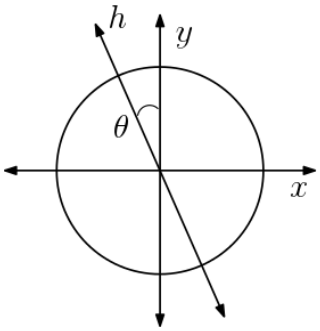
Homework 1

Christopher Mertin
CS6966: Theory of Machine Learning

February 1, 2017

1. Consider the problem of classifying points in the two-dimensional plane, *i.e.*, $\chi = \mathbb{R}^2$. Suppose that the (unknown) true label of a point (x, y) is given by $\text{sign}(x)$ (we define $\text{sign}(0) = \pm 1$ for convenience). Suppose the input distribution $\mathcal{D}$ is the uniform distribution over the unit circle centered at the origin.

   (a) Consider the hypothesis h as shown in the figure below (h classifies all the points on the right of the line as $+1$ and all the points to the left as -1). Compyte the risk $L_{\mathcal{D}}(h)$, as a function of θ (which is, as is standard, given in radians).



   (b) Suppose we obtain $1/\theta$ (whcih is given to be an integer ≥ 2) training samples (*i.e.*, samples from $\mathcal{D}$, along with their true labels). What is the probability that we find a point whose label is "inconsistent" with h? Can you bound this probability by a constant independent of θ?

   (c) Give an example of a distribution $\mathcal{D}$ under which h has risk zero.

2. Suppose $A_1, A_2, \ldots, A_n$ are events in a probability space.

   (a) Suppose $\Pr[A_i] = \frac{1}{2n}$ for all i. Then, show that the probability that none of the A_i's occur is at least $1/2$.

   (b) Give a concrete example of events A_i for which $\Pr[A_i] = \frac{1}{n-1}$ for all i, and the probability that none of them occur is zero.

   (c) Suppose $n \geq 3$, and $\Pr[A_i] = \frac{1}{n-1}$, but the events are all *independent*. Show that the probability that none of them occur is $\geq 1/8$.

3. In our proof of the no-free lunch theorem, we assumed the algorithm A to be deterministic. Let us now see how to allow randomized algorithms. Let A be a randomized map from set X to set Y. Formally, this means that for every $x \in X$, $A(x)$ is a random variabl, that takes values in Y. Suppose $|X| < c\,|Y|$, for some constant $c < 1$.

   (a) Show that there exists $y \in Y$ such that $\max_{x \in X} \Pr[A(x) = y] \leq c$.

   (b) Show that this implies that for any distribution $\mathcal{D}$ over X, $\Pr_{x \sim \mathcal{D}}(A(x) = y) \leq c$.

4. Recall that the VC dimension of a hypothesis class $\mathcal{H}$ is the size fo the largest set that it can "shatter."

   (a) Consider the task of classifying points on a 2D plane, and let $\mathcal{H}$ be the class of axis parallel rectangles (points inside the rectangle are "+" and the poitns outside are "−"). Prove that the VC dimension of $\mathcal{H}$ is 4.

   (b) This time, let $\chi = \mathbb{R}^d \backslash \{0\}$ (origin excluded), and let $\mathcal{H}$ be the set of all hyperplanes through the origin (points on one side are "+" and the other side are "−"). Prove that the VC dimension of $\mathcal{H}$ is $\leq d$.

   Hint: Consider *any* set of $d+1$ points. They need to be linearly dependent. Now, could it happen that u, v are "+", but $\alpha u + \beta v$ is "−" for $\alpha,\ \beta \geq 0$? Can you generalize this?

   (c) **(BONUS)** Let χ be the points on the real line, and let $\mathcal{H}$ be the class of hypotheses of the form $\text{sign}(p(x))$, where $p(x)$ is a polynomial of degree at most d (for convenience, define $\text{sign}(0) = +1$). Prove that the VC dimension of this class is $d + 1$.

   Hint: The tricky part is the uppoer bound. Here, suppose $d = 2$, and suppose we consider any four points $x_1 < x_2 < x_3 < x_4$. Can the sign pattern $+, -, +, -$ arise from a degree 2 polynomial?

5. In the examples above (and in general), a good rule of thumb for VC dimension of a function class is the *number of parameters* involved in defining a function in that class. However, this is not universally true, as illustrated in this problem: Let χ be the points on the real line, and define $\mathcal{H}$ to be the class of functions of the form $h_\theta \coloneqq \text{sign}(\sin(\theta x))$, for $\theta \in \mathbb{R}$. Note that each hypothesis is defined by the single parameter θ.

   Prove that the VC dimension of $\mathcal{H}$ is infinity.

   So where does the "complexity" of the function class come from? **(BONUS)** Prove that if we restrict θ to be a rational number whose numerator and denominator have at most n bits, then the VC dimension is $\mathcal{O}(n)$.