mapsto \rho_h(x, y) = h(x, y) - \max_{y' \neq y} h(x, y') \ ,$$

and ρ_h belongs to the class $\mathcal{H}_\rho$ of functions induced by $\mathcal{H}$.

(a) Set the empirical Rademacher complexity of the class $\mathcal{H}_\rho$ to be :

$$\widehat{R}_n(\mathcal{H}_\rho) \leq \frac{1}{n}\mathbb{E}\left(\sup_{h \in \mathcal{H}} \sum_{i=1}^{n} \varepsilon_i \rho_h(X_i, Y_i) \middle| (X_1, Y_1), \ldots, (X_n, Y_n) \right) \ .$$

Note that, for any Λ, we have that :

$$\Lambda(X_i, Y_i) = \sum_{y=1}^{K} \Lambda(X_i, y)\mathbb{I}\{y = Y_i\} = \sum_{y=1}^{K} \Lambda(y) \left(\frac{2\mathbb{I}\{y = Y_i\} - 1}{2} + \frac{1}{2} \right) \ ,$$

and show that :

$$\widehat{R}_n(\mathcal{H}_\rho) \leq \frac{1}{n} \sum_{y=1}^{K} \mathbb{E}\left(\sup_{h \in \mathcal{H}} \sum_{i=1}^{n} \varepsilon_i \rho_h(X_i, y) \middle| X_1, \ldots, X_n \right)$$

(b) Set $\mathcal{H}_X = \{x \mapsto h(x, y) \ : \ y \in \{1, \ldots, K\}, \ h \in \mathcal{H}\}$. Using the definition of ρ_h and the main result of Part A, prove that :

$$\widehat{R}_n(\mathcal{H}_\rho) \leq K^\alpha \widehat{R}_n(\mathcal{H}_X)$$

where α will be made explicit.

(c) Set $\varphi_\gamma(u) = (1 - u/\gamma)\mathbb{I}\{u \in (0, \gamma]\} + \mathbb{I}\{u \geq 0\}$ and compute its Lipschitz constant.

(d) Relate the multiclass classification error $L(g_h) = \mathbb{P}\{Y \neq g_h(X)\}$ to the multiclass margin error $L_\gamma(h) = \mathbb{E}\{\varphi_\gamma\big(\rho_h(x, y)\big)\}$.

(e) We introduce $\widehat{L}_\gamma(h) = \frac{1}{n}\sum_{i=1}^{n} \varphi_\gamma(\rho_h(X_i, Y_i))$. Use a concentration inequality to derive an upper bound on the quantity :

$$\sup_{h \in \mathcal{H}} (L_\gamma(h) - \widehat{L}_\gamma(h)) \ .$$

(f) Give a sketch of proof that the following inequality holds, with probability at least $1 - \delta$, for any $h \in \mathcal{H}$:

$$L(g_h) \leq \widehat{L}_\gamma(h) + c_1(K, \gamma)\mathbb{E}(\widehat{R}_n(\mathcal{H}_X)) + c_2(n, \delta)$$

where c_1 and c_2 will have to be computed explicitly.

Exercise 2 - VC dimension

1. Recall the definition of the VC dimension of a class of subsets of $\mathbb{R}^d$.
2. What is the VC dimension of the class of hyperplanes of $\mathbb{R}^d$?
3. Compute the VC dimension of the class $\mathcal{A}$ of orthants of $\mathbb{R}^2$:

$$\mathcal{A}_2 = \{\ ]-\infty, a_1] \times]-\infty, a_2]\ :\ (a_1, a_2) \in \mathbb{R}^2\}\ .$$

4. Generalize the previous result to the VC dimension of the class $\mathcal{A}_d$ of orthants of $\mathbb{R}^d$.
5. Back to $d = 2$. Let the class of piecewise constant functions :

$$\mathcal{G} = \left\{ \sum_{k=1}^{K} w_k \mathbb{I}\{A_k\}\ :\ \ K \geq 1, w_k > 0, A_k \in \mathcal{A}_2 \right\}\ .$$

We consider the class of subsets of $\mathbb{R}^2$ determined by :

$$\mathcal{C} = \left\{ C_g = \{x \in \mathbb{R}^2\ :\ \ g(x) > 1\}\ :\ g \in \mathcal{G} \right\}\ .$$

What is the VC dimension of $\mathcal{C}$? Indication : consider n points on a line of $\mathbb{R}^2$ with negative slope and crossing the origin.

Exercise 3 - Convex risk minimization

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 ?