

Introduction to Statistical Learning

Exercise sheet n°2

DEFINITIONS

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

- Bounded differences function - A real-valued function h of n variables over a metric space \mathcal{Z} is said to be a function with bounded differences if there exist $c_1, \dots, c_n > 0$ such that :

$$\sup_{z_1, \dots, z_n, z'_i \in \mathcal{Z}} |h(z_1, \dots, z_n) - h(z_1, \dots, z_{i-1}, z'_i, z_{i+1}, \dots, z_n)| \leq c_i$$

- McDiarmid's concentration inequality - Assume h is a function with bounded differences with bounding constants c_1, \dots, c_n then, we have, for any $t > 0$

$$\mathbb{P}(h(Z_1, \dots, Z_n) - \mathbb{E}(h(Z_1, \dots, Z_n)) > t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right)$$

and

$$\mathbb{P}(h(Z_1, \dots, Z_n) - \mathbb{E}(h(Z_1, \dots, Z_n)) < -t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right)$$

- 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 :

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

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

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

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

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

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

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

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

Exercise 1 - (Hoeffding's lemma) Consider Z a random variable such that : $\mathbb{E}(Z) = 0$ and $\mathbb{P}(Z \in [a, b]) = 1$ almost surely. Prove the following upper bound : for any $s > 0$,

$$\mathbb{E}(e^{sZ}) \leq \exp\left(\frac{s^2(b-a)^2}{8}\right)$$

Exercise 2 - (Bernstein inequality) Let X_1, \dots, X_n independent random variables that are centered and uniformly bounded ($\forall i, |X_i| \leq c$ p.s.). Denote by $\sigma^2 = \frac{1}{n} \text{Var}(X_i)$. Show that $\forall t > 0$:

$$\mathbb{P}\left\{\sum_{i=1}^n X_i > t\right\} \leq \exp\left\{-\frac{t^2}{2n\sigma^2 + 2ct/3}\right\}.$$

Comment on the two typical behaviors of the tail probability of the empirical average that are captured by this exponential upper bound. Compare to Hoeffding's inequality.

Exercise 3 - (Application of McDiarmid's concentration inequality) Let \mathcal{F} be a class of $[0, 1]$ -valued functions. Show that, with probability at least $1 - \delta$:

$$R_n(\mathcal{F}) \leq \widehat{R}_n(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{2n}}$$

and also that :

$$\sup_{f \in \mathcal{F}} \left(\mathbb{E}(f(Z_1)) - \frac{1}{n} \sum_{i=1}^n f(Z_i) \right) \leq 2\widehat{R}_n(\mathcal{F}) + 3\sqrt{\frac{\log(2/\delta)}{2n}}$$

Exercise 4 - (Sauer's lemma) Consider \mathcal{A} a class of subsets of \mathbb{R}^d with VC dimension $V < +\infty$ and growth function $\gamma(\mathcal{A}, n), \forall n \geq 1$. Show that :

$$\forall n \geq 1, \quad \gamma(\mathcal{A}, n) \leq \sum_{i=0}^V \binom{n}{i}.$$

Exercice 5 - (VC dimension of half-spaces) Consider the class \mathcal{A} of half-spaces in \mathbb{R}^d and show that its VC dimension $V(\mathcal{A}) = d + 1$.

Hint : first prove the upper bound by Radon's theorem, and then build a separating hyper-plane for any arbitrary labeling for some set of $d + 1$ points.

Exercice 6 - Assume that \mathcal{T} is an m -dimensional vector space of real-valued functions. Compute the VC dimension of the class of sets :

$$\mathcal{A} = \{ \{x \in \mathbb{R}^d : t(x) \geq 0\} : t \in \mathcal{T} \}$$

Application : what is the VC dimension of the class of closed balls in \mathbb{R}^d :

$$\left\{ x = (x_1, \dots, x_d)^T \in \mathbb{R}^d : \sum_{i=1}^d |x_i - a_i|^2 \leq b \right\}$$

where $a_1, \dots, a_d, b \in \mathbb{R}$.

Exercice 7 - Compute the VC dimension $V(\mathcal{A})$ in the following cases :

- (a) $\mathcal{A} = \{] - \infty, x_1] \times \dots \times] - \infty, x_d] : (x_1, \dots, x_d) \in \mathbb{R}^d \}$,
 - (b) \mathcal{A} is the class of all rectangles of \mathbb{R}^2 with axis-orthogonal edges.
 - (c) \mathcal{A} is the class of all rectangles of \mathbb{R}^2 .
 - (d) \mathcal{A} is the class of all triangles of \mathbb{R}^2 .
-