

Introduction to Statistical Learning

Mid-term exam

Duration : 1h30 - No documents allowed

Reminder/Notations

- Law of iterated expectation : $\mathbb{E}(U) = \mathbb{E}(\mathbb{E}(U | V))$ where U, V are random variables.
- McDiarmid inequality : let h be a function of n variables x_1, \dots, x_n satisfying the uniform bounded differences assumption with constant c, \dots, c : for any index i ,

$$\sup_{x_1, \dots, x_n, x'_i} |h(x_1, \dots, x_n) - h(x_1, \dots, x_{i-1}, x'_i, x_{i+1}, \dots, x_n)| \leq c. \quad (1)$$

Then, we have that : for any $t > 0$,

$$\mathbb{P}\{h(X_1, \dots, X_n) - \mathbb{E}(h(X_1, \dots, X_n)) \geq t\} \leq \exp\left(-\frac{2t^2}{nc^2}\right). \quad (2)$$

and

$$\mathbb{P}\{h(X_1, \dots, X_n) - \mathbb{E}(h(X_1, \dots, X_n)) \leq -t\} \leq \exp\left(-\frac{2t^2}{nc^2}\right). \quad (3)$$

- The empirical Rademacher complexity of \mathcal{G} wrt to the sample $Z_1^n = \{Z_1, \dots, Z_n\}$ is defined as :

$$\widehat{R}_n(\mathcal{G}, Z) = \mathbb{E}\left(\sup_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \varepsilon_i g(Z_i) \middle| Z_1^n\right) \quad (4)$$

where $\varepsilon_1, \dots, \varepsilon_n$ are *IID* Rademacher random variables, and they also are independent of Z_1^n .

- The Rademacher complexity of \mathcal{G} is defined as :

$$R_n(\mathcal{G}, Z) = \mathbb{E}(\widehat{R}_n(\mathcal{G}, Z)) \quad (5)$$

- Growth function of a class \mathcal{C} of sets of \mathbb{R}^d of order n :

$$\gamma(\mathcal{C}, n) = \max_{K_n = \{x_1, \dots, x_n\} \subset \mathbb{R}^d} |\{K_n \cap C : C \in \mathcal{C}\}| \quad (6)$$

- VC dimension of a class \mathcal{C} of sets of \mathbb{R}^d :

$$V(\mathcal{C}) = \max \{n \in \mathbb{N} : \gamma(\mathcal{C}, n) = 2^n\}. \quad (7)$$

Exercise 1 - Consider (X, Y) a random pair that models classification data with labels in $\{0, 1\}$. Define the following classification error

$$L_\omega(g) = \mathbb{E}(2\omega(Y) \cdot \mathbb{I}\{Y \neq g(X)\})$$

where $\omega(0) + \omega(1) = 1$.

1. Find the optimal elements (minimizer, error) for this risk.
2. Justify the interest of such an L_ω in practice?
3. Now consider the unit square in \mathbb{R}^2 .
 - (a) Plot the curves defined by $g \mapsto (\mathbb{P}\{g(X) = 1 \mid Y = 0\}, \mathbb{P}\{g(X) = 1 \mid Y = 1\})$ when g varies such that $L_\omega(g) = C$ with C fixed, for different values of C .
 - (b) Same question but assuming now that $\mathbb{P}\{g(X) = 1\} = C$ with C fixed.

Exercise 2 - Let \mathcal{G} be a class of $\{0, 1\}$ -valued functions over \mathbb{R}^d . Let $(X_1, Y_1), \dots, (X_n, Y_n)$ an IID sample of classification data in $\mathbb{R}^d \times \{0, 1\}$. Set $\delta > 0$.

1. Show that, with probability at least $1 - \delta$:

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

2. Set $\mathcal{F} = \{(x, y) \mapsto \mathbb{I}\{y \neq g(x)\} : g \in \mathcal{G}\}$ and relate $R_n(\mathcal{F}, (X, Y))$ to $R_n(\mathcal{G}, X)$.
3. Consider the binary classification problem. Given a class \mathcal{G} of candidate classifiers, what is the strategy that selects a classifier out of \mathcal{G} and for which performance can be explained by a control of the Rademacher average? Provide a mathematical argument for performance prediction of the learning strategy.

Exercise 3 - Consider the two following types of sets of \mathbb{R}^d , with $d \geq 1$:

- $C(\theta, b) = \{x \in \mathbb{R}^d : \theta^T x \leq b\}$
- $S(j, a, b) = \{x = (x^{(1)}, \dots, x^{(d)}) \in \mathbb{R}^d : ax^{(j)} \leq b\}$

where $\theta \in \mathbb{R}^d$, $b \in \mathbb{R}$, $a \in \{-1, +1\}$ and $j \in \{1, \dots, d\}$.

We define the two collections :

- $\Gamma_1 = \{C(\theta, b) : \theta \in \mathbb{R}^d, b \in \mathbb{R}\}$
- $\Gamma_2 = \{S(j, a, b) : a \in \{-1, +1\}, j \in \{1, \dots, d\}, b \in \mathbb{R}\}$

We propose to show that $V(\Gamma_2) < V(\Gamma_1)$ when $d \geq d_0$, for some d_0 :

1. Describe what happens in the case $d = 1$. What does it imply for d_0 ?
2. Prove a tight lower bound on $V(\Gamma_1)$.
3. Given a set K_n of n points $\{x_1, \dots, x_n\}$ in \mathbb{R}^d , what is the maximal number of subsets of K_n obtained as $K_n \cap S$, where $S \in \Gamma_2$.
4. Give an upper bound for $V(\Gamma_2)$ and conclude.