15-859(B) Machine Learning Theory

Homework # 6 **Due: April 30, 2012**

Groundrules: Same as before. You should work on the exercises by yourself but may work with others on the problems (just write down who you worked with). Also if you use material from outside sources, say where you got it.

Exercises:

1. **Online investing.** You can think of the "combining expert advice" setting as modeling a kind of online investment problem: each day you have $\$M$ to invest, you probabilistically choose one of n investments to put it in, and then at the end of the day you find out how well you did, along with how well you would have done had you chosen each of the other investments. Suppose, however, that you are required to split your $\$M$ equally among k investments each day. You may probabilistically decide how to do it, but every day you must choose exactly k investments to put your money in for that day. We could model this as having $\binom{n}{k}$ experts, but that is exponential in k. Show how this can instead be modeled in the Kalai-Vempala framework to get a polynomial-time regret-minimizing algorithm. Make sure to argue you can solve the offline problem.

Problems:

2. **Policy iteration.** The goal of this problem is to prove that a method called "policy iteration" will eventually reach an optimal policy in an MDP. In policy iteration, given some policy π_i (a mapping of states to actions), you solve a linear system to compute the state values under that policy:

$$V^{\pi_i}(s) = R(s, \pi_i(s)) + \gamma \sum_{s'} \Pr_{s,\pi_i(s)}(s') V^{\pi_i}(s').$$

(Here, "$R(s,a)$" is the expected reward of executing action a from state s.) Then, we define policy π_{i+1} to be the greedy policy with respect to those values. That is,

$$\pi_{i+1}(s) = \arg\max_a \left[R(s,a) + \gamma \sum_{s'} \Pr_{s,a}(s') V^{\pi_i}(s') \right],$$

and so on to $\pi_{i+2}, \pi_{i+3}, \ldots$.

   (a) As an easy first step, argue that if $\pi_{i+1} = \pi_i$ (i.e., $\pi_{i+1}(s) = \pi_i(s)$ for all states s), then π_i is optimal.

   (b) As the harder second step, argue that the values never decrease (i.e., for all s, $V^{\pi_{i+1}}(s) \geq V^{\pi_i}(s)$). This completes the argument because there are only a finite number of different policies.

   Hint: what about a hybrid policy that uses π_{i+1} for one step and then π_i from then on? How about π_{i+1} for two steps?

3. **Sample complexity bounds.** For some learning algorithms, the hypothesis produced can be uniquely described by a small subset of k of the training examples. E.g., if you are learning an interval on the line using the simple algorithm "take the smallest interval that encloses all the positive examples," then the hypothesis can be reconstructed from just the outermost positive examples, so $k = 2$. For a conservative Mistake-Bound learning algorithm, you can reconstruct the hypothesis by just looking at the examples on which a mistake was made, so $k \leq M$, where M is the algorithm's mistake-bound. (In this case, you may also care about the *order* in which those examples arrived.)

   Prove a PAC guarantee based on k. Specifically, fixing a description language (reconstruction procedure), so for a given set S' of examples we have a well-defined hypothesis $h_{S'}$, show that

$$\Pr_{S \sim D^n} \left(\exists S' \subseteq S, |S'| = k, \text{ such that } h_{S'} \text{ has 0 error on } S - S' \text{ but true error } > \epsilon \right) \leq \delta,$$

   so long as

$$n \geq \frac{1}{\epsilon} \left(k \ln n + \epsilon k + \ln \frac{1}{\delta} \right).$$

   Hint: This problem is not hard, but it requires care, so you should be very clear in your analysis what events you are taking a union bound over. In particular, there are potentially an infinite number of possible hypotheses $h_{S'}$ so you don't want to do a union bound over all sets $S' \sim D^k$. Instead you may want to think about sets of indices of the examples in S.

   Note the similarity of the form of this bound to VC-dimension and other bounds we have seen. These are often called "compression bounds".