10-806 Foundations of Machine Learning and Data Science

Homework # 2 **Due: October 12, 2015**

Groundrules:

- Your work will be graded on correctness, clarity, and conciseness. You should only submit work that you believe to be correct; if you cannot solve a problem completely, you will get significantly more partial credit if you clearly identify the gap(s) in your solution. It is good practice to start any long solution with an informal (but accurate) proof summary that describes the main idea.
- You may collaborate with others on this problem set and consult external sources. However, you must *write your own solutions* and *list your collaborators/sources* for each problem.

Problems:

1. [20 pts] **An Incorrect Proof.** Explain the bug in the following attempt to prove a sample complexity guarantee for the non-realizable case without dependence on the concept class C.

   Let A be an algorithm that takes the training sample S and outputs the hypothesis $h_A \in C$ of minimum empirical error. Let h^* denote the $h \in C$ of minimum true error.

   Now, for any hypothesis h, let B_h denote the bad event that $|\text{err}_S(h) - \text{err}_D(h)| > \epsilon$. We know by Hoeffding bounds that for a sample S of size at least $\frac{1}{2\epsilon^2}\ln(4/\delta)$, we have $\Pr_S(B_h) < \delta/2$.

   So, simply plug in $h = h^*$ and we have $\Pr_S(B_{h^*}) < \delta/2$, and also plug in $h = h_A$ and we have $\Pr_S(B_{h_A}) < \delta/2$. So, with probability at least $1-\delta$, neither bad event occurs, and so we have

   $$\text{err}_D(h_A) \leq \text{err}_S(h_A) + \epsilon \leq \text{err}_S(h^*) + \epsilon \leq \text{err}_D(h^*) + 2\epsilon$$

   as desired, with a sample size that has no dependence on C.

2. [40 pts] **VC-dimension of linear separators**. Let LTF_n denote the set of linear separators (linear threshold functions) in R^n. That is, LTF_n consists of all Boolean functions that can be described as "$f(\vec{x})$ = positive iff $a_1x_1 + \ldots + a_nx_n \geq a_0$," where $a_0, \ldots, a_n$ are real-valued.

   (a) Prove that VC-dim$(\text{LTF}_n) \geq n+1$ by presenting a set S of $n+1$ points in R^n such that one can label S in all 2^{n+1} possible ways using linear separators (and show how one can label S in any desired way.)

   (b) The following is "Radon's Theorem," from the 1920's. Note: 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 $\sum_i \lambda_i = 1$.

   Theorem. *Let S be a set of $n+2$ points in R^n. 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 VC-dim$(\text{LTF}_n) \leq n+1$.

(c) Now 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 R^n, 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 R^n, 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...

Thus, combining these parts together, the VC-dimension of linear separators in R^n is $n+1$.

3. [20 pts] **Rademacher complexity.** Consider 1-dimensional data (each example is a point on the real line). For real-valued a, define the function $f_a(x) = 1$ if $x \leq a$ and $f_a(x) = 0$ otherwise. Let $\mathcal{F} = \{f_a\}$. Consider a set S of m distinct examples on the line. What is the empirical Rademacher complexity $\hat{\mathcal{R}}_m(\mathcal{F})$ of $\mathcal{F}$ on S?

   There are several ways to analyze this. If you want, you may use the following interesting fact about gambling. Suppose at each time $t = 1, 2, 3, \ldots$ you bet \$1 on a fair game (with probability 1/2 you win \$1 and with probability 1/2 you lose \$1). After T total rounds, by linearity of expectation, your expected total winnings is \$0. However, if you look back and imagine you had stopped at the best possible time in hindsight (the time $t \leq T$ at which your total winnings were highest), the expected value of your winnings then is $\Theta(\sqrt{T})$; i.e., this is the expected maximum, over all $t \leq T$ of your winnings by time t.

4. [20 pts] **Sample Complexity Lower Bounds.** Prove that any algorithm for learning a concept class C with $|C| \geq 3$ must use $\Omega(\frac{1}{\epsilon} \log \frac{1}{\delta})$ examples in the worst case to learn with error parameter ϵ and confidence parameter δ.

   Hint: as a first step, show there must exist two examples x_1, x_2 and two functions $c_1, c_2 \in C$ such that $c_1(x_1) = c_2(x_1)$ but $c_1(x_2) \neq c_2(x_2)$.[1] Then prove that the distribution D that places $1 - 2\epsilon$ probability on x_1 and 2ϵ probability on x_2 has the property that any algorithm seeing $o(\frac{1}{\epsilon} \log \frac{1}{\delta})$ examples would have probability at least δ of having error at least ϵ for at least one of the two target functions c_1, c_2.

5. [20 pts extra credit] **VC-dimension of MAJ(H).** Show that if hypothesis class H has VC-dimension d, then the class $\text{MAJ}_k(H)$ has VC-dimension $O(kd \log kd)$. Here, we define $\text{MAJ}_k(H)$ to be the class of functions achievable by taking majority votes over k functions in H. Note that we are only asking for an upper bound here, not a lower bound.

6. [20 pts extra credit] **VC-dimension of boxes**. What is the VC-dimension V of the class $\mathcal{H}$ of axis-parallel boxes in R^d? That is, $\mathcal{H} = \{h_{\mathbf{a},\mathbf{b}} : \mathbf{a}, \mathbf{b} \in R^d\}$ where $h_{\mathbf{a},\mathbf{b}}(\mathbf{x}) = 1$ if $a_i \leq x_i \leq b_i$ for all $i = 1, \ldots, d$ and $h_{\mathbf{a},\mathbf{b}}(\mathbf{x}) = -1$ otherwise.

   (a) Prove that the VC-dimension is at least your chosen V by giving a set of V points that is shattered by the class (and explaining why it is shattered).

   (b) Prove that the VC-dimension is at most your chosen V by proving that no set of $V+1$ points can be shattered.

[1] This will use the fact that $|C| \geq 3$. If C had only two functions with one the negation of the other, then this would not be the case and indeed one could learn perfectly from just a single example.