Introduction to Statistical Learning

Exercise sheet n°3

Definitions

Let $\mathcal{F}$ be a class of bounded real-valued functions and $\mathcal{A}$ a class of subsets of $\mathbb{R}^d$.

— The empirical Rademacher complexity of $\mathcal{F}$ wrt to the sample $D_n = \{Z_1, \dots, Z_n\}$ is defined as :

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

— The Rademacher complexity of $\mathcal{F}$ is defined as :

$$R_n(\mathcal{F}) = \mathbb{E}\big(\widehat{R}_n(\mathcal{F})\big)$$

— Trace $\mathrm{Tr}(\mathcal{A}, \mathbf{x}_1^n)$ of $\mathcal{A}$ over a set of point $\mathbf{x}_1^n = \{x_1, \dots, x_n\}$ in $\mathbb{R}^d$:

$$\mathrm{Tr}(\mathcal{A}, \mathbf{x}_1^n) = \{A \cap \mathbf{x}_1^n \ : \ A \in \mathcal{A}\}$$

— Growth function $n \mapsto \gamma(\mathcal{A}, n)$ of $\mathcal{A}$

$$\gamma(\mathcal{A}, n) = \max_{\mathbf{x}_1^n} |\mathrm{Tr}(\mathcal{A}, \mathbf{x}_1^n)|$$

where $|\cdot|$ denotes the cardinality of the set.

— Vapnik-Chervonenkis dimension $V(\mathcal{A})$ or VC dimension of $\mathcal{A}$

$$V(\mathcal{A}) = \max n \in \mathbb{N} \ : \ s(\mathcal{A}, n) = 2^n$$

— Massart's Lemma :

Let C be a finite subset of $\mathbb{R}^n$ and $R = \max_{z \in C} \|z\|_2$, where $z = (z_1, \dots, z_n)$ and $\varepsilon_1, \dots, \varepsilon_n$ a sample of i.i.d. Rademacher random variables. Then

$$\mathbb{E}\left[\sup_{z \in C} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i z_i \right] \leq \frac{R\sqrt{2 \log |C|}}{n}$$

Exercise 1 - (Properties of Rademacher averages) Let $\mathcal{T}$, $\mathcal{T}_1$, $\mathcal{T}_2$, be classes of real-valued functions. Prove the following properties :

1. If $c \in \mathbb{R}$, then $\hat{R}_n(c\mathcal{T}) = |c|\hat{R}_n(\mathcal{T})$.
2. If $\mathcal{T}_1 \subseteq \mathcal{T}_2$, then $\hat{R}_n(\mathcal{T}_1) \leq \hat{R}_n(\mathcal{T}_2)$
3. $\hat{R}_n(\mathcal{T}_1 + \mathcal{T}_2) = \hat{R}_n(\mathcal{T}_1) + \hat{R}_n(\mathcal{T}_2)$
4. Let $\text{conv}(\mathcal{T})$ be the convex hull of $\mathcal{T}$. Prove that : $\hat{R}_n\big(\text{conv}(\mathcal{T})\big) = \hat{R}_n(\mathcal{T})$.
5. If $\psi : \mathbb{R} \to \mathbb{R}$ is k-Lipschitz, then $\hat{R}_n(\psi \circ \mathcal{T}) \leq k\hat{R}_n(\mathcal{T})$.

Exercise 2 - (Rademacher average for linear and kernel classes)

1. Consider $\mathcal{B}_\infty(C) = \{x \in \mathbb{R}^d \mid \|x\|_\infty \leq C\}$ and $\mathcal{G} = \{x \in \mathcal{B}_\infty(C) \mapsto w^T x : \|w\|_1 \leq B\}$. Show that the following bound holds :

$$\hat{R}_n(\mathcal{G}) \leq \frac{BC\sqrt{2\ln(2d)}}{\sqrt{n}} .$$

2. Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ a positive definite and symmetric kernel function with feature mapping Φ that is : for any (x, x'), we have $K(x, x') =< \Phi(x), \Phi(x') >$ where $<,>$ is the product of some Hilbert space. Given a sample $X_1, \ldots, X_n$, define its Gram matrix as $K = \big(k(X_,, X_j)\big)_{1 \leq i,j \leq n}$. Consider the class of functions $\mathcal{H} = \{x \mapsto < w, \Phi(x) > : \|w\|_k \leq M\}$, and prove that :

$$\hat{R}_n(\mathcal{H}) \leq \frac{M\sqrt{\text{Tr}(K)}}{n} .$$

Exercise 3 - (Rademacher average for neural networks) Consider an i.i.d. sample $X_1, \ldots, X_n$ of observations over some space $\mathcal{X}$ and $\mathcal{F}_0$ is a set of real-valued functions over $\mathcal{X}$ that includes the zero function. Assume $\psi : \mathbb{R} \to \mathbb{R}$ is k-Lispchitz and define, for fixed positive real numbers V and B :

— a one layer network as : $\mathcal{F}_1 = \{x \mapsto \psi(v + \sum_{j=1}^m w_j f_j(x)) : |v| \leq V, \|w\|_1 \leq B, f_j \in \mathcal{F}_0\}$

— a p-layer network as (iterative definition with fixed layer size) : $\mathcal{F}_p = \{x \mapsto \psi(v + \sum_{j=1}^m w_j f_j(x)) : |v| \leq V, \|w\|_1 \leq B, f_j \in \mathcal{F}_{p-1}\}$

Prove the following upper bounds on the empirical Rademacher average :

1. $\hat{R}_n(\mathcal{F}_1) \leq k\left(\frac{V}{\sqrt{n}} + 2B\hat{R}_n(\mathcal{F}_0)\right)$.
2. Assume in addition that $\psi(-u) = -\psi(u)$ and $k = 1$ then show that on $\mathcal{X} = \{x \in \mathbb{R}^d, \ \|x\|_\infty \leq 1\}$:

$$\hat{R}_n(\mathcal{F}_p) \leq \frac{1}{\sqrt{n}}\left(B^p\sqrt{2\ln(2d)} + V\sum_{l=1}^{p-1} B^l\right) .$$

Exercise 4 - (Relation between Rademacher average and combinatorial complexities) Consider a class $\mathcal{G}$ of binary valued functions. Show that :

$$\overline{R}_n(\mathcal{G}) \leq \sqrt{\frac{2\ln(\gamma(\mathcal{G}, n))}{n}}$$

and if $V(\mathcal{G}) < +\infty$ then show that, for some constant C, we have :

$$\overline{R}_n(\mathcal{G}) \leq C\sqrt{\frac{V(\mathcal{G})\log(n)}{n}}$$

Exercise 5 - (Relation between Rademacher average and metric complexities) Consider two sets A and D in a pseudometric space equipped with pseudometric d and $\epsilon > 0$. The set A is called an ϵ-cover of D, if for any $u \in D$, there exists some $a \in A$ such that $d(u, a) \leq \epsilon$. The ϵ-covering number of D, denoted by $N(\epsilon, D)$, is the smallest cardinality among all ϵ-covers of D.

Now, consider $\mathcal{F}$ the pseudometric space of real-valued functions over $\mathbb{R}^d$ with the empirical $L_2(X)$ semi-norm over a sample $X_1, \ldots, X_n$ and such that $\gamma_0 = \sup_{f\in\mathcal{F}} \|f\|_{2,X}$. Then, show that the following bound holds :

$$\hat{R}_n(\mathcal{F}) \leq \inf_{\epsilon\in[0,\gamma/2)} \left\{ 4\epsilon + \frac{12}{\sqrt{n}} \int_{\epsilon}^{\gamma_0} \sqrt{\ln N(u, \mathcal{F})}\, du \right\}$$

Deduce that, if $\mathcal{F}$ is a class of binary-valued functions with finite VC dimension V, that we obtain

$$\hat{R}_n(\mathcal{F}) \leq C\sqrt{\frac{V}{n}}$$

for some universal constant C.

Hint : Use that $N(\epsilon, D) \leq \left(\frac{9}{\epsilon^2}\log\frac{2e}{\epsilon^2}\right)^V$