Introduction to Statistical Learning

Final exam

Duration : 2h - Lecture notes allowed

Exercise 1 -

Consider $\lambda > 0$ and $\mathcal{G}$ a family of $\{-1, +1\}$-classifiers with finite VC dimension V. We introduce the λ-blown-up convex hull of $\mathcal{G}$ to be defined as :

$$\mathcal{F}_\lambda = \left\{ f = \sum_{j=1}^{N} w_j g_j \ : \ N \in \mathbb{N},\ g_j \in \mathcal{G},\ w_j \in \mathbb{R},\ \sum_{j=1}^{N} |w_j| \leq \lambda \right\}$$

1. Consider $X_1, \ldots, X_n$ an IID sample in $\mathbb{R}^d$ and recall the definition of the Rademacher average :
$$R_n\big(\mathcal{F}_\lambda\big) = \mathbb{E}\left(\sup_{f \in \mathcal{F}_\lambda} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(X_i) \right)$$
where $\varepsilon_1, \ldots, \varepsilon_n$ are *IID* Rademacher random variables, and they also are independent of $X_1, \ldots, X_n$. Provide an upper bound of $R_n\big(\mathcal{F}_\lambda\big)$ that depends on V, n, and λ and give the main arguments of the computation.
2. 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 :
$$A(f^*) = \mathbb{E}(H(\eta(X)))$$
3. State some simple properties of H and find $c > 0$ such that : for any $t \in [0, 1]$, we have
$$H(t) \leq 1 - \left(\frac{1 - 2t}{2c} \right)^2$$
4. 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.

Exercice 2 -

1. Consider a random variable X such that $\mathbb{E}(X) = 0$ and $X \in [a, b]$ almost surely. Give a sketch of proof evoking the main arguments of the following result : for any $t > 0$, we have :
$$\mathbb{E}(e^{tX}) \leq e^{t^2 (b-a)^2 / 8}$$
2. Consider $Q \subset \mathbb{R}^k$ a finite set of points. We assume that they are all contained in the Euclidean ball with center the origin and radius R. Then show that : for any $t > 0$
$$\mathbb{E}\left(\sup_{q=(q_1,\ldots,q_k) \in Q} \sum_{i=1}^{k} \varepsilon_i q_i \right) \leq \frac{tR^2}{2} + \frac{\log |Q|}{t}$$
where $|Q|$ is the number of points in Q.
3. Provide the optimal choice of t in the previous question and give the expression of the optimal bound.

Exercice 2 - Let $Z, Z_1, \ldots, Z_n, , Z'_1, \ldots, Z'_n$ IID random variables with distribution P over $\mathcal{Z}$ and $\mathcal{F}$ be a closed and convex class of functions which is a subset of a Hilbert class $\mathcal{H}$ with norm $\|\cdot\|$. Let $\ell \ : \ \mathcal{F} \times \mathcal{Z} \to \mathbb{RR}$ be a loss function to assess the quality of $f \in \mathcal{F}$ on a sample Z. We denote by : (i) $L(f) = \mathbb{E}\big(\ell(f, Z)\big)$ the expected error of element f on average, (ii) $\hat{L}_n(f) = \frac{1}{n}\sum_{i=1}^n \ell(f, Z_i)$ the empirical risk over the sample $Z_1, \ldots, Z_n$, (iii) $\bar{L}_{\mathcal{F}} = \inf_{f\in\mathcal{F}} L(f)$. We consider a learning algorithm $A \ : \ \mathcal{Z}^n \to \mathcal{F}$ which based on the sample $Z_1, \ldots, Z_n$ outputs a random function $\hat{f}_n = A(Z_1, \ldots, Z_n)$. We want to give an estimate of the excess of risk $L(\hat{f}_n) - \bar{L}_{\mathcal{F}}$ which holds with high probability, where $L(\hat{f}_n) = \mathbb{E}\big(\ell(\hat{f}_n, Z) \mid Z_1, \ldots, Z_n\big)$.

Reminder. Consider a function $\phi \ : \ \mathcal{F} \to \mathbb{R}$.

— A vector $g \in \mathcal{H}$ is a subgradient of ϕ at $f \in \mathcal{F}$ if, for any $f' \in \mathcal{F}$, we have :

$$\phi(f') \geq \phi(f) + < g, f' - f >$$

— The function ϕ is said to be subdifferentiable at f if the set $\partial\phi(f)$ of all subgradients of ϕ at f is not empty.

— The function ϕ is said to be α-strongly convex is ϕ is convex, subdifferentiable and, for any $f, f' \in \mathcal{F}$, and $g \in \partial\phi(f)$, we have :

$$\phi(f') \geq \phi(f) + < g, f' - f > + \frac{\alpha}{2}\|f - f'\|^2 \ .$$

Part A. We consider here the algorithm A such that $A(Z_1, \ldots, Z_n) = \arg\min_{f\in\mathcal{F}} \hat{L}_n(f)$. We assume that $f \mapsto \ell(f, z)$ is α-strongly convex and L-Lipschitz.

1. Prove that $f \mapsto \frac{1}{n}\left(\ell(f, Z'_i) - \ell(f, Z_i)\right)$ is Lipschitz where the Lipschitz constant will be provided.
2. Assume that φ is α-strongly convex and ψ L-Lipschitz over $\mathcal{F}$. Show that there is a unique f^* that minimizes φ and assume that $\tilde{f}$ is a minimizer of $\varphi + \psi$ over $\mathcal{F}$. Show that $\|f^* - \tilde{f}\| \leq \lambda(\alpha, L, n)$ where λ will be computed.
3. Denote by $\hat{f}_n^{(i)}$ the minimizer of the empirical risk over the sample Z_1, …, Z_{i-1}, Z'_i, Z_{i+1}, …, Z_n. Provide a bound of $\left|\ell(\hat{f}_n, z) - \ell(\hat{f}_n^{(i)}, z)\right|$ which holds for any $z \in \mathcal{Z}$.
4. Derive a bound on the quantity $\mathbb{E}\big(L(\hat{f}_n) - \hat{L}_n(\hat{f}_n)\big)$ and then, for $\mathbb{E}\big(L(\hat{f}_n) - \bar{L}_{\mathcal{F}}\big)$.
5. Conclude on a probabilistic bound for $L(\hat{f}_n) - \bar{L}_{\mathcal{F}}$ which holds with probability at least $1 - \delta$.

Part B. We assume here that $\mathcal{F}$ is convex and bounded, i.e. for any $f \in \mathcal{F}$, we have $\|f\| \leq M$ for some $M < \infty$. Fix $\lambda > 0$. We consider here the algorithm A_β such that $\hat{f}_{n,\beta} = A_\beta(Z_1, \ldots, Z_n) = \arg\min_{f\in\mathcal{F}} \left\{\hat{L}_n(f) + \frac{\beta}{2}\|f\|^2\right\}$. We assume here that $f \mapsto \ell(f, z)$ is simply convex and L-Lipschitz.

1. Show that $f \mapsto \ell(f, z) + \frac{\beta}{2}\|f\|^2$ is strongly convex and Lipschitz with constants to be determined. Use part A to derive a bound on $L(\hat{f}_{n,\beta}) - \inf_{f\in\mathcal{F}}\{L(f) + \frac{\beta}{2}\|f\|^2\}$ which holds with probability at least $1 - \delta$.
2. Optimize the bound with respect to β.
   Hint : use $\beta = \frac{\kappa(M,L)}{\sqrt{n}}$ with a properly tuned κ.