Homework 5
APPM 4490/5490 Theory of Machine Learning, Spring 2024

Due date: Friday, Feb 23 '24, before 11 AM, via paper or via Gradescope

Instructor: Prof. Becker
Revision date: 2/16/2024

Theme: VC dimension

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 6 of "Understanding Machine Learning" by Shai Shalev-Shwartz and Shai Ben-David (2014, Cambridge University Press). Note: you can buy the book for about $45 on Amazon (or less for an e-book), and the authors host a free PDF copy on their website (but note that this PDF has different page numbers).

Problem 1: Exercise 6.11a in Shalev-Shwartz and Ben-David: let $\mathcal{H}_1, \ldots, \mathcal{H}_r$ be hypothesis classes over the same domain $\mathcal{X}$ and let $\mathrm{VCdim}(\mathcal{H}_i) \le d$ for all $i \in [r]$. Assume $d \ge 3$. Prove that

$$\mathrm{VCdim}\left(\mathcal{H} \stackrel{\text{def}}{=} \bigcup_{i=1}^{r} \mathcal{H}_i\right) \le 4d\log(2d) + 2\log(r).$$

Note: the hint in the book is helpful but slightly confusing and is a bit imprecise on whether inequalities are strict. Recall $\mathrm{VCdim}(\mathcal{H})$ is the largest value of m such that the growth function has the trivial bound, $\tau_{\mathcal{H}}(m) = 2^m$. Hence, bounding the growth function can bound the VC dimension. So, prove $\tau_{\mathcal{H}}(m) < rm^d$ (hint: use Sauer's lemma), and thus if $\mathrm{VCdim}(\mathcal{H}) \ge m$ means $2^m < rm^d$, and then use Lemma A.2 (hint: look at Lemma A.1). An alternative to this last step is to take $\tau_{\mathcal{H}}(m) \le rm^d$ and then prove a variant of Lemma A.2 that has strict inequalities.

Note: Sauer's lemma in Shalev-Shwartz and Ben-David requires $m > d + 1$ where $d = \mathrm{VCdim}(\mathcal{H})$, but as shown in Mohri et al. this can be relaxed to $m \ge d$ (for example, observe for $d = m$, Sauer's lemma is that $\tau_{\mathcal{H}}(m) \le e^m$ which is always true since $\tau_{\mathcal{H}}(m) \le 2^m$ and $2 < e$).

Note: Lemma A.1 (and hence Lemma A.2) are given for the natural logarithm, but under similar conditions on a and b, they also hold for log base 2; you can assume this without proof [this is true, though you may need to adjust the conditions on a and b for them to hold, but don't worry about that for this problem]. Also, you may interpret the logarithms in the bound as base 2 as well. Furthermore, note that we can actually show $\mathrm{VCdim}(\mathcal{H}) < 4d\log_2(2d) + 2\log_2(r)$ (i.e., a strict inequality).