Homework 2, due Tuesday 10/17

1. *McDiarmid's bounded-difference inequality.* Suppose function $f : \mathcal{X}^n \to \mathbb{R}$ has the property that changing the ith input can change f by at most c_i; formally, $|f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)| \leq c_i$ for any $x_1, \ldots, x_n, x_i'$. Then, it is known that for independent random variables $X_1, \ldots, X_n$,

$$\mathbb{P}[f(X_1, \ldots, X_n) \geq \mathbb{E}f + \epsilon] \;\leq\; e^{-2\epsilon^2/\sum_i c_i^2}.$$

   Show that this inequality is an immediate consequence of our general concentration result for metric spaces (Lecture 1, Theorem 23).

2. Show that the class of axis-parallel rectangles in $\mathbb{R}^n$ is efficiently PAC-learnable. Make sure to bound the number of samples needed.

3. A particular hypothesis class H is PAC-learnable when the δ parameter is constrained to be 0.9. Show that H is PAC-learnable with arbitrary $0 < \delta < 1$.

4. Consider a variant of the PAC model in which there are *two* example oracles: one which generates positive examples and one which generates negative examples, both according to the underlying distribution $\mathcal{D}$ on $\mathcal{X} \times \{-1, +1\}$. In other words, these two oracles generate samples from distributions $\mathcal{D}^+$ and $\mathcal{D}^-$ respectively, where

$$\mathcal{D}^+(x) \propto \mathcal{D}(x, 1) \quad \text{ and } \quad \mathcal{D}^+(x) \propto \mathcal{D}(x, -1).$$

   The learning algorithm is now given as input ϵ, δ, and the two oracles, and it must return a hypothesis $h \in H$ such that with probability at least $1 - \delta$, both $\mathbb{P}_{x \in \mathcal{D}^+}(h(x) = -1) \leq \epsilon$ and $\mathbb{P}_{x \in \mathcal{D}^-}(h(x) = 1) \leq \epsilon$.

   Suppose H includes h_+, the always-plus concept, and h_-, the always-minus concept. Show that H is efficiently PAC-learnable (in the standard one-oracle model) if and only if H is efficiently PAC-learnable in the two-oracle model.

5. Show that over input space $\mathcal{X} = \{0, 1\}^n$, the class of decision lists is expressive enough to include all conjunctions and disjunctions, but not all 3CNF formulas.

6. Suppose we are getting data points that are independent but not identically distributed; the ith point (X_i, Y_i) comes from some (unknown) distribution $\mathcal{D}_i$. Let $\overline{\mathcal{D}}_t$ denote the *average* distribution upto time t, that is, $\overline{\mathcal{D}}_t = (\mathcal{D}_1 + \cdots + \mathcal{D}_t)/t$.

   Show that if we get m points in this manner, then for any $0 < \epsilon < 1$,

$$\mathbb{P}[\text{some } h \in H \text{ with } \mathrm{err}_{\overline{\mathcal{D}}_m}(h) > \epsilon \text{ is consistent with the } m \text{ points}] \;\leq\; |H| e^{-\epsilon m}.$$