15-854 Machine Learning Theory

Homework # 4 **Due: March 2, 1998**

In class, we proved that if see at least $2/\epsilon[\log_2(2H[2m]) + \log_2(1/\delta)]$ examples, then with high probability, all $h \in H$ with true error greater than ϵ will have observed error greater than 0 (i.e., will be inconsistent with the training data). In this assignment, we will prove a similar kind of bound for the stronger goal of ensuring that with high probability, all $h \in H$ have their observed error within an additive ϵ of their true error.

To recap, the argument in class went like this. The event A that we care about is the event that a "bad" hypothesis is consistent with the m training examples. An easier event to reason about, however, is the event B that some hypothesis is consistent with m training examples and then makes more that $\varepsilon m/2$ mistakes on the following m test examples. If event A occurs, then B probably will occur too (since the bad hypothesis will probably make a lot of mistakes on the test examples) so it suffices to show that event B is unlikely. We then analyzed event B by imagining that we first selected a set of $2m$ points (which by definition can be split in at most $H[2m]$ ways by concepts in H) and then we randomly split this set into training and testing portions. We then bounded the probability that any of these $H[2m]$ hypotheses fooled us over the choice of random splits.

Exercises:

1. Let A be the event that, given m training examples drawn from distribution $\mathcal{D}$,

$$\text{for some } h \in H,\ |\text{training-error}(h) - \text{true-error}(h)| > \varepsilon.$$

Let B be the event that, given m training examples and m test examples from $\mathcal{D}$,

$$\text{for some } h \in H,\ |\text{training-error}(h) - \text{test-error}(h)| > \varepsilon/2.$$

Prove that $\Pr[A] \leq 2\Pr[B]$. Specifically, suppose h has true error p, observed training error $\tilde{p}$, and $|p - \tilde{p}| > \varepsilon$. Give a bound on m such that with probability at least 1/2, the observed error $\hat{p}$ of h on m new test examples will satisfy $|\tilde{p} - \hat{p}| > \varepsilon/2$.

(In class we did this in a hand-wavy way for the case that $\tilde{p} = 0$.)

Problems:

2. Let's fix a set of examples S and fix a hypothesis h that makes M mistakes over S. Suppose we randomly color each example in S red or blue and let S_1 be the set of red examples and S_2 be the set of blue examples. Let M_1 be the number of mistakes of h on S_1 and let M_2 be the number of mistakes of h on S_2. Prove that

$$\Pr(|M_1 - M_2| > L) \leq 2e^{-L^2/(2M)}.$$

Make sure to be clear about how you are legally applying Chernoff/Hoeffding bounds.

3. Same setup as Problem 2, but instead of randomly coloring each example red or blue independently, we pick a random fixed fraction f of the points in S to color red. Let $P_f(L, M) = \Pr(|M_1 - M_2| > L)$ under this experiment.

   (a) (extra credit) Prove that for any given L and M, $P_f(L, M)$ is minimized at $f = 1/2$ over all values of f. For fewer extra credit points, prove that $P_{1/2}(L, M) \leq 2P_f(L, M)$ for any f.

   In other words the probability that the number of training mistakes is very different from the number of testing mistakes will only go up if you don't make an even split of the examples into training and testing portions.

   (b) Using part (a), prove that $P_{1/2}(L, M) \leq 2e^{-L^2/(2M)}$.

4. Combining all of the above, prove that the probability of event A from Exercise 1 is at most $4H[2m]e^{-\varepsilon^2 m/8}$. Therefore, if

$$m \geq \frac{8}{\varepsilon^2}\left[\ln H[2m] + \ln \frac{4}{\delta}\right],$$

   then with high probability all $h \in H$ have their observed error within ε of their true error.