Homework 8
APPM 4490/5490 Spr 2022 Theoretical ML

Due date: Friday, Mar 18, in class or on Canvas/Gradescope by midnight **Instructor**: Prof. Becker
Theme: Stability

Instructions Collaboration with your fellow students is OK and in fact recommended, although direct copying is not allowed. The internet is allowed for basic tasks (e.g., looking up definitions on wikipedia) but it is not permissible to search for proofs or to *post* requests for help on forums such as `http://math.stackexchange.com/` or to look at solution manuals. Please write down the names of the students that you worked with.

An arbitrary subset of these questions will be graded.

Reading You are responsible for reading chapter 13 about "regularization and stability" of Understanding Machine Learning by Shai Shalev-Shwartz and Shai Ben-David (2014, Cambridge University Press).

Problem 1: Exercise 13.1 in Shalev-Shwartz and Ben-David, going from a bound on the "expected risk" to a bound (with high probability) on the true risk and hence the usual agnostic PAC learning. Let $\mathtt{A}$ be an algorithm that guarantees that $m \geq m_{\mathcal{H}}^{\text{expected}}(\epsilon)$ then $\forall \mathcal{D}$ it holds

$$\mathop{\mathbb{E}}_{S \overset{\text{iid}}{\sim} \mathcal{D}^m} [L_{\mathcal{D}}(\mathtt{A}(S))] \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon.$$

a) Show that $\forall \delta \in (0,1)$, if $m \geq m_{\mathcal{H}}^{\text{expected}}(\epsilon \cdot \delta)$ then with probability at least $1-\delta$ it holds that

$$L_{\mathcal{D}}(\mathtt{A}(S)) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon.$$

Hint: use Markov's inequality

b) For every $\delta \in (0,1)$, let

$$m_{\mathcal{H}}^{\text{agnostic}}(\epsilon, \delta) = m_{\mathcal{H}}^{\text{expected}}(\epsilon/2) \cdot \lceil \log_2(1/\delta) \rceil + \left\lceil \frac{\ln(4/\delta) + \ln(\lceil \log_2(1/\delta) \rceil)}{\epsilon^2} \right\rceil \tag{1}$$

and describe a corresponding procedure that agnostic PAC learns the problem $(\mathcal{H}, \mathcal{Z}, \ell)$ with sample complexity $m_{\mathcal{H}}^{\text{agnostic}}$, assuming the loss ℓ has output within $[0,1]$.

Note: The book suggests Eq. (1), but I think they have typos (they accounted for some but not all factors of 2). The expression I got was:

$$m_{\mathcal{H}}^{\text{agnostic}}(2\epsilon, 2\delta) = m_{\mathcal{H}}^{\text{expected}}(\epsilon/2) \cdot \lceil \log_2(1/\delta) \rceil + \left\lceil 2\frac{\ln(2/\delta) + \ln(\lceil \log_2(1/\delta) \rceil)}{\epsilon^2} \right\rceil \tag{2}$$

You can prove either equation (1) or (2).

Hint: Similar to exercise 10.1, let $k = \lceil \log_2(1/\delta) \rceil$ and divide the data into $k+1$ chunks, the first k chunks of size $m_{\mathcal{H}}^{\text{expected}}(\epsilon/2)$, and train each of these first k chunks $S^{(j)}$ (separately) via the algorithm $\mathtt{A}$. On the basis of part (a), reason that

$$\mathbb{P}\left(\bigwedge_{j=1}^{k} \left(L_{\mathcal{D}}(\mathtt{A}(S^{(j)})) > \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon \right) \right) < 2^{-k}$$

and $2^{-k} \leq \delta$ because of how we defined k. Finally, use the last chunk as a validation set.

c) **(No work required, just an observation)** From part (a), we can immediately deduce $m_{\mathcal{H}}^{\text{agnostic}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{\text{expected}}(\epsilon \cdot \delta)$. So why bother with part (b)? Because the bound is much better. For simplicity, let's assume $m_{\mathcal{H}}^{\text{expected}}(\epsilon) = 1/\epsilon^2$, which is a fairly common complexity (and we'll ignore constants and such for now). Then using part (a) by itself, we have $m_{\mathcal{H}}^{\text{agnostic}}(\epsilon, \delta) \leq \frac{1}{\epsilon^2 \delta^2}$, which grows large quite quickly as $\epsilon, \delta \to 0$.

But using part (b) (and ignoring negligible terms and constants), we have instead $m_{\mathcal{H}}^{\text{agnostic}}(\epsilon, \delta) \leq \frac{\log(1/\delta)}{\epsilon^2}$, a difference of $\log(1/\delta)$ vs $1/\delta^2$ (so for base-10 log and $\delta = 0.01$, this is a value of 2 vs a value of 10,000). That's a huge gain!

Problem 2: Let $\mathcal{X} = \mathbb{R}^d$ for some $d \in \mathbb{N}$, $\mathcal{Y} = \{0, 1\}$ and $\mathcal{H} = \{h_r \mid r \geq 0\}$ where

$$h_r(\mathbf{x}) = \begin{cases} 1 & \|\mathbf{x}\| \leq r \\ 0 & \text{else} \end{cases}$$

is the class of indicator functions of balls centered at 0 (and $\|\cdot\|$ is an arbitrary norm).

a) Prove $\mathcal{H}$ is PAC-learnable with sample complexity at most $m_{\mathcal{H}}(\epsilon, \delta) \leq \lceil \epsilon^{-1} \log(1/\delta) \rceil$ (so assume realizability). *(Note: this is a better sample complexity than you get via the generic bound in the Fundamental Theorem of Statistical Learning)*

This was the bonus problem on the midterm. *Hint: Try a technique like we did in problem 2 from HW 1*

b) Compute the exact VC dimension of $\mathcal{H}$ (not a bound).

c) Compute a bound for the growth function $\tau_{\mathcal{H}}(m)$ in terms of the value for $d = \text{VCdim}(\mathcal{H})$ that you just computed.

d) Compute an exact expression for the growth function $\tau_{\mathcal{H}}(m)$ (not a bound).

Problem 3: Let $\boldsymbol{\sigma} = (\sigma_i)_{i=1}^m$ be a Rademacher random variable (iid entries, $\mathbb{P}[\sigma_i = 1] = \mathbb{P}(\sigma_i = -1] = \frac{1}{2}$). Give a non-trivial upper bound on $\mathbb{E}_{\boldsymbol{\sigma}}[\max_{j \in [m+1]} \sum_{i=j}^m \sigma_i]$ where we define $\sum_{i=m+1}^m \sigma_i = 0$. (i.e., m is a trivial bound, so give something sharper than this). *Hint: use your results from Problem 2*