Introduction to Statistical Learning

Final exam (3 pages)

Duration : 2h00 - 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 $Z_1^n = (Z_1, \ldots, Z_n)$ and let $\sigma_1, \ldots, \sigma_n$ an IID 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 scalar product $< \cdot, \cdot >_k$ and 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 $X_1, \ldots, X_n$, we denote by $K = \big(k(X_i, X_j)\big)_{1\leq i,j\leq n}$ the Gram matrix induced by the kernel function k.

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

Exercice 2 - Consider an IID sample of Rademacher random variables ($\mathbb{P}\{\sigma_1 = +1\} = \mathbb{P}\{\sigma_1 = -1\} = 1/2$).

1. Consider a random variable X such that $\mathbb{E}(X) = 0$ and $X \in [a, b]$ almost surely. Give a sketch of proof evoking the main arguments of the following result : for any $t > 0$, we have :
$$\mathbb{E}(e^{tX}) \leq e^{t^2(b-a)^2/8}$$
2. Consider $Q \subset \mathbb{R}^k$ a finite set of points. We assume that they are all contained in the Euclidean ball with center the origin and radius R. Then show that : for any $t > 0$
$$\mathbb{E}\left(\sup_{q=(q_1,\ldots,q_k)\in Q} \sum_{i=1}^{k} \sigma_i q_i\right) \leq \frac{tR^2}{2} + \frac{\log|Q|}{t}$$
where $|Q|$ is the number of points in Q.
3. Provide the optimal choice of t in the previous question and give the expression of the optimal bound.

Exercice 3 - Consider the following :

— $D_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$ an IID sample of supervised training data over $\mathcal{X} \times \mathcal{Y}$,

— $\mathcal{F}$ a class of predictors from $\mathcal{X}$ to $\mathcal{Y}$,

— $A : D_n \mapsto \hat{f}_n \in \mathcal{F}$ a learning algorithm,

— $\ell : \mathcal{Y}^2 \to \mathbb{R}_+$ a cost function such that $\ell(y, y') \leq \Lambda$ for any $y, y' \in \mathcal{Y}$, with $\Lambda > 0$,

— $L(\hat{f}) = \mathbb{E}(\ell(Y, \hat{f}(X)) \mid D_n)$ is the risk of any data-driven predictor $\hat{f}$,

— $\hat{L}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i))$ is the empirical risk of any predictor $f \in \mathcal{F}$.

We consider the notation D'_n for a sample of size n which differs from D_n by a single point, and $\hat{f}'_n = A(D'_n)$. We assume that, for any n, there exists a $\beta_n \geq 0$ such that for any samples D_n and D'_n and for any pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we have : $|\ell(y, \hat{f}_n(x)) - \ell(y, \hat{f}'_n(x))| \leq \beta_n$.

1. Find an upper bound on $|L(\hat{f}_n) - L(\hat{f}'_n)|$ depending on β_n.
2. Find an upper bound on $|\hat{L}_n(\hat{f}_n) - \hat{L}_n(\hat{f}'_n)|$ depending on β_n, Λ and n.
3. Show that the quantity $L(\hat{f}_n) - \hat{L}_n(\hat{f}_n)$ satisfies the bounded differences condition and apply a well-known concentration inequality.
4. Then, show that we have, with probability at least $1 - \delta$:
$$L(\hat{f}_n) \leq \hat{L}_n(\hat{f}_n) + \beta_n + (2n\beta_n + \Lambda)\sqrt{\frac{\log(1/\delta)}{2n}}$$
5. What would be an appropriate order of magnitude for the coefficient β_n ? Can you give examples of algorithms that would display such values for β_n ?

Exercice 4 - 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 :

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

and the empirical margin ranking error as :

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

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. Which inequality relates the empirical Rademacher average of the loss class $\varphi_\rho \circ \widetilde{\mathcal{G}}$ to the empirical Rademacher average of $\widetilde{\mathcal{G}}$? Provide a proof of this inequality.
3. Show that, for any $\delta \in (0, 1)$, we have, with probability at least $1 - \delta$: for any $g \in \mathcal{G}$

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

   for some c_1 and $c_2(n, \delta)$ that will have to be given explicitly.
4. Deduce from the previous question a margin error bound for $L^R(g)$ 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 $\mathcal{G}$.
5. Specify the previous result to the case of a kernel class of functions with $\mathcal{G} = \mathcal{F}_M$ as defined in **Exercise 1**.
6. Which algorithms can be justified by the inequalities obtained in the two previous questions.