TTIC 31250 An Introduction to the Theory of Machine Learning

Homework # 3 **Solutions**

Groundrules: Same as before. You should work on the exercises by yourself but may work with others on the problems (just write down who you worked with). Also if you use material from outside sources, say where you got it.

Exercises:

1. Think about what you would like to do for your project and propose it.

2. Consider the class $\mathcal{H}$ of axis-parallel rectangles in R^3. Specifically, a legal target function is specified by three intervals $[x_1^{min}, x_1^{max}]$, $[x_2^{min}, x_2^{max}]$, and $[x_3^{min}, x_3^{max}]$, and classifies an example (x_1, x_2, x_3) as positive if $x_1 \in [x_1^{min}, x_1^{max}]$ and $x_2 \in [x_2^{min}, x_2^{max}]$ and $x_3 \in [x_3^{min}, x_3^{max}]$; otherwise, the example is classified as negative. Argue that $\mathcal{H}[m] = O(m^6)$.

   Solution: Given m data points, there are at most $m + 1$ meaningful locations for x_{min}. More precisely, any two choices for x_{min} that agree in terms of which datapoints are less than x_{min} are equivalent in terms of the classification produced, and there are at most $m + 1$ equivalence classes. This holds for the other 5 parameters as well. So $\mathcal{C}[m] \leq (m + 1)^6 = O(m^6)$.

Problems:

3. [**VC-dimension of Two-Layer Networks**] Suppose that hypothesis class $\mathcal{H}$ has VC-dimension d. Now suppose we create a 2-layer network by choosing k functions $h_1, h_2, \ldots, h_k$ from $\mathcal{H}$ and then running their output through some other fixed Boolean function f. That is, given an input x, the network outputs $f(h_1(x), ..., h_k(x))$. For a given f, call the class of all such functions TWO-LAYER$_{f,k}(\mathcal{H})$. Show that TWO-LAYER$_{f,k}(\mathcal{H})$ has VC-dimension $O(kd \log kd)$. Note that we are only asking for an upper bound here, not a lower bound.

   Hint: Suppose you have a set S of m data points. By Sauer's lemma, we know there are at most $O(m^d)$ ways of labeling those points using functions in $\mathcal{H}$. Use that to get an upper bound on the number of ways of labeling those points using functions in TWO-LAYER$_{f,k}(\mathcal{H})$. Now select m so that this is less than 2^m which means the VC-dimension must be less than m.

   Solution: Let m be the VC-dimension of TWO-LAYER$_k(\mathcal{H})$, so by definition, there must exist a set S of m points shattered by TWO-LAYER$_k(\mathcal{H})$. We know by Sauer's lemma that there are at most m^d ways of partitioning the points in S using functions in $\mathcal{H}$. Since each function in TWO-LAYER$_k(\mathcal{H})$ is determined by k functions in $\mathcal{H}$, this means there are at most $(m^d)^k = m^{kd}$ ways of partitioning the points using functions in TWO-LAYER$_k(\mathcal{H})$. Since S is shattered, we therefore must have $2^m \leq m^{kd}$, or equivalently $m \leq kd\lg(m)$. We can solve this as follows. First, assuming $m \geq 16$ we have $\lg(m) \leq \sqrt{m}$ so $kd\lg(m) \leq kd\sqrt{m}$ which implies that $m \leq (kd)^2$. To get the better bound, we can then just plug back in to our original inequality, saying that since $m \leq (kd)^2$, it must be that $\lg(m) \leq 2\lg(kd)$, so our original inequality implies $m \leq 2kd\lg(kd)$.

In problems 4-6, you will prove that the VC-dimension of the class $\mathcal{C}_n$ of halfspaces in n dimensions is $n+1$. ($\mathcal{C}_n$ is the set of functions $a_1x_1 + \ldots + a_nx_n \geq a_0$, where $a_0, \ldots, a_n$ are real-valued.) We will use the following definition: The *convex hull* of a set of points S is the set of all convex combinations of points in S; this is the set of all points that can be written as $\sum_{x_i \in S} \lambda_i x_i$, where each $\lambda_i \geq 0$, and $\sum_i \lambda_i = 1$. It is not hard to see that if a halfspace has all points from a set S on one side, then it must have the entire convex hull of S on that side as well.

4. [**lower bound**] Prove that VC-dim$(\mathcal{C}_n) \geq n+1$ by presenting a set of $n+1$ points in n-dimensional space such that one can partition that set with halfspaces in all possible ways. (And, explain how one can partition the set in any desired way.)

   Solution: One good set of $n+1$ points is: the origin and all points with a 1 in one cooridinate and zeros in the rest (i.e, all neighbors of the origin on the Boolean cube). Let p_i be the point with a 1 in the ith coordinate. Suppose we wish to partition this set into two pieces S_1 and S_2 with a hyperplane (and, say, the origin is in S_1). Then just use the hyperplane:

   $$\sum_{\{i:p_i \in S_2\}} x_i = 1/2.$$

   (The halfspace would then replace "=" with "≤" or "≥" depending on which we want to be positive and which we want to be negative.)

5. [**upper bound part 1**] The following is "Radon's Theorem," from the 1920's.

   Theorem. *Let S be a set of $n+2$ points in n dimensions. Then S can be partitioned into two (disjoint) subsets S_1 and S_2 whose convex hulls intersect.*

   Show that Radon's Theorem implies that the VC-dimension of halfspaces is *at most* $n+1$. Conclude that VC-dim$(\mathcal{C}_n) = n+1$.

   Solution: If S is a set of $n+2$ points, then by Radon's theorem we may partition S into sets S_1 and S_2 whose convex hulls intersect. Let p be a point in that intersection. No hyperplane can have S_1 on one side and S_2 on the other since that would imply that the convex hull of S_1 is on one side and the convex hull of S_2 is on the other, which means that p is on both sides. So, no set of $n+2$ points can be shattered.

6. [**upper bound part 2**] Now we prove Radon's Theorem. We will need the following standard fact from linear algebra. If $x_1, \ldots, x_{n+1}$ are $n+1$ points in n-dimensional

space, then they are linearly dependent. That is, there exist real values $\lambda_1, \ldots, \lambda_{n+1}$ *not all zero* such that $\lambda_1 x_1 + \ldots + \lambda_{n+1} x_{n+1} = 0$.

You may now prove Radon's Theorem however you wish. However, as a suggested first step, prove the following. For any set of $n+2$ points $x_1, \ldots, x_{n+2}$ in n-dimensional space, there exist $\lambda_1, \ldots, \lambda_{n+2}$ *not all zero* such that $\sum_i \lambda_i x_i = 0$ and $\sum_i \lambda_i = 0$. (This is called *affine dependence.*) Now, think about the lambdas...

Solution: We first prove the theorem about affine dependence. Let S be a set of $n+2$ points $x_1, \ldots, x_{n+2}$. Now, add a new cooordinate and consider the points: $(x_1, 1), \ldots, (x_{n+2}, 1)$. We know that there exist real values $\lambda_1, \ldots, \lambda_{n+2}$ not all zero such that $\lambda_1(x_1, 1) + \ldots + \lambda_{n+2}(x_{n+2}, 1) = 0$. But this means that

$$\lambda_1 x_1 + \ldots + \lambda_{n+2} x_{n+2} = 0$$

and

$$\lambda_1 + \ldots + \lambda_{n+2} = 0.$$

We now prove Radon's Theorem given this fact. Let $S = \{x_1, \ldots, x_{n+2}\}$ and let $\lambda_1, \ldots, \lambda_{n+2}$ be as in the definition of affine dependence. Let $S_1 = \{x_i : \lambda_i > 0\}$ and let $S_2 = \{x_i : \lambda_i \leq 0\}$. Note that $\displaystyle\sum_{\{i : x_i \in S_1\}} \lambda_i x_i = \sum_{\{i : x_i \in S_2\}} -\lambda_i x_i$.

Let L be the sum of the λ_i in S_1 (so $-L$ is the sum of the λ_i in S_2). Then,

$$p = \sum_{\{i : x_i \in S_1\}} \frac{\lambda_i}{L} x_i$$

is in the convex hull of S_1 (using here the fact that the terms $\frac{\lambda_i}{L}$ are non-negative). But, p also is in the convex hull of S_2 since p is also the sum over $x_i \in S_2$ of $\frac{-\lambda_i}{L} x_i$ (where the $\frac{-\lambda_i}{L}$ terms are all non-negative). So, the convex hulls intersect.