Homework 7
APPM 4490/5490 Theory of Machine Learning, Spring 2024

Due date: Friday, Mar 8 '24, before 11 AM, via paper or via Gradescope

Instructor: Prof. Becker

Revision date: 3/1/2024

Theme: Boosting

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 10 about "boosting" of Understanding Machine Learning by Shai Shalev-Shwartz and Shai Ben-David (2014, Cambridge University Press), as well as chapter 7 in Foundations of Machine Learning, 2nd edition, by Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar.

Problem 1: Show that the VC dimension of decision stumps in d dimensions is bounded $\text{VCdim}(\mathcal{H}) \leq c_1 + c_2 \log_2(d)$ for some constants c_1 and c_2. (Any value of constants work, but you might try to get $c_1 = 6 - 4\log_2(\ln(2)) \approx 8.12$ and $c_2 = 2$, or better). Note: Exercise 9.5 in Mohri et al. shows more generally that binary Decision Trees of k nodes in d dimensions have $\text{VCdim}(\mathcal{H}) = \Omega(k \log(d))$.

Hint: After picking a dimension, how many dichotomies can a stump achieve? It may help to think of the ERM algorithm we derived for decision stumps; also, decision stumps have nice graphical interpretations, especially for $d = 2$. Use this to bound the size of the growth function $\tau(m)$, and then do the usual trick: argue that if $\text{VCdim} \geq m$, then we must have $\tau(m) \geq 2^m$, and use Lemma A.2.

Problem 2: Bounding the VC dimension of AdaBoost. Let $\text{VCdim}(\mathcal{H}) = d$, and consider boosting the base class $\mathcal{H}$ by T rounds of AdaBoost (this $\mathcal{H}$ is what we called $\mathcal{B}$ in lecture). Then Shalev-Shwartz and Ben-David shows that the boosted hypothesis class has VC dimension bounded by (Lemma 10.3)

$$\text{VCdim}(\mathcal{H}_{\text{boosted}}) \leq T(d+1)(3\log(T(d+1)) + 2) \quad \text{if} \quad T \geq 3, d \geq 3$$

or via eq (7.9) in Mohri et al.:

$$\text{VCdim}(\mathcal{H}_{\text{boosted}}) \leq 2(d+1)(T+1)\log_2((T+1)e).$$

We can make tighter bounds by not doing approximations. The proof of Lemma 10.3 bounds the growth function

$$\tau(m) \leq \left(\frac{em}{d}\right)^{dT} \left(\frac{em}{T}\right)^{T} \tag{1}$$

and then gives some simplifying lower bounds to this, and then argues that if this is to be at least as big as 2^m (so that VCdim is m), it leads to the bounds we have on d. As an alternative, we can *numerically* solve $\tau(m) = 2^m$ for m, and then rely on monotonicity when m is sufficiently large (no need to prove that for this exercise) that this is the VC dimension. Let d be the VC dimension of decision stumps in $\mathbb{R}^{10}$ (or actually the integer bound on it we derived in problem 1), and letting $T = 3$, compute the m that solves $\tau(m) = 2^m$ (or better, take the logarithm

of both side before solving), and use this to get a more exact bound to the VC dimension for 10-dimensional decision trees that are boosted 3 times. (We don't actually solve $\tau(m) = 2^m$ since we don't know $\tau(m)$ exactly, but use the bound in Eq. (1) as a proxy for $\tau(m)$). Use numerical rootfinding, such as `fzero` in Matlab or `scipy.optimize.root_scalar` in Python.

Finally, compare with the generic bounds from the textbooks that are stated above.

Problem 3:

a) Using the VC dimension of 10-dimensional decision stumps boosted 3 times (as calculated in the previous problem), estimate how many iid samples $m = |S|$ are needed to ensure that $L_{\mathcal{D}}(h) \leq \widehat{L}_S(h) + \epsilon$ with probability at least $1 - \delta$, for $\epsilon = \delta = 0.05$? [If you don't trust your VC dimension calculation, you can use one of the textbook bounds]

Hint: Using results from Mohri et al. gives the best bound, but unlike Shalev-Shwartz and Ben-David, you do not have an explicit formula for $m(\epsilon, \delta) \ldots$ but that's OK, because you can solve for it numerically using root finding again!

b) How large of a validation set S_V would we need so that $L_{\mathcal{D}}(h) \leq \widehat{L}_{S_V}(h) + \epsilon$ with probability at least $1 - \delta$, for the same values of ϵ and δ as in the previous part of the problem? [Note: this holds regardless of whether we used an ERM or boosting or any other algorithm]. Compare this number of samples to the number of samples from part (a).