Introduction to Statistical Learning

Final exam (3 pages)

Duration : 2h00 - Lecture notes allowed

Notations

— **Indicator function**. The indicator function $\mathbb{I}\{\Omega\}$ takes the value 1 if Ω is true, and 0 otherwise.

— **IID**. Independent and Identically Distributed.

— **Empirical Rademacher average**. Consider an IID sample $Z_1^n = (Z_1, \ldots, Z_n)$ and let $\sigma_1, \ldots, \sigma_n$ be an IID sample of Rademacher random variables ($\mathbb{P}\{\sigma_1 = +1\} = \mathbb{P}\{\sigma_1 = -1\} = 1/2$) independent of Z_1^n. Given a class $\mathcal{T}$ of functions, we denote its empirical Rademacher average by :

$$\hat{R}_n(\mathcal{T}) = \mathbb{E}\left(\sup_{t\in\mathcal{T}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i t(Z_i) \mid Z_1^n\right)$$

— **Kernel function - definitions and properties.** Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a positive definite and symmetric kernel function. We recall that k has the property that there exist : (i) a Hilbert space $\mathcal{H}_k$ equipped with scalar product $< \cdot, \cdot >_k$ and norm $\|\cdot\|_k$ and (ii) a feature mapping $\Phi : \mathbb{R}^d \to \mathcal{H}_k$ such that $k(x, x') =< \Phi(x), \Phi(x') >$ and $k(x, x) = \|\Phi(x)\|_k$ for any x, x'. Given a sample $X_1, \ldots, X_n$, we denote by $K = \big(k(X_i, X_j)\big)_{1\leq i,j\leq n}$ the Gram matrix induced by the kernel function k.

— **Subdifferentiability, strong convexity.** Let $\mathcal{F}$ be a closed and convex class of functions which is a subset of a Hilbert class $\mathcal{H}$ with scalar product $< \cdot, \cdot >$ and norm $\|\cdot\|$. 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* if ϕ is convex, subdifferentiable and, for any $f, f' \in \mathcal{F}$, and $g \in \partial\phi(f)$, we have :

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

Exercise 1 - Consider an IID sample $X_1, \dots, X_n$ of observations over the space $\mathcal{X}$ and $\mathcal{F}_0$ 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 $\mathcal{X} = \{x \in \mathbb{R}^d, \ \|x\|_\infty \leq 1\}$:
$$\hat{R}_n(\mathcal{F}_p) \leq \frac{1}{\sqrt{n}} \left(B^{p+1}\sqrt{2\ln(2d)} + V \sum_{l=0}^{p-1} B^l \right) .$$

Exercise 2 - Consider the setup of preference learning where we observe an IID sample of triples $(X_1, X_1', Y_1), \dots, (X_n, X_n', Y_n)$. The probabilistic model assumes that, for each i, the triple (X_i, X_i', Y_i) is such that X_i, X_i' are IID random vectors over $\mathbb{R}^d$ and Y_i is a random variable over $\{-1, 0, +1\}$. Consider the margin loss function as $\varphi_\rho(u) = (1 - (u/\rho))\mathbb{I}\{0 < u \leq \rho\} + \mathbb{I}\{u \leq 0\}$ for any real number u. We define the ranking error of a preference rule $g : \mathbb{R}^d \to \{-1, 0, +1\}$ as :

$$L^R(g) = \mathbb{P}\{Y \neq 0, \ Y \cdot (g(X') - g(X)) \leq 0\}$$

and the empirical margin ranking error as :

$$\widehat{L}^R_{n,\rho}(g) = \frac{1}{n} \sum_{i=1}^n \varphi_\rho(Y_i \cdot (g(X_i') - g(X_i)) \ ,$$

Now consider a class $\mathcal{G}$ of preference rules and define :

$$\widetilde{\mathcal{G}} = \{(x, x', y) \mapsto y(g(x') - g(x)) \ : \ g \in \mathcal{G}\} \ .$$

1. Provide an upper bound of the empirical Rademacher average of $\widetilde{\mathcal{G}}$ in terms of the empirical Rademacher average of $\mathcal{G}$.
2. Which inequality relates the empirical Rademacher average of the loss class $\varphi_\rho \circ \widetilde{\mathcal{G}}$ to the empirical Rademacher average of $\widetilde{\mathcal{G}}$? Provide a proof of this inequality.

3. Show that, for any $\delta \in (0,1)$, we have, with probability at least $1-\delta$: for any $g \in \mathcal{G}$

$$\mathbb{E}(\varphi_\rho(Y \cdot (g(X') - g(X))) \leq \widehat{L}^R_{n,\rho}(g) + c_1 \hat{R}_n(\varphi_\rho \circ \tilde{\mathcal{G}}) + c_2(n,\delta)$$

   for some c_1 and $c_2(n,\delta)$ that will have to be given explicitly.
4. Deduce from the previous question a margin error bound for $L^R(g)$ that holds with large probability for any $g \in \mathcal{G}$ and which involves the empirical ranking error of g over the sample and the complexity of $\mathcal{G}$.
5. Specify the previous result to the case of a kernel class of functions with $\mathcal{G} = \mathcal{F}_{k,M} = \{x \mapsto < w, \Phi(x) >_k : \ w \in \mathcal{H}, \|w\|_k \leq M\}$ as defined in the notations for a kernel function k.

Problem - Let $Z, Z_1, \ldots, Z_n, Z'_1, \ldots, Z'_n$ be 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 $\ell : \mathcal{F} \times \mathcal{Z} \to \mathbb{R}$ 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 $L(\hat{f}_n) = \mathbb{E}\big(\ell(\hat{f}_n, Z) \mid Z_1,\ldots,Z_n\big)$.

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.