Introduction to Statistical Learning

Final exam

Duration : 2h - Lecture notes not allowed

Reminder on some definitions and results

— IID means Independent and Identically Distributed.

— Jensen's inequality : if ψ is a convex function, then we have $\psi(\mathbb{E}(U)) \leq \mathbb{E}\big(\psi(U)\big)$.

— 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 $t > 0$

$$\mathbb{P}\{\overline{Z}_n - \mathbb{E}(Z_1) > t\} \leq \exp(-2nt^2)$$

— 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 $t > 0$,

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

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

$$\widehat{R}_n(\mathcal{G}, Z) = \mathbb{E}\left(\sup_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i g(Z_i) \,\middle|\, Z_1^n\right) \tag{3}$$

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 :

$$R_n(\mathcal{G}, Z) = \mathbb{E}\big(\widehat{R}_n(\mathcal{G}, Z)\big) \tag{4}$$

— Growth function (or shattering coefficient) of a class $\mathcal{C}$ of sets of $\mathbb{R}^d$ of order n :

$$\gamma(n) = \max_{K_n=\{x_1,\ldots,x_n\}\subset\mathbb{R}^d} |\{K_n \cap C \;:\; C \in \mathcal{C}\}| \tag{5}$$

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

$$V = \max\{n \in \mathbb{N} \;:\; \gamma(n) = 2^n\} \; . \tag{6}$$

— Sauer's lemma : for all $n \geq V$, $\gamma(n) \leq (ne/V)^V$.

Exercice 1 - Consider an IID sample of Rademacher random variables ($\mathbb{P}\{\varepsilon_1 = +1\} = \mathbb{P}\{\varepsilon_1 = -1\} = 1/2$).

1. Consider $Q \subset \mathbb{R}^n$ a finite set of points. We assume that they are all contained in the Euclidean ball with center the origin and radius R. Using Jensen's inequality with the exponential function, show that : for any $t > 0$
$$\mathbb{E}\left(\sup_{q=(q_1,\dots,q_n)\in Q} \sum_{i=1}^{k} \varepsilon_i q_i\right) \leq \frac{tR^2}{2} + \frac{\log |Q|}{t}$$
where $|Q|$ is the number of points in Q.
2. Provide the optimal choice of t in the previous question and give the expression of the optimal bound.
3. Based on the previous inequality, provide a bound for the Rademacher average of a class $\mathcal{G}$ of binary $\{-1, +1\}$-classifiers, first in terms of the growth function of the class, then in terms of the VC dimension of the class.
4. In order to assess the learning complexity for the convex hull of a class $\mathcal{G}$ of classifiers, which notion should be used ? Explain why.

Exercise 2 - In this exercise, we consider a binary classification problem with labels in $\{-1, +1\}$, and the elements f are soft classifiers (real-valued functions over $\mathbb{R}^d$).

1. Set $\varphi(x) = \log_2\big(1 + \exp(x)\big)$ and consider the convex cost function $A(f) = \mathbb{E}\varphi(-Y \cdot f(X))$. Define f^* the optimal element wrt to the functional A and find an explicit function H such that :
$$A(f^*) = \mathbb{E}(H(\eta(X)))$$
2. State some simple properties of H and find $c > 0$ such that : for any $t \in [0, 1]$, we have
$$H(t) \leq 1 - \left(\frac{1-2t}{2c}\right)^2$$
3. We introduce : $L(f) = \mathbb{P}(Y \cdot f(X) < 0)$ and L^* its optimal value. Find α such that the ratio $(L(f) - L^*)/(A(f) - A^*)^\alpha$ is uniformly bounded over all f's.
4. Consider $\lambda > 0$ and $\mathcal{G}$ a family of $\{-1, +1\}$-classifiers with finite VC dimension V. We introduce the following functional class :
$$\mathcal{F}_\lambda = \left\{ f = \sum_{j=1}^{N} w_j g_j \ : \ N \in \mathbb{N},\ g_j \in \mathcal{G},\ w_j \in \mathbb{R},\ \sum_{j=1}^{N} |w_j| \leq \lambda \right\}$$
We set $\widehat{A}_n$ to be the empirical version of A. Show that :
$$\sup_{f\in\mathcal{F}_\lambda} |\widehat{A}_n(f) - A(f)| \leq c_1(\lambda)\sqrt{\frac{V\log(en/V)}{n}} + c_2(\lambda)\sqrt{\frac{\log(1/\delta)}{n}}$$
where c_1 and c_2 will be found explicitly.
5. Consider $\widehat{f}_{n,\lambda}$ the minimizer of $\widehat{A}_n$ over $\mathcal{F}_\lambda$. Provide an explicit upper bound on its classification error $L\big(\widehat{f}_{n,\lambda}\big) - L^*$ which will depend on V, n, and λ, but also on the approximation error wrt to the convex risk : $\inf_{f\in\mathcal{F}_\lambda} A(f) - A^*$.

Exercise 3 - In this exercise, we consider a binary classification problem with labels in $\{-1,+1\}$ over $\mathbb{R}^d$ with IID training data $(X_1, Y_1) \ldots, (X_n, Y_n)$. We define an n-vector of convex weights $\pi(i)$ over the sample points : for any $i = 1, \ldots, n$, $\pi(i) \geq 0$ and $\sum_{i=1}^n \pi(i) = 1$. We introduce the following functionals :

— for any $\{-1,1\}$-classifier g,

$$\epsilon(g) = \sum_{i=1}^n \pi(i)\mathbb{I}\{Y_i \cdot g(X_i) = -1\}$$

— for any real-valued f,

$$\widehat{A}_n(f) = \frac{1}{n}\sum_{i=1}^n \exp\left(-Y_i \cdot f(X_i)\right) \ .$$

1. Provide an expression of $\pi(i)$ such that : for any fixed f, minimizing

$$g \mapsto \left.\frac{\partial A_n(f+\alpha g)}{\partial \alpha}\right|_{\alpha=0}$$

is equivalent to minimizing $\epsilon(g)$.

2. We propose to build decision rules f of the form $f_T = \sum_{t=1}^T \alpha_t g_t$ where the α_t's are real-valued coefficients and g_t's are simple classifiers taking their values in $\{-1,1\}$. Propose an algorithm relying on an iterative principle to determine the updates of (α_t, g_t).
3. Give the explicit expression of α_t at every iteration of the algorithm. *Hint :* We may consider the zero of the function $\alpha \mapsto \frac{\partial A_n(f_{t-1}+\alpha g_t)}{\partial \alpha}$.
4. Provide some practical choices in order to develop a numerical implementation of this algorithm.

Exercise 4 - Let $\mathcal{F}$ be a class of real-valued functions and a fixed value of $\rho > 0$. We assume $(X,Y), (X_1, Y_1) \ldots, (X_n, Y_n)$ are IID binary classification data with labels in $\{-1,+1\}$. Consider the following error functions $L(f) = \mathbb{P}(Y \cdot f(X) < 0)$ and $\widehat{L}_{n,\rho}(f) = \frac{1}{n}\sum_{i=1}^n \psi_\rho(Y_i \cdot f(X_i))$ where : for any $\rho > 0$,

$$\psi_\rho(t) = (1 - t/\rho)\mathbb{I}\{0 \leq t \leq \rho\} + \mathbb{I}\{t \leq 0\}$$

1. For any $\delta > 0$, show that with probability at least $1-\delta$, the two following inequalities hold :

$$\sup_{f\in\mathcal{F}}(L(f) - \widehat{L}_{n,\rho}(f)) \leq \frac{2}{\rho}\mathbb{E}(\widehat{R}_n(\mathcal{F})) + \sqrt{\frac{\log(1/\delta)}{2n}}$$

and

$$\sup_{f\in\mathcal{F}}(L(f) - \widehat{L}_{n,\rho}(f)) \leq \frac{2}{\rho}\widehat{R}_n(\mathcal{F}) + 3\sqrt{\frac{\log(2/\delta)}{2n}}$$

2. From the previous question, derive generalization error guarantees for the Empirical Risk Minimization (ERM) estimator $\widehat{f}_n = \arg\min_{f\in\mathcal{F}} \sum_{i=1}^n \mathbb{I}(Y_i \cdot f(X_i) < 0)$.