CS7545, Spring 2023: Machine Learning Theory - Homework #2

Jacob Abernethy, Tyler LaBonte, and Yeojoon Youn Due: Tuesday, February 21 at 11:59 p.m.

Homework Policy: Working in groups is fine, but *every student* must submit their own writeup, *i.e.,* write the solutions on their own and not submit a shared document. Please write the members of your group on your solutions. There is no strict limit to the size of the group but we may find it a bit suspicious if there are more than 4 to a team. Questions labelled with **(Challenge)** are not strictly required, but you'll get some participation credit if you have something interesting to add, even if it's only a partial answer.

1) **Properties of Rademacher complexity.** Suppose $A \subseteq \mathbb{R}^m$.

(a) Prove that $\mathfrak{R}(A+b) = \mathfrak{R}(A)$ where $A + b = \{a + b : a \in A\}$ for any $b \in \mathbb{R}^m$.

(b) Prove that $\mathfrak{R}(cA) = |c|\mathfrak{R}(A)$ where $cA = \{c \cdot a : a \in A\}$ for any $c \in \mathbb{R}$.

(c) In lecture we proved the following one-sided generalization bound: for $\mathcal{F}$ containing functions $f : \mathcal{X} \to [0, 1]$ and any $\delta > 0$, with probability at least $1 - \delta$ over the choice of a sample $S \sim \mathcal{D}^m$, the following holds for all $f \in \mathcal{F}$:

$$L(f) \leq \widehat{L}_S(f) + 2\mathfrak{R}(\mathcal{F}) + \sqrt{\frac{\log 1/\delta}{m}}. \tag{1}$$

However, to show a bound on the estimation error of ERM we actually needed a two-sided bound, on $\sup_{f\in\mathcal{F}} \left|L(f) - \widehat{L}_S(f)\right|$. Use parts **(a)** and **(b)** to prove one.

2) **Rademacher complexity of ℓ_1-bounded neural networks.** Suppose the input space is $\mathcal{X} = \mathbb{R}^n$ and we have a training set $S = \{(x_i, y_i)\}_{i=1}^m$. For an L-Lipschitz activation function ϕ, define the class of neural networks of depth $2 \leq j \leq D$ and width H with ℓ_1-bounded weights recursively as

$$\mathcal{F}_j := \left\{ x \mapsto \sum_{k=1}^{H} w_k \phi(f_k(x)) : f_k \in \mathcal{F}_{j-1}, \|w\|_1 \leq B_j \right\} \tag{2}$$

For example, ϕ could be the sigmoid or ReLU nonlinearities, which are 1-Lipschitz.

(a) Define $\mathcal{F}_1 := \{x \mapsto \langle w, x\rangle : \|w\|_1 \leq B_1\}$ and suppose $\|x_i\|_\infty \leq C$ for all $1 \leq i \leq m$. Prove that

$$\widehat{\mathfrak{R}}_S(\mathcal{F}_1) \leq B_1 C \sqrt{\frac{2\log 2n}{m}}. \tag{3}$$

Hint. Use Hölder's Inequality and Massart's lemma.

(b) Prove that $\widehat{\mathfrak{R}}_S(\mathcal{F}_j) \leq 2LB_j\widehat{\mathfrak{R}}_S(\mathcal{F}_{j-1})$ for $2 \leq j \leq D$. **Hint.** Use Hölder's Inequality and Talagrand's contraction lemma.

(c) Use parts **(a)** and **(b)** to show an upper bound on the Rademacher complexity of $\widehat{\mathfrak{R}}_S(\mathcal{F}_D)$.

(d) If we have time near the end of the semester, we will prove that the VC-dimension of depth D neural networks with piecewise linear activations and W weights is $\mathcal{O}(WD \log W)$. What are the advantages and disadvantages of your Rademacher complexity bound compared to this VC-dimension bound?

3) **Massart's Lemma, Take 2.** If you recall when we proved Massart's Lemma, it looked suspiciously similar to the proof of Hoeffding's Inequality – indeed, you can reduce it directly. We will show a slightly worse result; see the textbook for a more sophisticated application which achieves the tight bound.

Let $A \subseteq [-1, 1]^m$ be a finite set. Prove the following via a reduction to Hoeffding's Inequality:

$$\mathfrak{R}(A) = O\left(\sqrt{\frac{\log m + \log |A|}{m}}\right). \tag{4}$$

Hint. For any real random variable Z and any real t, we have $\mathbb{E}[Z] = \mathbb{E}[Z \cdot \mathbb{1}[Z \leq t]] + \mathbb{E}[Z \cdot \mathbb{1}[Z > t]]$.

4) **Growth function.** In lecture we studied the growth function for classes of functions taking values in the set $\{-1, 1\}$, but the same definition applies to classes of functions taking values in the finite set $\mathcal{Y}$. In this case, $\Pi_{\mathcal{H}}(m) \leq |\mathcal{Y}|^m$ (analogous to 2^m in the original setup).

(a) Let $\mathcal{H}_1 \subseteq \{h : \mathcal{X} \to \mathcal{Y}_1\}$ and $\mathcal{H}_2 \subseteq \{h : \mathcal{X} \to \mathcal{Y}_2\}$ be function classes and let $\mathcal{H}_3 \subseteq \{h : \mathcal{X} \times \mathcal{X} \to \mathcal{Y}_1 \times \mathcal{Y}_2\}$ such that $\mathcal{H}_3 = \{(h_1, h_2) : h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2\}$. Show that

$$\Pi_{\mathcal{H}_3}(m) \leq \Pi_{\mathcal{H}_1}(m) \cdot \Pi_{\mathcal{H}_2}(m). \tag{5}$$

(b) Let $\mathcal{H}_1 \subseteq \{h : \mathcal{X} \to \mathcal{Y}_1\}$ and $\mathcal{H}_2 \subseteq \{h : \mathcal{Y}_1 \to \mathcal{Y}_2\}$ be function classes and let $\mathcal{H}_3 \subseteq \{h : \mathcal{X} \to \mathcal{Y}_2\}$ such that $\mathcal{H}_3 = \{h_2 \circ h_1 : h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2\}$. Show that

$$\Pi_{\mathcal{H}_3}(m) \leq \Pi_{\mathcal{H}_1}(m) \cdot \Pi_{\mathcal{H}_2}(m). \tag{6}$$

5) **VC-dimension.**

(a) What is the VC-dimension of a union of k intervals on the real line?

(b) What is the VC-dimension of axis-aligned hyperrectangles in $\mathbb{R}^n$?

(c) A simplex in $\mathbb{R}^n$ is the intersection of $n+1$ halfspaces (not necessarily bounded). Prove that the VC-dimension of simplices in $\mathbb{R}^n$ is $\mathcal{O}(n^2 \log n)$. **Hint.** Use the VC-dimension of halfspaces in $\mathbb{R}^n$.

(d) Prove the best lower bound you can on the VC-dimension of simplices in $\mathbb{R}^n$ **(Challenge).**

6) **Desymmetrization (Challenge).** Let $S = \{x_1, \ldots, x_m\} \sim \mathcal{D}^m$ and suppose $\mathcal{F}$ contains functions $f : \mathcal{X} \to [0, 1]$. Prove the symmetrization lower bound, also called the desymmetrization inequality:

$$\frac{1}{2}\mathfrak{R}(\mathcal{F}) - \sqrt{\frac{\log 2}{2m}} \leq \mathop{\mathbb{E}}_{S}\left[\sup_{f \in \mathcal{F}} \left|L(f) - \widehat{L}_S(f)\right|\right]. \tag{7}$$