Introduction to Statistical Learning

Mid-term exam

Duration : 2h - Lecture notes 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 :

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

Exercice 1 - Consider IID random pairs (X, Y) and (X', Y') over $\mathbb{R}^d \times \mathcal{Y}$. Set the following posterior probabilities :

$$\begin{array}{lll} \forall x, x' \in \mathbb{R}^d, & \rho_+(x, x') & = \mathbb{P}\{Y - Y' > 0 \mid X = x, X' = x'\} \\ & \rho_-(x, x') & = \mathbb{P}\{Y - Y' < 0 \mid X = x, X' = x'\} \end{array}$$

and for any preference rule $\pi : \mathbb{R}^d \times \mathbb{R}^d \to \{-1, 1\}$, consider the pairwise error measure

$$L(\pi) = \mathbb{P}\left\{(Y - Y') \cdot \pi(X, X') < 0\right\} \ .$$

1. Find the Bayes rule π^* and the Bayes error $L^* = L(\pi^*)$ for this problem, as well as the excess of risk $L(\pi) - L^*$ for any preference rule π (will involve ρ_+ and ρ_-).
2. Assume $\mathcal{Y} = \{-1, +1\}$ and denote by $\eta(x) = \mathbb{P}\{Y = +1 \mid X = x\}$. Provide the expressions for $\rho_+(x, x')$ and $\rho_-(x, x')$ and discuss how the behavior of η could lead to difficult situations for the learning process to be efficient.
3. Assume now that $\mathcal{Y} = \mathbb{R}$ and that $Y = m(X) + \sigma(X) \cdot N$ where m and σ are P_X-measurable functions, N is a random noise variable with normal distribution $\mathcal{N}(0, 1)$, while N and X are independent random variables. Provide the expressions for $\rho_+(x, x')$ and $\rho_-(x, x')$ in this case and discuss the relation between properties of the model and the learning process.

Exercise 2 - Consider the plain binary classification problem with a *finite* class $\mathcal{G}$ of candidate decision rules (classifiers) from supervised data $(X_1, Y_1), \ldots, (X_n, Y_n)$ where the (X_i, Y_i)'s are IID random pairs over $\mathbb{R}^d \times \{0, 1\}$. We set the empirical classification error of a classifier g to be $\widehat{L}_n(g) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}\{g(X_i) \neq Y_i\}$. We want to provide guarantees over the true classification error $L(g)$ for any $g \in \mathcal{G}$.

1. Give an upper bound on the deviation probability $\mathbb{P}\{\max_{g \in \mathcal{G}} |L(g) - \widehat{L}_n(g)| > t\}$ for any $t > 0$.
2. Deduce an upper bound on $L(g)$ that holds for any $g \in \mathcal{G}$ with a probability at least of $1 - \delta$. The upper bound shall depend on $\widehat{L}_n(g)$, δ, $|\mathcal{G}|$, n.
3. Consider the algorithm that outputs $\widehat{g}_n = \arg\min_{g \in \mathcal{G}} \widehat{L}_n(g)$. Propose a meaningful upper bound of the difference $L(\widehat{g}_n) - L(g^*)$ where g^* is the Bayes classifier. Comment on the choice of the class $\mathcal{G}$.
4. Assume for this question that $\mathcal{G}$ is countable and not a finite class of classifiers. Is it possible to derive a similar bound as in the previous question ? Please provide precise arguments to support your answer.

Exercise 3 - We consider the setup of binary classification where X is a random vector over $\mathbb{R}^d$ and Y is a random variable taking values in $\{-1, +1\}$.
We denote $\eta(x) = \mathbb{P}\{Y = 1 \mid X = x\}$.

1. Express the minimizing function of the criterion $A(f) = \mathbb{E}(\log_2(1 + e^{-Yf(X)}))$ as a function of η in the following cases :
   (i) among functions $f : \mathbb{R}^d \to \{-1, +1\}$
   (ii) among functions $f : \mathbb{R}^d \to \mathbb{R} \cup \{-\infty, +\infty\}$
2. In the case (ii), compute the minimum of $A(f)$.
3. Explain why the minimizers obtained are relevant if the criterion of interest is the classification error ? Please provide precise technical arguments with explicit constants.

Exercise 4 - 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 for fixed g, the empirical Rademacher complexity $\widehat{R}_n(\mathcal{G}, X)$ seen as a function of $X_1, \ldots, X_n$ satisfies the bounded differences condition.
2. Show that, with probability at least $1 - \delta$:

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

3. 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 $R_n(\mathcal{G}, X)$.
4. 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.