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.

— 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.

— 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 $x_1, \dots, x_n$ satisfying the uniform bounded differences assumption with constant $c, \dots, c$: for any index i,

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

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

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

and

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

Exercise 1 - Consider the binary classification model where the random pair (X, Y) has distribution P over $\mathbb{R}_+ \times \{0,1\}$ and :

— the marginal distribution of X over $\mathbb{R}_+$ is denoted P_X

— the conditional distribution of Y given $X = x$ is a Bernoulli distribution with parameter $\eta(x) = \dfrac{x}{x+\theta}$, for any $x \in \mathbb{R}_+$, and for fixed $\theta > 0$.

1. Assume that the marginal distribution follows a uniform distribution $P_X = \mathcal{U}([0, \alpha\theta])$ over $\mathbb{R}_+$ with $\alpha > 1$.
   (a) Find the minimizing argument g^* of $L(g) = \mathbb{P}(Y \neq g(X))$ over all measurable classifiers $g : \mathbb{R}_+ \to \{0,1\}$.
   (b) Compute $L^* = L(g^*)$ in the case where the marginal distribution $P_X = \mathcal{U}([0, \alpha\theta])$ with $\alpha > 1$.
2. Now assume we have the following IID data $(X_1, Y_1), \ldots, (X_n, Y_n)$ available.
   (a) Assuming P_X as before (with α fixed), propose an empirical estimate $\widehat{\theta}$ of θ based only on $X_1, \ldots, X_n$.
   (b) For general P_X over $\mathbb{R}_+$ (unspecified, not necessarily uniformly distributed, and not depending on θ), what is a possible empirical estimate $\widehat{\theta}$ for θ ?
   (c) Denote by $\widehat{\eta}$ the plugin estimate of η based on $\widehat{\theta}$, what is the plugin classifier $\widehat{g}$ based on $\widehat{\eta}$? Find a bound on $L(\widehat{g}) - L^*$ depending on the quantity $\mathbb{E}(|\widehat{\eta}(X) - \eta(X)|)$.

Exercice 2 - Find the optimal elements g^* and $\mathcal{L}^* = \mathcal{L}(g^*)$ in the following cases of error measures with binary classification data :

1. Set $\omega \in (0,1)$, and consider $\mathcal{L}(g) = L_\omega(g)$ such that

$$L_\omega(g) = 2\mathbb{E}\big((1-\omega)\mathbb{I}\{Y = +1\}\mathbb{I}\{g(X) = -1\} + \omega\mathbb{I}\{Y = -1\}\mathbb{I}\{g(X) = +1\}\big) \ .$$

2. Set $u \in (0,1)$, and consider $\mathcal{L}(g) = \mathbb{P}(Y \neq g(X))$ with the constraint

$$\mathbb{P}(g(X) = 1) = u \ .$$

Exercice 3 - 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 $\{-1,+1\}$. We denote by $\eta(x) = \mathbb{P}\{Y = +1 \mid X = x\}$ the posterior probability.

1. Find the optimal decision rule f^* which minimizes the criterion $A(f)$ over the class of measurable functions $f : \mathbb{R}^d \to \mathbb{R} \cup \{-\infty, +\infty\}$ in the following cases :
   (a) Criterion to minimize : $A(f) = \mathbb{E}\exp(-Yf(X)))$.
   (b) Criterion to minimize : $A(f) = \mathbb{E}\big(\log_2(1+\exp(-Yf(X)))\big)$.
2. Consider $L(f) = \mathbb{P}(Y \cdot f(X) < 0)$ for any measurable real-valued function f and $A(f)$ a surrogate criterion such as those introduced in Questions 1.(a) and 1.(b).
   (a) What is the relationship between minimizing $L(f)$ and $A(f)$?
   (b) Given a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, provide a *numerically plausible* procedure to approximate $L(f^*)$. A sketch of proof is expected.

Exercise 4 - Let $\mathcal{F}$ a class of functions from $\mathbb{R}^d$ to $[-B, B]$, with $B > 0$. Consider random sign variables $\varepsilon_1, \ldots, \varepsilon_n$ IID such that $\mathbb{P}\{\varepsilon_1 = -1\} = \mathbb{P}\{\varepsilon_1 = +1\} = 1/2$. Consider the *empirical Rademacher complexity* defined as

$$\widehat{R}_n(\mathcal{F}) = \frac{1}{n}\mathbb{E}\left(\sup_{f\in\mathcal{F}} \sum_{i=1}^n \varepsilon_i f(X_i) \,\middle|\, X_1, \ldots, X_n\right)$$

and the *average Rademacher complexity* as :

$$\bar{R}_n(\mathcal{F}) = \frac{1}{n}\mathbb{E}\left(\sup_{f\in\mathcal{F}} \sum_{i=1}^n \varepsilon_i f(X_i)\right)$$

1. Show that for fixed $\mathcal{F}$, the empirical Rademacher complexity seen as a function of $X_1, \ldots, X_n$ satisfies the bounded differences condition.
2. Provide an upper bound on the average Rademacher complexity in terms of the empirical Rademacher complexity that holds with high probability.
3. How the Rademacher average can be applied to provide theoretical guarantees for the consistency of Empirical Risk Minimization to solve the classification problem ?
