Introduction to Statistical Learning

Mid-term exam

Duration : 2h - Lecture notes not allowed

Reminder on main definitions and results

- The indicator function $\mathbb{I}\{\Omega\}$ takes the value 1 if Ω is true, and 0 otherwise.
- If A denotes a set, then the notation $|A|$ denotes the cardinality of A.
- Union bound : $\mathbb{P}\{A \cup B\} \leq \mathbb{P}\{A\} + \mathbb{P}\{B\}$ where A and B are events.
- IID means Independent and Identically Distributed.
- Law of iterated expectation : $\mathbb{E}(U) = \mathbb{E}(\mathbb{E}(U \mid V))$ where U, V are random variables.
- Hoeffing's inequality : Consider $Z_1, \ldots, Z_n$ IID over $[0,1]$ and $\overline{Z}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i$. We have, for any $t > 0$
$$\mathbb{P}\{\overline{Z}_n - \mathbb{E}(Z_1) > t\} \leq \exp(-2nt^2)$$
and
$$\mathbb{P}\{\overline{Z}_n - \mathbb{E}(Z_1) < -t\} \leq \exp(-2nt^2)$$
- Subadditivity of supremum operator : $\sup(f+g) \leq \sup(f) + \sup(g)$ and $\sup(f) - \sup(g) \leq \sup(f-g)$.
- 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 (X, Y) a random pair that models classification data with labels in $\{0, 1\}$.

1. For a classifier $g \ : \ \mathbb{R}^d \to \{0, 1\}$, define $L(g) = \mathbb{P}\{Y \neq g(X)\}$. What is the minimizing argument g^* (called the Bayes classifier) of $L(g)$ over all possible classifiers g ? What is the minimal value of $L(g)$ over all possible classifiers g ?
2. Now define $L_c(g) = c_0 \mathbb{P}\{Y \neq g(X), Y = 1\} + c_1 \mathbb{P}\{Y \neq g(X), Y = 0\}$ where $c_0, c_1 > 0$. What is the minimizing argument g_c^* of $L_c(g)$ over all possible classifiers g ? What is the minimal value of $L_c(g)$ over all possible classifiers g ?
3. Using the same notations as in the previous question, we set $c_0 = 1$ and $c_1 = \lambda$ where $\lambda \in [0; +\infty]$. After this reparameterization of the binary classification problem with asymmetric costs, we denote L_c by L_λ and we consider the sequence of binary classification problems $\{\min_g L_\lambda(g) \ : \ \lambda \in [0; +\infty]\}$ with sequence of solutions $\{g_\lambda^* \ : \ \lambda \in [0; +\infty]\}$. Consider the decision rule $h^*(x) = \int_0^\infty g_\lambda^*(x) \, d\lambda$.
   (a) What is the learning problem solved by h^* ?
   (b) Which is the criterion that h^* optimizes ? Give a proof of that fact.
   (c) Is h^* the unique optimal element for the performance criterion ?

Exercice 2 - We 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 $\{-1, +1\}$.

1. We consider the following problems for which the question is to compute the optimal decision rule g^*, h^* or f^*.
   (a) $R(g) = \mathbb{E}\big((Y - g(X))^2\big)$ where $g \ : \ \mathbb{R}^d \to \{-1, +1\}$
   (b) $R(h) = \mathbb{E}\big((Y - h(X))^2\big)$ where $h \ : \ \mathbb{R}^d \to \mathbb{R}$
   (c) $A(f) = \mathbb{E}\big(\log_2(1 + e^{-Yf(X)})\big)$ where $f : \mathbb{R}^d \to \mathbb{R} \cup \{-\infty, +\infty\}$ and $\log_2(u) = \log(u)/\log(2)$ for any $u > 0$.

   Explain why such criteria are relevant for the binary classification problem.
2. We now consider the case of 1(c).
   (a) Determine the function H such that : $A(f^*) = \mathbb{E}(H(\eta(X)))$.
   (b) Plot H and state its main properties. Compare $(u - 1/2)^2$ and $(1 - H(u))$.
   (c) Consider $L(g) = \mathbb{P}\{Y \neq g(X)\}$ and $L^* = \inf_g L(g)$. What upper bound can be given on the quantity $L(\mathrm{sgn}(f)) - L^*$ in terms of $A(f) - A(f^*)$?

Exercise 3 - 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 4 - Consider the two following types of sets of $\mathbb{R}^d$, with $d \geq 1$:

— $C(\theta, b) = \{x \in \mathbb{R}^d \ : \ \theta^T x \leq b\}$

— $S(j, a, b) = \{x = (x^{(1)}, \ldots, x^{(d)}) \in \mathbb{R}^d \ : \ a x^{(j)} \leq b\}$

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.