CS260: Machine Learning Theory
Problem Set 1
Due Monday, October 10, 2011

Ground Rules:

- This problem set is due at the beginning of class on October 10. Please bring a **hard copy** of your solutions with you to class. Slightly late assignments should be submitted directly to the grader, and will be penalized 25% (i.e., 25 points). No assignments will be accepted more than 24 hours late.
- You are strongly encouraged to discuss the problem set with other students in the class, as long as you follow the rules outlined in the course academic honesty policy. Don't forget to **list all of your collaborators** and **properly credit any sources** that you consult.
- All solutions must be typed; LaTeX is strongly recommended. Hand-written solutions will be penalized 25%, and unreadable answers will not be graded.
- You will be graded on both correctness and clarity. Be concise and clear, especially when writing proofs! If you cannot solve a problem completely, you will get more partial credit if you identify the gaps in your argument rather than trying to cover them up.

Problems:

1. **Finding a consistent hypothesis** *(20 points)*

   Let S be a set of m points labeled by an unknown function $c \in \mathcal{C}$. Suppose you are given an efficient algorithm $\mathcal{A}$ that PAC-learns $\mathcal{C}$ using $\mathcal{H}$. Prove that you can use $\mathcal{A}$ to find in polynomial time (in m and a confidence parameter δ) a hypothesis $h \in \mathcal{H}$ that is consistent with S with high probability.

2. **Learning n-dimensional axis-aligned boxes** *(35 points total)*

   In this problem, we consider the class of n-dimensional axis-aligned boxes. Each function c in this class is specified by a set of $2n$ values $\ell_1^c, \ell_2^c, \cdots, \ell_n^c$ and $u_1^c, u_2^c, \cdots, u_n^c$ which define an axis-aligned n-dimensional box. Given an n-dimensional input vector $\vec{x}$, $c(\vec{x})$ is defined to be 1 if for every $i \in \{1, \cdots, n\}$ the ith coordinate of $\vec{x}$ lies in $[\ell_i^c, u_i^c]$. Otherwise, $c(\vec{x})$ is defined to be 0.

   An example of a 2-dimensional axis-aligned box is shown in Figure 1.

   Hint: For each problem below, first think about the case when $n = 1$. Then move on to the $n = 2$ case before trying to find a general solution. (You do not have to provide separate proofs for these cases.)

   (a) Let $\mathcal{C}$ be the class of n-dimensional axis-aligned boxes, and $c \in \mathcal{C}$ be the unknown target function. State an efficient algorithm that takes as input an arbitrary set of points $\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_m$, each in $\mathbb{R}^n$, and their corresponding labels $c(\vec{x}_1), c(\vec{x}_2), \cdots, c(\vec{x}_m)$, and outputs a function $h \in \mathcal{C}$ that is consistent with this data. You may assume a model of computation in which real numbers can be stored in constant memory and in which basic operations on real numbers (addition, comparisons, etc.) take constant time. Your algorithm should be simple enough to implement in a few lines of code. *(10 points)*

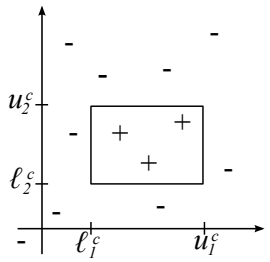

Figure 1: An example of a 2-dimensional axis-aligned box and some labeled points.

(b) Let $\mathcal{D}$ be an arbitrary, unknown distribution over points in $\mathbb{R}^n$, and let c be an arbitrary, unknown n-dimensional axis-aligned box. Suppose your algorithm from part a is given as input m points $\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_m$ drawn i.i.d. from $\mathcal{D}$ and their corresponding labels $c(\vec{x}_1), c(\vec{x}_2), \cdots, c(\vec{x}_m)$. Let h be the function output by your algorithm. Let δ be a fixed parameter in $(0, 1/2)$. Without referring to VC dimension (which we will get to a little later in the class), state and prove an upper bound on $\Pr_{\vec{x}\sim\mathcal{D}}[h(\vec{x}) \neq c(\vec{x})]$ that holds with probability at least $1-\delta$. Don't worry about getting the best possible constant factors in your bound, just focus on the important parameters (n, δ, m). *(25 points)*

3. **A Two-Distribution PAC Model** *(45 points total)*

In this problem, we consider a variant of the PAC model in which the learning algorithm may explicitly request positive examples or negative examples, but must find a hypothesis that performs well on both the marginal distribution over positive examples and the marginal distribution over negative examples.

Formally, we say that an algorithm $\mathcal{A}$ PAC-learns a concept class $\mathcal{C}$ using a hypothesis class $\mathcal{H}$ in the two-distribution variant of the PAC model if for any $c \in \mathcal{C}$, for any distribution $\mathcal{D}_c^+$ over the the subset of instance space that c labels positively and any distribution $\mathcal{D}_c^-$ over the subset of instance space that c labels negatively, for any $\epsilon \in (0, 1/2)$ and $\delta \in (0, 1/2)$, given access to a polynomial (in $1/\epsilon$ and $1/\delta$) number of examples drawn i.i.d. from $\mathcal{D}_c^+$ and a polynomial number of examples drawn i.i.d. from $\mathcal{D}_c^-$, $\mathcal{A}$ outputs a function $h \in \mathcal{H}$ such that with probability at least $1-\delta$, $\Pr_{x\sim\mathcal{D}_c^+}[h(x) = 0] \leq \epsilon$ and $\Pr_{x\sim\mathcal{D}_c^-}[h(x) = 1] \leq \epsilon$.

$\mathcal{A}$ efficiently PAC-learns $\mathcal{C}$ using $\mathcal{H}$ in the two-distribution variant if its running time is bounded by a polynomial in $1/\delta$ and $1/\epsilon$.

In the following problems, use the basic ("preliminary") definition of the PAC model that was given in class on September 28.

(a) Prove that if $\mathcal{C}$ is efficiently PAC-learnable using $\mathcal{H}$ in the basic (one distribution) model, then $\mathcal{C}$ is efficiently PAC-learnable using $\mathcal{H}$ in the two-distribution model. *(20 points)*

(b) Let h_0 be a function that always outputs 0, and h_1 be a function that always outputs 1. Prove that if $\mathcal{C}$ is efficiently PAC-learnable using $\mathcal{H}$ in the two-distribution model, then C is efficiently PAC-learnable using $\mathcal{H} \cup \{h_0, h_1\}$ in the basic model. Hint: You may wish to use Hoeffding's inequality (which we will cover in class, but not until October 5) or Chebyshev's inequality in your solution. Try to give a formal proof, but if you cannot, provide a sketch of how the argument should go. *(25 points)*