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 $X_1, \dots, X_n$ 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 $\varepsilon_1, \dots, \varepsilon_n$ are *IID* Rademacher random variables, and they also are independent of $X_1, \dots, X_n$. 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.

5. 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.

6. 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: $\inf_{f \in \mathcal{F}_\lambda} A(f) - A^*$.

Exercice 2 - Consider the following setup:

- Data: let the triple (X, X', Y) with X, X' being IID random vectors over $\mathbb{R}^d$ and Y a random variable over $\{-1, 0, +1\}$. We assume that there exists a measurable function f^* such that $Y = f^*(X, X')$.
- Error: For any decision rule $h : \mathbb{R}^d \to \mathbb{R}$, we consider the criterion:

$$L(h) = \mathbb{P}\{f^*(X, X') \neq 0,\ f^*(X, X')(h(X') - h(X)) \leq 0\}$$

1. What is the goal of learning here?
2. Under which conditions on f^* do we have a minimal error $L^* = \inf L$ equal to 0?
3. Justify briefly the use of a convex formulation of the risk:

$$A_n(h) = \frac{1}{n} \sum_{i=1}^{n} \exp\big(-Y_i(h(X'_i) - h(X_i))\big)$$

   where the $(X_1, X'_1, Y_1), \ldots, (X_n, X'_n, Y_n)$ are IID.

4. Consider 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 binary $\{0, 1\}$-valued classifier g,

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

   and

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

   Provide an expression of $\pi(i)$ such that: for any fixed h, for $\alpha \in \mathbb{R}$, minimizing

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

   is equivalent to minimizing $\epsilon^-(g) - \epsilon^+(g)$.

5. We propose to build decision rules h of the form $h_T = \sum_{t=1}^{T} \alpha_t g_t$ where the α_t's are real-valued coefficients and g_t's are simple classifiers taking their values in $\{0, 1\}$. Propose an algorithm relying on an iterative principle to determine the updates of (α_t, g_t).
6. Give the explicit expression of α_t at every iteration of the algorithm. *Hint:* We may consider the zero of the function $\alpha \mapsto \frac{\partial A_n(h_{t-1} + \alpha g_t)}{\partial \alpha}$.
7. What are the parameters of the algorithm and how to calibrate them?