Introduction to Statistical Learning

Final exam

Duration : 2h - Lecture notes allowed

Exercice 1 - Consider the following :

- $D_n = ((X_1, Y_1), \dots, (X_n, Y_n))$ an IID sample of supervised training data over $\mathcal{X} \times \mathcal{Y}$,
- $\mathcal{F}$ a class of predictors from $\mathcal{X}$ to $\mathcal{Y}$,
- $A : D_n \mapsto \hat{f}_n \in \mathcal{F}$ a learning algorithm,
- $\ell : \mathcal{Y}^2 \to \mathbb{R}_+$ a cost function such that $\ell(y, y') \leq \Lambda$ for any $y, y' \in \mathcal{Y}$, with $\Lambda > 0$,
- $L(\hat{f}) = \mathbb{E}(\ell(Y, \hat{f}(X)) \mid D_n)$ is the risk of any data-driven predictor $\hat{f}$,
- $\hat{L}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i))$ is the empirical risk of any predictor $f \in \mathcal{F}$.

We consider the notation D'_n for a sample of size n which differs from D_n by a single point, and $\hat{f}'_n = A(D'_n)$. We assume that, for any n, there exists a $\beta_n \geq 0$ such that for any samples D_n and D'_n and for any pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we have : $|\ell(y, \hat{f}_n(x)) - \ell(y, \hat{f}'_n(x))| \leq \beta_n$.

1. Find an upper bound on $|L(\hat{f}_n) - L(\hat{f}'_n)|$ depending on β_n.
2. Find an upper bound on $|\hat{L}_n(\hat{f}_n) - \hat{L}_n(\hat{f}'_n)|$ depending on β_n, Λ and n.
3. Show that the quantity $L(\hat{f}_n) - \hat{L}_n(\hat{f}_n)$ satisfies the bounded differences condition and apply a well-known concentration inequality.
4. Then, show that we have, with probability at least $1 - \delta$:

$$L(\hat{f}_n) \leq \hat{L}_n(\hat{f}_n) + \beta_n + (2n\beta_n + \Lambda)\sqrt{\frac{\log(1/\delta)}{2n}}$$

5. What would be an appropriate order of magnitude for the coefficient β_n ? Can you give examples of algorithms that would display such values for β_n ?

Exercise 2 - Consider an IID sample $X_1, \dots, X_n$ of observations over the space $\mathcal{X}$ and $\mathcal{F}_0$ is a set of real-valued functions over $\mathcal{X}$ that includes the zero function. Assume $\psi : \mathbb{R} \to \mathbb{R}$ is k-Lispchitz and define, for fixed positive real numbers V and B :

- the class $\mathcal{F}_0$ is a linear perceptron with bounded weights : $\mathcal{F}_0 = \{x \mapsto w^T x \ : \ \|w\|_1 \leq B\}$
- a one layer network as : $\mathcal{F}_1 = \{x \mapsto \psi(v + \sum_{j=1}^{m} w_j f_j(x)) \ : \ |v| \leq V, \|w\|_1 \leq B, f_j \in \mathcal{F}_0\}$
- a p-layer network as (iterative definition with fixed layer size) : $\mathcal{F}_p = \{x \mapsto \psi(v + \sum_{j=1}^{m} w_j f_j(x)) \ : \ |v| \leq V, \|w\|_1 \leq B, f_j \in \mathcal{F}_{p-1}\}$

Prove the following upper bounds on the empirical Rademacher average :

1. $\hat{R}_n(\mathcal{F}_1) \leq k\left(\frac{V}{\sqrt{n}} + 2B\hat{R}_n(\mathcal{F}_0)\right)$.
2. We assume now that $\mathcal{X}$ is the ℓ_∞ unit ball : $\mathcal{X} = \{x \in \mathbb{R}^d \ : \ \|x\|_\infty \leq 1\}$ and show that :
$$\hat{R}_n(\mathcal{F}_0) \leq \frac{B\sqrt{2\ln(2d)}}{\sqrt{n}}$$
3. Assume in addition that $\psi(-u) = -\psi(u)$ and $k = 1$ then show that on $\mathcal{X} = \{x \in \mathbb{R}^d,\ \|x\|_\infty \leq 1\}$:
$$\hat{R}_n(\mathcal{F}_p) \leq \frac{1}{\sqrt{n}}\left(B^{p+1}\sqrt{2\ln(2d)} + V\sum_{l=0}^{p-1} B^l\right) .$$

Problem

A. Given a sample of IID random vectors $X_1, \ldots, X_n$ on $\mathbb{R}^d$ and a class $\mathcal{F}$ of bounded real-valued functions, the empirical Rademacher average is defined as the random quantity :
$$\widehat{R}_n(\mathcal{F}) = \frac{1}{n}\mathbb{E}\left(\sup_{f\in\mathcal{F}}\sum_{i=1}^n \varepsilon_i f(X_i)\middle| X_1, \ldots, X_n\right)$$
where $\varepsilon_1, \ldots, \varepsilon_n$ 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 :
$$\frac{1}{n}\mathbb{E}\left(\sup_{f_1\in\mathcal{F}_1, f_2\in\mathcal{F}_2}\sum_{i=1}^n \varepsilon_i|f_1(X_i) - f_2(X_i)|\middle| X_1, \ldots, X_n\right)$$
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 $(X_1, Y_1), \ldots, (X_n, Y_n)$ IID copies of the random pair (X, Y) where the output variable Y takes values in $\{1, \ldots, K\}$. The decision rules are functions g_h of the form :
$$g_h \ : \ x \mapsto \operatorname*{arg\,max}_{y\in\{1,\ldots,K\}} h(x, y)$$
where h is a real-valued function in a class $\mathcal{H}$ of functions over the set $\mathbb{R}^d \times \{1, \ldots, K\}$. 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 :

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

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.
