Introduction to Statistical Learning

Exercise sheet n°4

Definitions :

— Define the margin loss as follows : for any $\rho > 0$,

$$\varphi_\rho(t) = \mathbb{I}\{t \leq 0\} + \mathbb{I}\{0 \leq t \leq \rho\} \left(1 - \frac{t}{\rho}\right) .$$

and the related cost function as : $\ell(y, f) = \varphi_\rho(y \cdot f)$.

— The empirical margin φ-risk is given by :

$$\widehat{A}_{\rho,n}(f) = \frac{1}{n} \sum_{i=1}^{n} \varphi_\rho(Y_i \cdot f(X_i))$$

and the margin φ-risk is denoted by $A_\rho(f) = \mathbb{E}\big(\varphi_\rho(Y \cdot f(X))\big)$.

— The empirical margin error is given by :

$$\widehat{L}_{\rho,n}(f) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(Y_i \cdot f(X_i) \leq \rho)$$

and the margin φ-risk is denoted by $L_\rho(f) = \mathbb{P}\big(Y \cdot f(X) \leq \rho\big)$.

Exercise 1 - [Generic Margin Analysis] Prove the following statements :

1. Let $\mathcal{F}$ be a class of real-valued functions. Fix $\rho > 0$, then for any $\delta > 0$, with probability at least $1 - \delta$, show that, for any $f \in \mathcal{F}$, we have :

$$L\big(\mathrm{sgn}(f)\big) \leq \widehat{L}_{\rho,n}(f) + \frac{2}{\rho}\overline{R}_n\big(\mathcal{F}\big) + \sqrt{\frac{\log(1/\delta)}{2n}}$$

where $\overline{R}_n\big(\mathcal{F}\big)$ is the expected Rademacher average of the class $\mathcal{F}$, and

$$L\big(\mathrm{sgn}(f)\big) \leq \widehat{L}_{\rho,n}(f) + \frac{2}{\rho}\widehat{R}_n\big(\mathcal{F}\big) + 3\sqrt{\frac{\log(2/\delta)}{2n}}$$

where $\widehat{R}_n\big(\mathcal{F}\big)$ is the empirical Rademacher average of the class $\mathcal{F}$.

2. Let $\mathcal{F}$ be a class of real-valued functions with values in $[-c, c]$. Fix $\delta > 0$, with probability at least $1 - \delta$, show that, for any $f \in \mathcal{F}$ and any $\rho > 0$, we have :

$$L\big(\mathrm{sgn}(f)\big) \leq \widehat{L}_{\rho,n}(f) + \frac{4}{\rho}\overline{R}_n\big(\mathcal{F}\big) + \sqrt{\frac{\log\log_2(2/\rho)}{n}} + \sqrt{\frac{\log(2/\delta)}{2n}}$$

where $\overline{R}_n\big(\mathcal{F}\big)$ is the expected Rademacher average of the class $\mathcal{F}$, and :

$$L\big(\mathrm{sgn}(f)\big) \leq \widehat{L}_{\rho,n}(f) + \frac{4}{\rho}\widehat{R}_n\big(\mathcal{F}\big) + \sqrt{\frac{\log\log_2(2c/\rho)}{n}} + 3\sqrt{\frac{\log(4/\delta)}{2n}}$$

where $\widehat{R}_n\big(\mathcal{F}\big)$ is the empirical Rademacher average of the class $\mathcal{F}$.
Hint : use a sequence $\rho_k = c2^{-k}$ and apply McDiarmid's concentration bound with threshold value $t_k = t + \sqrt{\frac{\log(k)}{n}}$.

3. Application to a kernel class - assume k is a uniformly bounded kernel symmetric and positive : $\sup_x k(x,x) = B^2$ and consider, for $M > 0$ that $B_k(M)$ is the RKHS ball for the metric induced by k with radius M. Prove that, for any $\delta > 0$, with probability at least $1-\delta$, we have, for any $f \in B_k(M)$ and any any $\rho > 0$, we have :

$$L\big(\mathrm{sgn}(f)\big) \leq \widehat{L}_{\rho,n}(f) + \frac{4}{\rho}\frac{BM}{\sqrt{n}} + +\sqrt{\frac{\log\log_2(2BM/\rho)}{n}} + \sqrt{\frac{\log(4/\delta)}{2n}}$$

Exercise 2 - [SVM consistency] Consider a kernel k, uniformly bounded by B^2 with associated RKHS $\mathcal{F}$ and norm $\|\cdot\|_{\mathcal{F}}$, and assume that $\inf_{f\in\mathcal{F}} A(f) = \inf_f A(f)$, where $A(f) = \mathbb{E}\big(\max\{0, 1 - Y \cdot f(X)\}\big)$, holds.

Consider also the following estimator for fixed $\lambda > 0$:

$$\widehat{f}_n^{\lambda} = \arg\min_{f\in\mathcal{F}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \max\{0, 1 - Y_i \cdot f(X_i)\} + \lambda\|f\|_{\mathcal{F}}^2 \right\}$$

1. Show that

$$A(\widehat{f}_n^{\lambda}) \leq \widehat{A}_n(\widehat{f}_n^{\lambda}) + \frac{2B}{\sqrt{n\lambda}} + (1 + B/\sqrt{\lambda})\sqrt{\frac{\log(2/\delta)}{2n}}$$

2. Now set $\lambda = \lambda_n$ such that $\lambda_n \to 0$ and $n\lambda_n \to \infty$ when $n \to \infty$ and prove that : for any $\epsilon > 0$, we have :

$$\sum_{n\geq 0} \mathbb{P}\left(A(\widehat{f}_n^{\lambda_n}) - \inf_f A(f) \geq \epsilon \right) < \infty$$

Deduce that $A(\widehat{f}_n^{\lambda_n})$ tends to $\inf_f A(f)$ almost surely. What can be said about $L(\mathrm{sgn}(\widehat{f}_n^{\lambda_n}))$?

Exercise 3 - [φ-risk analysis of boosting] 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.

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 4 - [Margin analysis for preference learning] Consider the setup of preference learning where we observe an IID sample of triples $(X_1, X_1', Y_1), \ldots, (X_n, X_n', Y_n)$. 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 $\{-1, 0, +1\}$. 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 $g \in \mathcal{G}$

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

   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 $\mathcal{G}$.