Introduction to Statistical Learning

Final exam

Duration : 2h - Lecture notes allowed

Notations

— **Indicator function**. The indicator function $\mathbb{I}\{\Omega\}$ takes the value 1 if Ω is true, and 0 otherwise.

— **Empirical Rademacher average**. Consider an IID sample $Z_1^n = (Z_1, \dots, Z_n)$ and let $\sigma_1, \dots, \sigma_n$ an i.i.d. sample of Rademacher random variables ($\mathbb{P}\{\sigma_1 = +1\} = \mathbb{P}\{\sigma_1 = -1\} = 1/2$) independent of Z_1^n. Given a class $\mathcal{T}$ of functions, we denote its empirical Rademacher average by :

$$\hat{R}_n(\mathcal{T}) = \mathbb{E}\left(\sup_{t\in\mathcal{T}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i t(Z_i) \mid Z_1^n\right)$$

— **Kernel function - definitions and properties.** Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a positive definite and symmetric kernel function. We recall that k has the property that there exist : (i) a Hilbert space $\mathcal{H}$ equipped with norm $\|\cdot\|_k$ and (ii) a feature mapping $\Phi : \mathbb{R}^d \to \mathcal{H}$ such that $k(x, x') =< \Phi(x), \Phi(x') >_k$ and $k(x, x) = \|\Phi(x)\|_k$ for any x, x'. Given a sample $X_1, \dots, X_n$, we denote by $K = \big(k(X_i, X_j)\big)_{1\le i,j\le n}$ the Gram matrix induced by the kernel function k.

— **Dual norm.** Consider a norm $\|\cdot\|$ on $\mathbb{R}^d$. Then its dual norm $\|\cdot\|_*$ is defined, for any $z \in \mathbb{R}^d$, as $\|z\|_* = \sup\{x^T z : x \in \mathbb{R}^d, \|x\| \le 1\}$. Moreover, we have the inequality : $x^T z \le \|x\| \cdot \|z\|_*$.

Exercise 1 - Consider an IID sample $X_1, \dots, X_n$ of random vectors in $\mathbb{R}^d$.

1. Consider $\mathcal{G}$ a class of functions with values in $\{-1, +1\}$ and its empirical Rademacher average $\hat{R}_n(\mathcal{G})$, and let $\mathcal{L}$ the class of classification loss functions :

   $$\mathcal{L} = \{(x, y) \mapsto \mathbb{I}\{g(x) \neq y\} : g \in \mathcal{G}\} .$$

   Assume an IID sample of pairs $(X_1, Y_1), \dots, (X_n, Y_n)$ is available. What is the relation between $\hat{R}_n(\mathcal{L})$ and $\hat{R}_n(\mathcal{G})$? Provide the proof of this relation.
2. Consider the class of linear functions $\mathcal{F}_1 = \{x \mapsto w^T x : w \in \mathbb{R}^d, \|w\|_2 \le M_1\}$ and find an upper bound for the empirical Rademacher average $\hat{R}_n(\mathcal{F}_1)$ in terms of M_1, n, and $\sum_{i=1}^n \|X_i\|_2^2$.
3. Consider the class of linear functions $\mathcal{F}_2 = \{x \mapsto w^T x : w \in \mathbb{R}^d, \|w\|_1 \le M_2\}$ and assume that, for any i, we have $\|X_i\|_\infty \le r$ almost surely. Find an upper bound for the empirical Rademacher average $\hat{R}_n(\mathcal{F}_2)$ in terms of M_2, n, r and d.
4. Consider the class of functions $\mathcal{F}_3 = \{x \mapsto < w, \Phi(x) > : w \in \mathcal{H}, \|w\|_k \le M_3\}$, and find an upper bound for $\hat{R}_n(\mathcal{F}_3)$ which depends on M_3, n, and K. Provide a simple condition on the kernel k such that the behavior of $\hat{R}_n(\mathcal{F}_3)$ as a function of n is at most $O(n^{-1/2})$.

Exercice 2 - Consider the setup of preference learning where we observe an IID sample of triples $(X_1, X_1', Y_1), \ldots, (X_n, X_n', Y_n)$. The probabilistic model assumes that, for each i, the triple (X_i, X_i', Y_i) is such that X_i, X_i' are IID random vectors over $\mathbb{R}^d$ and Y_i is a random variable over $\{-1, 0, +1\}$. We define the ranking error of a preference rule $g : \mathbb{R}^d \to \{-1, 0, +1\}$ as :

$$L(g) = \mathbb{P}\{Y \neq 0,\ Y \cdot (g(X') - g(X)) \leq 0\}$$

ans the empirical margin ranking error as :

$$\widehat{L}_{n,\rho}(g) = \frac{1}{n} \sum_{i=1}^{n} m_\rho(Y_i \cdot (g(X_i') - g(X_i))$$

where the margin loss is defined, for any $\rho > 0$, by

$$m_\rho(t) = \mathbb{I}\{t \leq 0\} + \mathbb{I}\{0 \leq t \leq \rho\} \left(1 - \frac{t}{\rho}\right) .$$

Now consider a class $\mathcal{G}$ of preference rules and define :

$$\widetilde{\mathcal{G}} = \{(x, x', y) \mapsto y(g(x') - g(x)) \ : \ g \in \mathcal{G}\} .$$

1. Provide an upper bound of the empirical Rademacher average of $\widetilde{\mathcal{G}}$ in terms of the empirical Rademacher average of $\mathcal{G}$.
2. Show that m_ρ is Lipschitz and provide its Lipschitz constant.
3. Which inequality relates the empirical Rademacher average of the loss class $m_\rho \circ \widetilde{\mathcal{G}}$ to the empirical Rademacher average of $\widetilde{\mathcal{G}}$? Provide a proof of this inequality.
4. Show that, for any $\delta \in (0, 1)$, we have, with probability at least $1 - \delta$: for any $g \in \mathcal{G}$

   $$\mathbb{E}(m_\rho(y(g(x') - g(x))) \leq \widehat{L}_{n,\rho}(g) + c_1 \hat{R}_n(m_\rho \circ \tilde{\mathcal{G}}) + c_2(n, \delta)$$

   for some c_1 and $c_2(n, \delta)$ that will have to be given explicitly.
5. Deduce from the previous question a margin error bound for $L(g)$ that holds with large probability for any $g \in \mathcal{G}$ and which involves the empirical ranking error of g over the sample and the complexity of $\mathcal{G}$.
6. Specify the previous result to the case of a kernel class of functions with $\mathcal{G} = \mathcal{F}_3$ as defined in **Exercise 1**.
7. What kind of algorithm can be justified by the inequalities obtained in the two previous questions.
8. Propose an algorithm based on convex risk minimization for the preference learning problem. Formulate it precisely in pseudocode and explain what theoretical justification could be provided.

Exercice 3 -

1. Recall the definition of the VC dimension.
2. How can the VC dimension be related to the Rademacher average ?
3. Compute the VC dimension $V(\mathcal{A})$ in the following cases :
   (a) $\mathcal{A}$ is the class of all half-spaces in $\mathbb{R}^d$.
   (b) $\mathcal{A} = \{\ ]-\infty, x_1] \times \ldots \times]-\infty, x_d]\ :\ (x_1, \ldots, x_d) \in \mathbb{R}^d\ \}$.
   (c) $\mathcal{A}$ is the class of all rectangles of $\mathbb{R}^2$ with axis-orthogonal edges.
   (d) $\mathcal{A}$ is the class of all rectangles of $\mathbb{R}^2$.
4. Is the notion of VC dimension relevant to show the consistency of machine learning algorithms ?