CS260: Machine Learning Theory
Problem Set 2
Due Monday, October 24, 2011

Ground Rules:

- This problem set is due at the beginning of class on October 24. 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. **Uniform convergence** *(25 points)*

   Consider a finite concept class $\mathcal{C}$ and finite hypothesis class $\mathcal{H}$ with $\mathcal{C} \subseteq \mathcal{H}$. Let P be a probability distribution over $\mathcal{H}$, with $P(h) > 0$ for every $h \in \mathcal{H}$. Let $\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_m$ be a set of points drawn i.i.d. from a fixed but unknown distribution $\mathcal{D}$. Define $\mathrm{err}(h) = \Pr_{\vec{x} \sim \mathcal{D}}(h(\vec{x}) \neq c(\vec{x}))$ and $\widehat{\mathrm{err}}(h) = (1/m)\,|\{i : h(\vec{x_i}) \neq c(\vec{x_i})\}|$.

   Let δ be a fixed parameter in $(0, 1/2)$. Prove that with probability at least $1 - \delta$, for every $h \in \mathcal{H}$,

   $$|\mathrm{err}(h) - \widehat{\mathrm{err}}(h)| \leq \sqrt{\frac{-k \ln\left(\frac{P(h)\delta}{2}\right)}{m}}$$

   for some constant k, where the probability is taken over the random choice of the sample $\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_m$.

2. **Learning n-dimensional axis-aligned boxes (continued)** *(40 points total)*

   Recall the class of n-dimensional axis-aligned boxes that you examined in Problem Set 1. 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 $u_1^c, u_2^c, \cdots, u_n^c$. 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.

(a) What is the VC dimension of the class of n-dimensional axis-aligned boxes? Provide a proof of your answer. This proof requires two parts: a proof that *at least* d points can be shattered, and a proof that *no more than* d points can be shattered for some d. *(30 points)*

(b) Let $\mathcal{A}$ be an algorithm that learns n-dimensional axis-aligned boxes in the consistency model, such as the algorithm that you defined in Problem Set 1. Let $\mathcal{D}$ be a fixed, unknown distribution over points in $\mathbb{R}^n$, let c be a fixed, unknown n-dimensional axis-aligned box, and let δ be a fixed parameter in $(0, 1/2)$. Suppose $\mathcal{A}$ 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 $c(\vec{x}_1), c(\vec{x}_2), \cdots, c(\vec{x}_m)$, and it outputs a function h. Using the VC dimension that you derived in part a, state a rough upper bound (big-O notation is ok) on $\Pr_{\vec{x}\sim\mathcal{D}}(h(\vec{x}) \neq c(\vec{x}))$ that holds with probability at least $1 - \delta$. How does this bound compare to the bound you derived in Problem Set 1? (Don't worry about constants in your comparison.) *(10 points)*

3. **VC dimension of linear threshold functions** *(35 points total)*

Consider the class of linear threshold functions which pass through the origin in n-dimensions. Each function c in this class can be represented as an n-dimensional weight vector $\vec{w}_c$ such that $c(\vec{x}) = 1$ if $\vec{w}_c \cdot \vec{x} \geq 0$ and $c(\vec{x}) = 0$ otherwise. In this problem, you will prove that the VC dimension of this class is n.

Hint: Once again, it might help you to first think about small values of n before trying to find a general solution.

(a) First prove that the VC dimension of the set of linear thresholds passing through the origin in n dimensions is *at least* n. That is, describe a set of n points that can be shattered by this class, and a give brief argument showing that they can be shattered. Your solution should hold for arbitrary values of n, not just one specific value. *(15 points)*

(b) Show that no set of $n + 1$ points can be shattered by this class. Hint: Recall that a set of k points $\vec{x}_1, \cdots, \vec{x}_k$ is linearly dependent if $\sum_{i=1}^{k} \lambda_i \vec{x}_i = 0$ for some values $\lambda_1, \cdots, \lambda_k$ such that at least one λ_i is non-zero. That is, there exists one point that can be written as a weighted sum of the others. Start by showing that no set of linearly dependent points can be shattered by this class. *(20 points)*