Introduction to Statistical Learning

Mid-term exam

Duration : 1h30 - No documents allowed

Reminder/Notations

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

$$\sup_{x_1,\ldots,x_n,x_i'} |h(x_1, \ldots, x_n) - h(x_1, \ldots, x_{i-1}, x_i', x_{i+1} \ldots, x_n)| \leq c \ . \tag{1}$$

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

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

  and

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

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

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

  where $\varepsilon_1, \ldots, \varepsilon_n$ are *IID* Rademacher random variables, and they also are independent of Z_1^n.
- The Rademacher complexity of $\mathcal{G}$ is defined as :

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

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

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

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

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

Exercise 1 - Consider the model for classification data where X is a random vector on $\mathbb{R}^d$ and Y is a random variable taking values in $\{-1,+1\}$.
Denote by $\eta(x) = \mathbb{P}\{Y = +1 \mid X = x\}$ the posterior probability. Find the optimal classifier under the two following risk scenarios :

1. Consider the classification error $L(g) = \mathbb{P}\{Y \neq g(X)\}$ for $g : \mathbb{R}^d \to \{-1,+1\}$. According to the value of the quantity $\mathbb{E}(Y \mid X = x)$ at $x \in \mathbb{R}^d$, what would be the optimal decision with respect to L ?
2. Fix $u \in (0,1)$. Now assume that we aim at minimizing $L(g)$ under the budget constraint $u = \mathbb{P}(g(X) = +1)$. Set $q \doteq q(u)$ such that $u = \mathbb{P}(\eta(X) > q)$ and express $L(g)$ as the expectation over X of a quantity depending on $\eta(X)$, u, and q. Then deduce what is the optimal classifier with respect to L under the budget constraint.

Exercise 2 - Let $\mathcal{G}$ be a class of $\{0,1\}$-valued functions over $\mathbb{R}^d$. Let $(X_1,Y_1),\ldots,(X_n,Y_n)$ 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 $R_n\big(\mathcal{F},(X,Y)\big)$ 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 $d \geq 1$:
- $C(\theta,b) = \{x \in \mathbb{R}^d \ : \ \theta^T x \leq b\}$
- $S(j,a,b) = \{x = (x^{(1)},\ldots,x^{(d)}) \in \mathbb{R}^d \ : \ ax^{(j)} \leq b\}$

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

We define the two collections :
- $\Gamma_1 = \{C(\theta,b) \ : \ \theta \in \mathbb{R}^d, b \in \mathbb{R}\}$
- $\Gamma_2 = \{S(j,a,b) \ : \ a \in \{-1,+1\}, j \in \{1,\ldots,d\}, b \in \mathbb{R}\}$

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 $\{x_1,\ldots,x_n\}$ 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.
