Introduction to Statistical Learning

Exercise set # 4

Exercise 1 - (Rademacher average for neural networks) Consider an i.i.d. sample $X_1, \ldots, X_n$ 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 $\mathcal{X} = \{x \in \mathbb{R}^d, \ \|x\|_\infty \leq 1\}$:
$$\hat{R}_n(\mathcal{F}_p) \leq \frac{1}{\sqrt{n}} \left(B^{p+1}\sqrt{2\ln(2d)} + V \sum_{l=0}^{p-1} B^l \right) \ .$$

Exercise 2 - [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, with probability at least $1 - \delta$:
$$\sup_{f \in \mathcal{F}_\lambda} |\widehat{A}_n(f) - A(f)| \leq c_1(\lambda) \sqrt{\frac{V \log(en/V)}{n}} + c_2(\lambda) \sqrt{\frac{\log(1/\delta)}{n}}$$
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 3 - [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 $\mathcal{G}$.