Introduction to Statistical Learning

Final exam

Duration : 2h - Lecture notes allowed

Problem

A. Given a sample of IID random vectors $X_1, \dots, X_n$ on $\mathbb{R}^d$ and a class $\mathcal{F}$ of bounded real-valued functions, the empirical Rademacher average is defined as the random quantity :

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

where $\varepsilon_1, \dots, \varepsilon_n$ are IID random sign variables such that $\mathbb{P}\{\varepsilon_1 = -1\} = \mathbb{P}\{\varepsilon_1 = +1\} = 1/2$. Also the ε sample and the X sample are assumed to be independent.

(a) Consider two classes of bounded real-valued functions $\mathcal{F}_1$, $\mathcal{F}_2$. Find a simple upper bound of the following quantity :

$$\frac{1}{n}\mathbb{E}\left(\sup_{f_1\in\mathcal{F}_1, f_2\in\mathcal{F}_2} \sum_{i=1}^{n} \varepsilon_i |f_1(X_i) - f_2(X_i)| \middle| X_1, \dots, X_n \right)$$

depending on $\widehat{R}_n(\mathcal{F}_1)$, $\widehat{R}_n(\mathcal{F}_2)$.

(b) Express $\max\{f_1, f_2\}$ as a linear relation involving $|f_1 - f_2|$.

(c) Consider the class $\mathcal{F} = \{\max\{f_1, f_2\} \ : \ f_1 \in \mathcal{F}_1, f_2 \in \mathcal{F}_2\}$ and provide a simple upper bound of $\widehat{R}_n(\mathcal{F})$ depending on $\widehat{R}_n(\mathcal{F}_1)$, $\widehat{R}_n(\mathcal{F}_2)$.

B. Consider a multiclass classification problem with observations $(X_1, Y_1), \dots, (X_n, Y_n)$ IID copies of the random pair (X, Y) where the output variable Y takes values in $\{1, \dots, K\}$. The decision rules are functions g_h of the form :

$$g_h \ : \ x \mapsto \operatorname*{arg\,max}_{y\in\{1,\dots,K\}} h(x, y)$$

where h is a real-valued function in a class $\mathcal{H}$ of functions over the set $\mathbb{R}^d \times \{1, \dots, K\}$. The complexity of learning in the multiclass classification setup relies on the complexity of the class $\mathcal{H}$ that will be considered here under the margin approach. We thus define the margin ρ_h of function h as :

$$(x, y) \mapsto \rho_h(x, y) = h(x, y) - \max_{y'\neq y} h(x, y') \ ,$$

and ρ_h belongs to the class $\mathcal{H}_\rho$ of functions induced by $\mathcal{H}$.

(a) Set the empirical Rademacher complexity of the class $\mathcal{H}_\rho$ to be :

$$\widehat{R}_n(\mathcal{H}_\rho) = \frac{1}{n}\mathbb{E}\left(\sup_{h\in\mathcal{H}}\sum_{i=1}^{n}\varepsilon_i\rho_h(X_i,Y_i)\middle|(X_1,Y_1),\dots,(X_n,Y_n)\right) .$$

Note that, for any Λ, we have that :

$$\Lambda(X_i,Y_i) = \sum_{y=1}^{K}\Lambda(X_i,y)\mathbb{I}\{y=Y_i\} = \sum_{y=1}^{K}\Lambda(y)\left(\frac{2\mathbb{I}\{y=Y_i\}-1}{2}+\frac{1}{2}\right) ,$$

and show that :

$$\widehat{R}_n(\mathcal{H}_\rho) \leq \frac{1}{n}\sum_{y=1}^{K}\mathbb{E}\left(\sup_{h\in\mathcal{H}}\sum_{i=1}^{n}\varepsilon_i\rho_h(X_i,y)\middle|X_1,\dots,X_n\right)$$

(b) Set $\mathcal{H}_X = \{x\mapsto h(x,y) \ : \ y\in\{1,\dots,K\},\ h\in\mathcal{H}\}$. Using the definition of ρ_h and the main result of Part A, prove that :

$$\widehat{R}_n(\mathcal{H}_\rho) \leq K^\alpha \widehat{R}_n(\mathcal{H}_X)$$

where α will be made explicit.

(c) Set $\varphi_\gamma(u) = (1-u/\gamma)\mathbb{I}\{u\in(0,\gamma]\} + \mathbb{I}\{u\leq 0\}$ and compute its Lipschitz constant.

(d) Relate the multiclass classification error $L(g_h) = \mathbb{P}\{Y\neq g_h(X)\}$ to the multiclass margin error $L_\gamma(h) = \mathbb{E}\{\varphi_\gamma\big(\rho_h(x,y)\big)\}$.

(e) We introduce $\widehat{L}_\gamma(h) = \frac{1}{n}\sum_{i=1}^{n}\varphi_\gamma(\rho_h(X_i,Y_i))$. Use a concentration inequality to derive an upper bound on the quantity :

$$\sup_{h\in\mathcal{H}}(L_\gamma(h)-\widehat{L}_\gamma(h)) .$$

(f) Give a sketch of proof that the following inequality holds, with probability at least $1-\delta$, for any $h\in\mathcal{H}$:

$$L(g_h) \leq \widehat{L}_\gamma(h) + c_1(K,\gamma)\mathbb{E}(\widehat{R}_n(\mathcal{H}_X)) + c_2(n,\delta)$$

where c_1 and c_2 will have to be computed explicitly.

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

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

and the *average Rademacher complexity* as :

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

1. Show that for fixed g, the empirical Rademacher complexity seen as a function of $X_1, \dots, X_n$ 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. Consider a sample $\{X_1, \dots, X_n\}$ of points included in the closed ball $\{x \in \mathbb{R}^d \,:\, \|x\| \leq M\}$ (with respect to Euclidean distance $\|\cdot\|$). Let $\mathcal{G}$ be the class of linear rules defined as $\{x \mapsto w^T x \,:\, w \in \mathbb{R}^d,\ \|w\| \leq L\}$. Derive an upper bound on the empirical Rademacher complexity that involves M, L and n.

Exercise 2 - 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 $\{-1, +1\}$.
Denote by $\eta(x) = \mathbb{P}\{Y = +1 \mid X = x\}$ the posterior probability. Find the optimal classifier under the two following risk scenarios :

1. Consider the classification error $L(g) = \mathbb{P}\{Y \neq g(X)\}$ for $g \,:\, \mathbb{R}^d \to \{-1, +1\}$. According to the value of the quantity $\mathbb{E}(Y \mid X = x)$ at $x \in \mathbb{R}^d$, what would be the optimal decision with respect to L ?
2. Fix $u \in (0, 1)$. Now assume that we aim at minimizing $L(g)$ under the budget constraint $u = \mathbb{P}(g(X) = +1)$. Set $q \doteq q(u)$ such that $u = \mathbb{P}(\eta(X) > q)$ and express $L(g)$ as the expectation over X of a quantity depending on $\eta(X)$, u, and q. Then deduce what is the optimal classifier with respect to L under the budget constraint.
