CS260: Machine Learning Theory
Problem Set 4
Due Monday, November 21, 2011

Ground Rules:

- This problem set is due at the beginning of class on November 21. 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. **Start early!**
- 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.
- 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. **Please make sure you define any notation that you use or variables that you introduce!!** You will be graded on what you wrote, not what you intended to write.

Problems:

1. **Distribution-Specific Weak Learning** *(15 points)*

   The PAC learning model is said to be *distribution-free* because of the requirement that it must be possible to achieve low error for *any* distribution $\mathcal{D}$ over the input space. It is possible to remove this requirement and define a distribution-specific notion of learnability as follows.

   Definition 1. *A concept class $\mathcal{C}$ is (strongly) PAC learnable for the distribution $\mathcal{D}$ using a hypothesis class $\mathcal{H}$ if there exists an algorithm $\mathcal{A}$ such that for any $c \in \mathcal{C}$, 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 $\mathcal{D}$ and labeled by c, $\mathcal{A}$ outputs a function $h \in \mathcal{H}$ such that with probability at least $1 - \delta$, $\mathrm{err}(h) \leq \epsilon$.*

   We could similarly define a distribution-specific notion of *weak* PAC learnability.

   In this problem, you will show that distribution-specific weak learnability *does not* imply distribution-specific strong learnability. More specifically, let $\mathcal{C}$ be the class of all 2^n functions from $\{-1, 1\}^n$ to $\{-1, 1\}$. Find a distribution $\mathcal{D}$ over $\{-1, 1\}^n$ such that $\mathcal{C}$ is weakly learnable for $\mathcal{D}$ using some hypothesis class $\mathcal{H}$ (e.g., $\mathcal{H} = \mathcal{C}$) but $\mathcal{C}$ is not strongly learnable for $\mathcal{D}$. Provide an informal proof sketch that shows that $\mathcal{C}$ is weakly learnable for $\mathcal{D}$ and not strongly learnable for $\mathcal{D}$. ("Informal" means you don't need to work out every detail, but your argument should be convincing.)

2. **Examining AdaBoost's Distributions** *(25 points)*

   Suppose that we run AdaBoost on a set of labeled points $(\vec{x}_1, y_1), \cdots, (\vec{x}_m, y_m)$ using a weak learning algorithm $\mathcal{A}$ that outputs a function from a class $\mathcal{H}$ that does *not* contain any function that perfectly classifies all m input points. Suppose further that after running AdaBoost for T rounds, we find that

the weak learning assumption holds at each round. That is, for every $t \in \{1, \cdots, T\}$, $\epsilon_t \leq 1/2 - \gamma$ for some constant $\gamma > 0$. Prove that there cannot be any round t such that $h_t = h_{t+1}$.

3. **Anytime Randomized Weighted Majority** *(60 points total)*

In class we proved a regret bound of $O(\sqrt{T \log n})$ for Randomized Weighted Majority under the assumption that the number of rounds T is known in advance and can therefore be used to set the parameter η. In this problem, we will prove a regret bound for an alternative version of Randomized Weighted Majority that holds when the number of rounds is unknown.

(a) Consider the class of Time-Sensitive Follow the Regularized Leader algorithms, that use weights of the form

$$\vec{p}_t = \arg\min_{\vec{p} \in \Delta_n} \left(\sum_{s=1}^{t-1} \vec{\ell}_s \cdot \vec{p} + \frac{1}{\eta_t} R(\vec{p}) \right),$$

where the η_t are arbitrary positive parameters satisfying $\eta_t \geq \eta_{t+1}$ for all t, and R is an arbitrary convex function. Prove the following lemma.

Lemma 1. *Let $\vec{p}_t$ be the distribution chosen by the Time-Sensitive Follow the Regularized Leader algorithm at time t. For any $\vec{p} \in \Delta_n$ and any time T,*

$$\sum_{t=1}^{T} \vec{\ell}_t \cdot \vec{p}_{t+1} - \sum_{t=1}^{T} \vec{\ell}_t \cdot \vec{p} \leq \frac{1}{\eta_{T+1}} \left(R(\vec{p}) - R(\vec{p}_1) \right) .$$

It might help you to define a quantity

$$\delta_t = \frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}.$$

to use in your analysis. (*20 points*)

(b) Define the Anytime RWM algorithm as follows. For each round $t = 1, 2, \cdots$, for each expert $i \in \{1, \cdots, n\}$, the Anytime RWM sets

$$p_{i,t} = \frac{e^{-\eta_t L_{i,t-1}}}{\sum_{j=1}^{n} e^{-\eta_t L_{j,t-1}}}$$

for some parameters $\eta_1, \eta_2, \cdots > 0$ that satisfy $\eta_t > \eta_{t+1}$ for all t. (We will set these parameters in the next part of the problem.)

Prove the following lemma for some constant c. (Aim for $c = 1$, but if you get a larger constant, it's ok.)

Lemma 2. *For any sequence of losses $\vec{\ell}_1, \cdots, \vec{\ell}_T$ with each $\ell_{i,t} \in [0, 1]$, let $\vec{p}_1, \cdots, \vec{p}_T$ be the distributions chosen by the Anytime Randomized Weighted Majority. For all t,*

$$\vec{\ell}_t \cdot (\vec{p}_t - \vec{p}_{t+1}) \leq c \left((t+1)\eta_t - t\eta_{t+1} \right).$$

It might help you to define another distribution $\vec{q}$ with

$$q_i = \frac{e^{-\eta_t L_{i,t}}}{\sum_{j=1}^{n} e^{-\eta_t L_{j,t}}}$$

and then separately bound $\vec{\ell}_t \cdot (\vec{p}_t - \vec{q})$ and $\vec{\ell}_t \cdot (\vec{q} - \vec{p}_{t+1})$. You also might want to use the old $1 + x \leq e^x$ trick. For parts of your analysis that are identical to proofs we went over in class, you can simply reference the relevant proofs. (*20 points*)

(c) Set the Anytime RWM parameters for each t to be

$$\eta_t = \sqrt{\frac{\log n}{t}}.$$

Use Lemmas 1 and 2 to bound the regret of the Anytime RWM algorithm after T time steps. Ignoring constants, your bound should be similar to the $O(\sqrt{T \log n})$ bound derived in class for the case that T is a specific known value.

It might help you to start by showing that

$$\sum_{t=1}^{T} \frac{1}{\sqrt{t}} = O(\sqrt{T}) \; .$$

You do not need to have a correct solution to parts a or b to complete this problem. (*20 points*)