CS 7545: Machine Learning Theory Due March 1, 2023, 11:59pm

Problem Set 2

Instructors: Jake Abernethy, Tyler LaBonte

Problem 1

Problem.

1. **Graded.** Is it possible for an ERM hypothesis $\hat{h} \in \mathcal{H}$ on a set $S \subseteq \mathcal{X}$ to have $\widehat{L}_S(\hat{h}) = 0$ but $L(\hat{h}) = 1$? Why or why not? Would this be overfitting or underfitting? What is the role of the complexity of $\mathcal{H}$ in this situation?

2. **Graded.** Let $\mathcal{H}$ be a hypothesis class, $\hat{h} \in \mathcal{H}$ be an ERM hypothesis for a sample S, and $h^\star = \arg\inf_{h \in \mathcal{H}} L(h)$. Show that

$$\mathbb{E}_{S \sim \mathcal{D}^m}[\widehat{L}_S(\hat{h})] \leq L(h^\star) \leq \mathbb{E}_{S \sim \mathcal{D}^m}[L(\hat{h})]. \tag{1}$$

3. **Ungraded, optional.** Here are some resources if you would like to study the proof of the no-free-lunch theorem. Most sources prove a simpler version; the one from class takes a bit more work. I highly recommend this illustrated proof. For a textbook version, see Section 5.1 "The No-Free-Lunch Theorem" in Understanding Machine Learning: From Theory to Algorithms. (Don't worry about the PAC-learning stuff in Corollary 5.2).

Problem 2

Problem. In lecture we showed the following one-sided uniform convergence generalization bound: for $\mathcal{H}$ containing functions $h : \mathcal{X} \to \{-1, 1\}$ such that $|\mathcal{H}| < \infty$ and any $\delta > 0$, with probability at least $1 - \delta$ over $S \sim \mathcal{D}^m$, the following holds for all $h \in \mathcal{H}$:

$$L(h) \leq \widehat{L}_S(h) + \sqrt{\frac{\log|\mathcal{H}| + \log 1/\delta}{2m}}. \tag{2}$$

This bound shows that the estimation error of the empirical risk minimizer goes to zero with $1/\sqrt{m}$, which is often called the "slow rate". Interestingly, we did not use any properties of $\mathcal{H}$ besides its size. One may wonder whether there is any advantage to choosing a hypothesis class $\mathcal{H}'$ which is the same size as $\mathcal{H}$ but contains "better" functions. In fact, it turns out that if all the functions in $\mathcal{H}'$ have sufficiently low generalization error, we can achieve a "fast rate" of $1/m$.

In order to prove the fast rate bound, we need a more sophisticated concentration bound than Hoeffding's inequality. Let us state *Bernstein's inequality*: Let $Z_1, \ldots, Z_m$ be *i.i.d.* random variables with zero mean such that $|Z_i| \leq C$ and $\mathrm{Var}(Z_i) \leq D$ for all i. Then for all $\epsilon > 0$,

$$\Pr\left[\frac{1}{m}\sum_{i=1}^{m} Z_i \geq \epsilon\right] \leq \exp\left(\frac{-(m\epsilon^2)/2}{D + (C\epsilon)/3}\right). \tag{3}$$

1. **Graded.** Let $\mathcal{H}$ contain functions $h : \mathcal{X} \to \{-1, 1\}$ with $|\mathcal{H}| < \infty$. Suppose there exists a function $q : \mathbb{R} \to \mathbb{R}$ such that $L(h) \leq q(m)$ for any $h \in \mathcal{H}$ and $S \subseteq \mathcal{X}$ with $|S| = m$. Use Bernstein's inequality to prove that for any $\delta > 0$, with probability at least $1 - \delta$ over $S \sim \mathcal{D}^m$, the following holds for all $h \in \mathcal{H}$:

$$L(h) \leq \widehat{L}_S(h) + \sqrt{\frac{2(\log|\mathcal{H}| + \log 1/\delta)q(m)}{m}} + \frac{2(\log|\mathcal{H}| + \log 1/\delta)}{3m}. \tag{4}$$

   Hint. First, define the Z_i's and compute C and D. Then, the rest will be similar to the proof of our original bound, but with Bernstein instead of Hoeffding. It's OK if you get different constants than are listed here.

2. **Graded.** How should $q(m)$ scale in order to obtain the fast rate? It is sufficient to give an answer like $q(m) = \mathcal{O}(?)$ and explain your reasoning.

3. **Ungraded, optional.** A more general form of Bernstein's inequality holds for a very large class of distributions called *subexponential* distributions. These distributions are roughly characterized by having heavier tails than a Gaussian – they decay with e^{-x} instead of e^{-x^2} – and they come up often in machine learning theory. If you would like to learn more, read sections 2.7 and 2.8 of High-Dimensional Probability.

Problem 3

Problem. In this problem, we will prove a classical bound on the Rademacher complexity of neural networks. Suppose the input space is $\mathcal{X} = \mathbb{R}^n$ and we have a training set $S = \{(x_i, y_i)\}_{i=1}^m$. Let $\phi : \mathbb{R} \to \mathbb{R}$ be an L-Lipschitz activation function such that $\phi(0) = 0$ (*e.g.*, the ReLU function). Define the class of neural networks of depth $2 \leq j \leq D$ and width H with ℓ_1-bounded weights recursively as

$$\mathcal{F}_j \coloneqq \left\{ x \mapsto \sum_{k=1}^{H} w_k \phi(f_k(x)) : f_k \in \mathcal{F}_{j-1}, \|w\|_1 \leq B_j \right\}. \tag{5}$$

Here, ϕ is applied elementwise, *i.e.*, $\phi(x) = (\phi(x_1), \ldots, \phi(x_n))$.

1. **Graded.** Define $\mathcal{F}_1 \coloneqq \{x \mapsto \langle w, x \rangle : \|w\|_1 \leq B_1\}$ and suppose $\|x_i\|_\infty \leq C$ for all $1 \leq i \leq m$. Prove that
$$\mathfrak{R}_S(\mathcal{F}_1) \leq B_1 C \sqrt{\frac{2 \log 2n}{m}}. \tag{6}$$
Hint. Use Hölder's inequality and Massart's lemma.

2. **Graded.** Prove that $\mathfrak{R}_S(\mathcal{F}_j) \leq 2LB_j \mathfrak{R}_S(\mathcal{F}_{j-1})$ for $2 \leq j \leq D$. **Hint.** Use Hölder's inequality and Talagrand's contraction lemma. You may use part **(4)** without proof.

3. **Graded.** Use parts **(1)** and **(2)** to show an upper bound on the Rademacher complexity of $\mathfrak{R}_S(\mathcal{F}_D)$. (You must use parts **(1)** and **(2)**).

4. **Ungraded, optional.** Prove that if a function class $\mathcal{G}$ contains the zero function, then
$$\mathbb{E}_\sigma \left[\sup_{g \in \mathcal{G}} \frac{1}{m} \left| \sum_{i=1}^{m} \sigma_i g(x_i) \right| \right] \leq 2\mathfrak{R}_S(\mathcal{G}). \tag{7}$$

Problem 4

Problem. Suppose $A \subseteq \mathbb{R}^m$.

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

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

3. **Graded.** In lecture we proved the following one-sided uniform convergence generalization bound: for $\mathcal{F}$ containing functions $f : \mathcal{X} \to [0, 1]$ and any $\delta > 0$, with probability at least $1 - \delta$ over $S \sim \mathcal{D}^m$, the following holds for all $f \in \mathcal{F}$:
$$L(f) \leq \widehat{L}_S(f) + 2\mathfrak{R}(\mathcal{F}) + \sqrt{\frac{\log 1/\delta}{2m}}. \tag{8}$$
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 **(1)** and **(2)** to prove one. (You must use parts **(1)** and **(2)**).

4. **Challenge, optional, 1 point extra credit.** Let $S \sim \mathcal{D}^m$ and suppose $\mathcal{F}$ contains functions $f : \mathcal{X} \to [0,1]$. Prove the symmetrization lower bound, also called the desymmetrization inequality:
$$\frac{1}{2}\mathfrak{R}(\mathcal{F}) - \sqrt{\frac{\log 2}{2m}} \leq \mathbb{E}_S \left[\sup_{f \in \mathcal{F}} \left| L(f) - \widehat{L}_S(f) \right| \right]. \tag{9}$$

Problem 5

Problem. 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).

1. **Graded.** 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
$$\Pi_{\mathcal{H}_3}(m) = \Pi_{\mathcal{H}_1}(m) \cdot \Pi_{\mathcal{H}_2}(m). \tag{10}$$

2. **Graded.** 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
$$\Pi_{\mathcal{H}_3}(m) \leq \Pi_{\mathcal{H}_1}(m) \cdot \Pi_{\mathcal{H}_2}(m). \tag{11}$$

3. **Ungraded, optional.** Prove that **(2)** is tight, *i.e.,* exhibit $\mathcal{X}, \mathcal{Y}_1, \mathcal{Y}_2, \mathcal{H}_1, \mathcal{H}_2, m$ such that $\Pi_{\mathcal{H}_3}(m) = \Pi_{\mathcal{H}_1}(m) \cdot \Pi_{\mathcal{H}_2}(m)$. **Hint.** You can take $|\mathcal{X}| = m = 1$.

Problem 6

Problem.

1. **Graded.** What is the VC-dimension of a union of k intervals on the real line?

2. **Graded.** What is the VC-dimension of axis-aligned hyperrectangles in $\mathbb{R}^n$?

3. **Graded.** 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 $\mathbb{R}^n$.

4. **Challenge, optional, 1 extra credit point.** Prove the best lower bound you can on the VC-dimension of simplices in $\mathbb{R}^n$. You will receive the extra credit point if you either (i) prove a lower bound of $\Omega(n)$ and show a reasonable attempt at improving it, or (ii) prove a lower bound better than $\Omega(n)$.