Introduction to Statistical Learning

Exercise set #3

Exercise 1 - (Properties of Rademacher averages) Let $\mathcal{T}$, $\mathcal{T}_1$, $\mathcal{T}_2$, be classes of real-valued functions. Prove the following properties :

1. If $c \in \mathbb{R}$, then $\hat{R}_n(c\mathcal{T}) = |c|\hat{R}_n(\mathcal{T})$.
2. If $\mathcal{T}_1 \subseteq \mathcal{T}_2$, then $\hat{R}_n(\mathcal{T}_1) \leq \hat{R}_n(\mathcal{T}_2)$
3. $\hat{R}_n(\mathcal{T}_1 + \mathcal{T}_2) = \hat{R}_n(\mathcal{T}_1) + \hat{R}_n(\mathcal{T}_2)$
4. Let $\text{conv}(\mathcal{T})$ be the convex hull of $\mathcal{T}$. Prove that : $\hat{R}_n\big(\text{conv}(\mathcal{T})\big) = \hat{R}_n(\mathcal{T})$.
5. If $\psi : \mathbb{R} \to \mathbb{R}$ is k-Lipschitz, then $\hat{R}_n(\psi \circ \mathcal{T}) \leq k\hat{R}_n(\mathcal{T})$.

Exercise 2 - (Rademacher average for linear and kernel classes)

1. Consider $\mathcal{B}_\infty(C) = \{x \in \mathbb{R}^d \mid \|x\|_\infty \leq C\}$ and $\mathcal{G} = \{x \in \mathcal{B}_\infty(C) \mapsto w^T x : \|w\|_1 \leq B\}$. Show that the following bound holds :

$$\hat{R}_n(\mathcal{G}) \leq \frac{BC\sqrt{2\ln(2d)}}{\sqrt{n}} .$$

2. Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ a positive definite and symmetric kernel function with feature mapping Φ that is : for any (x, x'), we have $k(x, x') = <\Phi(x), \Phi(x')>$ where $<,>$ is the product of some Hilbert space. Given a sample $X_1, \ldots, X_n$, define its Gram matrix as $K = \big(k(X_,, X_j)\big)_{1 \leq i,j \leq n}$. Consider the class of functions $\mathcal{H} = \{x \mapsto < w, \Phi(x) > : \|w\|_k \leq M\}$, and prove that :

$$\hat{R}_n(\mathcal{H}) \leq \frac{M\sqrt{\text{trace}(K)}}{n} .$$

Exercise 3 - (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 $\lambda > 0$:

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

1. Show that $\|\widehat{f}_n^\lambda\|^2 \leq 1/\lambda$.

2. Show that, with probability at least $1-\delta$:

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

3. 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 $L(\mathrm{sgn}(\widehat{f}_n^{\lambda_n}))$?

4. Present an alternative proof of SVM consistency based on the property of stability (first solve Exercise 4).

Exercice 4 - (Error bounds for stable algorithms)

Consider the following :

— $D_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$ 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 $\hat{f}$,
— $\hat{L}_n(f) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \ell(Y_i, f(X_i))$ 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 ?