Homework 8
APPM 7400 Spr 2020 Theoretical ML

Due date: Friday, April 3, before 1 PM **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.

Note: please upload a PDF (typed or scanned) directly to Canvas, since there are no more in-person meetings due to coronavirus

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 $m \ge m_{\mathcal{H}}^{\text{expected}}(\epsilon)$ then $\forall \mathcal{D}$ it holds

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

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

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

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:

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

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