Introduction to Statistical Learning

Exercise sheet n°3

Exercise 1 - (Rademacher average computations) Let $\sigma_1, \dots, \sigma_n$ an i.i.d. sample of Rademacher random variables ($\mathbb{P}\{\sigma_1 = +1\} = \mathbb{P}\{\sigma_1 = -1\} = 1/2$). Given a class $\mathcal{T}$ of functions, we denote its empirical Rademacher average by :

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

1. Consider $\mathcal{G}$ a class of functions with values in $\{-1, +1\}$ and its empirical Rademacher average $\hat{R}_n(\mathcal{G})$, and let $\mathcal{L}$ the class of classification loss functions :

   $$\mathcal{L} = \{(x, y) \mapsto \mathbb{I}\{g(x) \neq y\} \ : \ g \in \mathcal{G}\} \ .$$

   Prove that : $\hat{R}_n(\mathcal{L}) = \frac{1}{2}\hat{R}_n(\mathcal{G})$.
2. Let $k \ : \ \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ a positive definite and symmetric kernel function with feature mapping Φ, and, given a sample $X_1, \dots, X_n$, its Gram matrix $K = \big(k(X_,, X_j)\big)_{1\leq i,j\leq n}$. Consider the class of functions $\mathcal{H} = \{x \mapsto < w, \Phi(x) > : \ \|w\|_k \leq M\}$, and prove that :

   $$\hat{R}_n(\mathcal{H}) \leq \frac{M\sqrt{\mathrm{Tr}(K)}}{n} \ .$$
3. Consider $\mathcal{F}$ a class of real-valued functions and $\mathrm{conv}(\mathcal{F})$ its convex hull. Prove that : $\hat{R}_n\big(\mathrm{conv}(\mathcal{F})\big) = \hat{R}_n(\mathcal{F})$.

Exercise 2 - (Margin bounds) Consider a class $\mathcal{F}$ of real-valued functions with Rademacher average $R_n(\mathcal{F})$. We introduce the following definitions :

— Classification error : $L(f) = \mathbb{P}\{Yf(X) < 0\}$.

— Margin loss :

$$m_\rho(t) = \begin{cases} 0 & \text{if } \ \rho \leq t \\ 1 - \dfrac{t}{\rho} & \text{if } \ 0 \leq t \leq \rho \\ 1 & \text{if } \ t \leq 0 \end{cases}$$

— Empirical margin error :

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

The goal of the exercise is to derive margin-dependent upper bounds on the generalization error.

1. Fix $\rho \in (0,1)$, and $\delta > 0$, then show that, with probability at least $1-\delta$, we have, for any f :
$$L(f) \leq \widehat{L}_{n,\rho}(f) + \frac{2}{\rho} R_n(\mathcal{F}) + 3\sqrt{\frac{\log(1/\delta)}{2n}} \ .$$
2. We introduce two sequences $(\rho_k)_{k\geq 1}$, $(\epsilon_k)_{k\geq 1}$ in $(0,1)$. For any $k \geq 1$, give a simple bound on :
$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \big(L(f) - \widehat{L}_{n,\rho_k}(f) \big) > \frac{2}{\rho_k} R_n(\mathcal{F}) + \epsilon_k \right)$$
3. Set $\epsilon_k = \epsilon + \sqrt{\dfrac{\log k}{n}}$ and $\rho_k = 1/2^k$ to derive a simple upper bound on :
$$\mathbb{P}\left(\forall f, \forall \rho \ : \ L(f) - \widehat{L}_{n,\rho}(f) > \frac{4}{\rho} R_n(\mathcal{F}) + \sqrt{\frac{\log \log_2(2/\rho)}{n}} + \epsilon \right)$$

Exercise 3 - (AdaBoost as Coordinate Descent) Consider $\mathcal{G}$ a family of $\{-1,+1\}$-classifiers and the class of finite linear combinations of elements of $\mathcal{G}$:
$$\mathcal{F}_N = \left\{ f = \sum_{t=1}^{N} w_t g_t \ : \ w_1, \dots, w_N \in \mathbb{R}, \ g_1, \dots, g_N \in \mathcal{G} \right\} \ .$$
We define the boosting distribution over the sample through the recurrence :
$$\pi_{t+1}(i) = \pi_t(i) \cdot \frac{\exp(-w_t Y_i g_t(X_i))}{Z_t}$$
with Z_t the normalization factor and $\pi_1(i) = 1/n$.

1. We set $\epsilon_t = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} \pi_t(i) \mathbb{I}\{Y_i \neq g_t(X_i)\}$. Compute Z_t as a function of ϵ_t.
2. We set $\mathbf{w} = (w_1, \dots, w_T)^T$ and we introduce the function : for $\mathbf{w} \in \mathbb{R}^T$,
$$\widehat{A}_n(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} \exp\left(-Y_i \sum_{t=1}^{T} w_t g_t(X_i) \right) \ .$$
Now set $\mathbf{e_t}$ be the unit vector on the t-th coordinate and $\mathbf{w}_{t-1} = (w_1, \dots, w_{t-1}, 0, \dots, 0)^T$, and compute
$$\frac{d\widehat{A}_n(\mathbf{w}_{t-1} + \eta \mathbf{e_t})}{d\eta}$$
3. Consider $\eta \to 0$: what is the optimal direction for minimizing $\widehat{A}_n$ coordinate-wise ? Select the optimal η and highlight the connection to the AdaBoost algorithm.
4. Describe a variant of AdaBoost based on the minimization of
$$\widehat{A}_n(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + \exp\left(-2Y_i \sum_{t=1}^{T} w_t g_t(X_i) \right) \right) \ .$$

Exercise 4 - The setup is binary classification. Let $\mathcal{G}$ be a symmetric class of classifiers with finite VC dimension V. We introduce the class $\mathcal{F}$ of functions f that can be expressed as follows :

$$f(x) = \sum_{i=1}^{N} \alpha_i g_i(x) \ , \quad \forall x \in \mathbb{R}^d$$

where $N \geq 1$, the coefficients α_i are positive weights such that $\sum_i \alpha_i = 1$, and the g_i's are elements of $\mathcal{G}$. We denote $g_f = \text{sign}(f)$ the corresponding classifier.

The *empirical margin error* is defined, for any $\gamma > 0$, by :

$$\widehat{L}_n^{\gamma}(g_f) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{Y_i f(X_i) < \gamma\} \ .$$

1. Compare the empirical margin error and empirical classification error ? What is the interpretation of the margin ?
2. Show that there exists a function that is continuous and piecewise linear and bounded between 0 and 1, denoted by ϕ_γ and such that :

$$\sup_{f \in \mathcal{F}} \left(L(g_f) - \widehat{L}_n^{\gamma}(g_f) \right) \leq \sup_{f \in \mathcal{F}} \left(\mathbb{E}(\phi_\gamma(-Yf(X))) - \frac{1}{n} \sum_{i=1}^{n} \phi_\gamma(-Y_i f(X_i)) \right) \ .$$

3. Use McDiarmid's inequality to upper bound the previous quantity with its expectation.
4. Let $\sigma_1, \ldots, \sigma_n$ an i.i.d. sample of Rademacher random variables ($\mathbb{P}\{\sigma_1 = +1\} = \mathbb{P}\{\sigma_1 = -1\} = 1/2$). Prove that :

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left(\mathbb{E}(\phi_\gamma(-Yf(X))) - \frac{1}{n} \sum_{i=1}^{n} \phi_\gamma(-Y_i f(X_i)) \right) \leq 2\mathbb{E} \sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} \sigma_i (\phi_\gamma(-Y_i f(X_i)) - \phi_\gamma(0)) \right) \ .$$

5. We introduce the contraction principle (proved) : let ψ a Lipschitz function (Lipschitz constant equal to 1) such that $\psi(0) = 0$, then, we have :

$$\mathbb{E} \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \sigma_i \psi(-Y_i f(X_i)) \leq \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \sigma_i (-Y_i f(X_i)) \ .$$

   Infer an upper bound of

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} \sigma_i (\phi_\gamma(-Y_i f(X_i)) - \phi_\gamma(0)) \right) \ .$$

6. Prove the following inequality :

$$\mathbb{E} \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \sigma_i f(X_i) \leq \mathbb{E} \sup_{g \in \mathcal{G}} \sum_{i=1}^{n} \sigma_i g(X_i)$$

   where $\mathcal{F}$ is the convex hull of $\mathcal{G}$ as defined at the beginning of the text.
7. Derive an upper bound on $\mathbb{E} \sup_{g \in \mathcal{G}} \sum_{i=1}^{n} \sigma_i g(X_i)$ which depends only on V, n.

8. Provide an upper bound with high probability for the quantity :

$$\sup_{f\in\mathcal{F}}\Big(L(g_f)-\widehat{L}_n^{\gamma}(g_f)\Big)\ ,$$

that depends only on γ, V, n and the confidence parameter δ.

Exercise 5 - (Convex risk bounds) Consider $\mathcal{G}$ a family of $\{-1,+1\}$-classifiers with finite VC dimension V. We introduce the λ-blown-up convex hull of $\mathcal{G}$ as :

$$\mathcal{F}_\lambda=\left\{f=\sum_{j=1}^N w_j g_j\ :\ N\in\mathbb{N},\ \sum_{j=1}^N|w_j|\le\lambda,\ g_1,\dots,g_N\in\mathcal{G}\right\}$$

1. Show an upper bound of the empirical Rademacher average $\widehat{R}_n\big(\mathcal{F}_\lambda\big)$ of the family $\mathcal{F}_\lambda$ that depends on V, n, and λ.
2. Consider the convex cost function $A(f)=\mathbb{E}\varphi(-Y\cdot f(X))$ with $\varphi(x)=\exp(x)$. Define f^* the optimal element wrt to 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

$$H(t)\le 1-\left(\frac{1-2t}{2c}\right)^2$$

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 :

$$\sup_{f\in\mathcal{F}_\lambda}|\widehat{A}_n(f)-A(f)|\le c_1(\lambda)\sqrt{\frac{V\log(en/V)}{n}}+c_2(\lambda)\sqrt{\frac{\log(1/\delta)}{n}}$$

where c_1 and c_2 will be found explicitly.

6. We recall that for any positive real numbers u, v, we have : $\sqrt{u+v}\le\sqrt{u}+\sqrt{v}$. 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 : $\inf_{f\in\mathcal{F}_\lambda}A(f)-A^*$.
7. Propose a relevant scheme for the choice of the regularization parameter λ.
