

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 : $\hat{R}_n(\text{conv}(\mathcal{T})) = \hat{R}_n(\mathcal{T})$.
 5. If $\psi : \mathbb{R} \rightarrow \mathbb{R}$ 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 $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ a positive definite and symmetric kernel function with feature mapping Φ that is : for any (x, x') , we have $k(x, x') = \langle \Phi(x), \Phi(x') \rangle$ where \langle, \rangle is the product of the Hilbert space induced by k . Given a sample X_1, \dots, X_n , define its Gram matrix as $K = (k(X_i, X_j))_{1 \leq i, j \leq n}$. Consider the class of functions $\mathcal{H} = \{x \mapsto \langle w, \Phi(x) \rangle : \|w\|_k \leq M\}$, and prove that :

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

Exercise 3 - (Relation between Rademacher average and combinatorial complexities) Consider a class \mathcal{G} of binary valued functions with shattering coefficient function denoted by $\gamma(\mathcal{G}, n)$. Show that :

$$\mathbb{E}(\hat{R}_n(\mathcal{G})) \leq \sqrt{\frac{2\ln(\gamma(\mathcal{G}, n))}{n}}$$

and if $V(\mathcal{G}) < +\infty$ then show that, for some constant C , we have :

$$\mathbb{E}(\hat{R}_n(\mathcal{G})) \leq C\sqrt{\frac{V(\mathcal{G})\log(n)}{n}}$$

Exercise 4 - (Contraction principle)

1. Consider $\psi : \mathbb{R} \rightarrow \mathbb{R}$ a Lipschitz function with constant κ .

Show that, for any class \mathcal{F} of real-valued functions, we have :

$$\widehat{R}_n(\psi \circ \mathcal{F}) \leq \kappa \widehat{R}_n(\mathcal{F})$$

Hint : Use a recurrence argument and the definition of the supremum applied to the following functions of $f : u_{j-1}(f) + (\psi \circ f)(X_j)$ and $u_{j-1}(f) - (\psi \circ f)(X_j)$.

2. Consider the following function : for any $\rho > 0$,

$$\psi_\rho(t) = (1 - t/\rho)\mathbb{I}\{0 \leq t \leq \rho\} + \mathbb{I}\{t \leq 0\}$$

Let \mathcal{F} be a class of real-valued functions and a fixed value of ρ . We assume $(X, Y), (X_1, Y_1) \dots, (X_n, Y_n)$ are IID binary classification data with labels in $\{-1, +1\}$. Denote by $L(f) = \mathbb{P}(Y \cdot f(X) < 0)$ and $\widehat{L}_{n,\rho}(f) = \frac{1}{n} \sum_{i=1}^n \psi_\rho(Y_i \cdot f(X_i))$. For any $\delta > 0$, show that with probability at least $1 - \delta$, the two following inequalities hold :

$$\sup_{f \in \mathcal{F}} (L(f) - \widehat{L}_{n,\rho}(f)) \leq \frac{2}{\rho} \mathbb{E}(\widehat{R}_n(\mathcal{F})) + \sqrt{\frac{\log(1/\delta)}{2n}}$$

and

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

Exercise 5 - (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 $A(f) = \mathbb{E}(\max\{0, 1 - Y \cdot f(X)\})$, 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 $\|\widehat{f}_n^\lambda\|^2 \leq 1/\lambda$.
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 $\lambda_n \rightarrow 0$ and $n\lambda_n \rightarrow \infty$ when $n \rightarrow \infty$ 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 $L(\text{sgn}(\widehat{f}_n^{\lambda_n}))$?

4. Present an alternative proof of SVM consistency based on the property of stability.