15-859(B) Machine Learning Theory
Probabilistic inequalities

A common question that comes up in machine learning is: "Given some fixed hypothesis h, how much data do I need to see so that I can be confident that its observed error will be near to its true error?" We have already seen this as: "What's the chance that a given hypothesis of true error ϵ or more will have observed error of 0?"

In this case, from first principles we calculated the probability to be $(1-\epsilon)^m$, where m is the number of examples. We then noticed that if there are N hypotheses under consideration, and this quantity is δ/N, then there is at most a δ chance there exists *any* hypothesis that fools us. Solving for m, we got that $\frac{1}{\epsilon}\left[\ln(N)+\ln(\frac{1}{\delta})\right]$ examples suffice.

More generally, suppose we consider a hypothesis with true error p, and let $q = 1-p$. If we see m examples, then the expected fraction of mistakes is p. The *standard deviation* σ of this quantity is $\sqrt{pq/m}$. We can now use a convenient rule for independent identically distributed Bernoulli trials, which in our terminology is:

$$\Pr[|\text{observed error} - \text{true error}| > 1.96\sigma] < 0.05.$$

For instance, if we want with 95% confidence for our true and observed errors to differ by only ϵ, then we need to see only $4pq/\epsilon^2$ examples (approximating 1.96^2 by 4), which is at most $1/\epsilon^2$. (We want to get rid of the p and q since we don't know them).

The above rule is convenient for testing out a single hypothesis on test data. More generally, we'd like a rule that can be applied for any desired confidence level (like the first case above) but holds even when the observed error is not 0 (like the second case above). For instance, we might want to prove uniform convergence results, or we might just want to test out several hypotheses. Luckily, there are some convenient inequalities for doing this, known as Hoeffding and Chernoff bounds. These bounds state the following. (Actually, they hold for a variety of situations we won't describe here; a good book that discusses how these are derived and many other things is Alon and Spencer's *The Probabilistic Method.*)

Consider a hypothesis with true error rate p (or a coin of bias p) observed on m examples (the coin is flipped m times). Let S be the number of observed errors (the number of heads seen) so S/m is the observed error rate.

Hoeffding bounds state that for any $\epsilon \in [0,1]$,

1. $\Pr[\frac{S}{m} > p+\epsilon] \leq e^{-2m\epsilon^2}$, and
2. $\Pr[\frac{S}{m} < p-\epsilon] \leq e^{-2m\epsilon^2}$.

Chernoff bounds state that under the same conditions (using α to avoid confusion, $\alpha \in [0,1]$),

1. $\Pr[\frac{S}{m} > p(1+\alpha)] \leq e^{-mp\alpha^2/3}$, and
2. $\Pr[\frac{S}{m} < p(1-\alpha)] \leq e^{-mp\alpha^2/2}$.

Let's do a quick example of using Hoeffding bounds. Suppose we'd like to say that with high probability $1-\delta$, all hypotheses in H have their observed error within ϵ of their true error. To achieve this, we just need for each hypothesis individually to have confidence parameter $\delta/|H|$. So, we set $2e^{-2m\epsilon^2} \leq \delta/|H|$, or:

$$2m\epsilon^2 \geq \ln(2|H|/\delta).$$

We can now either solve for ϵ as a function of m and δ or solve for m as a function of ϵ and δ. Solving for ϵ, we find that our goal is achieved for ϵ satisfying:

$$\epsilon \geq \sqrt{\frac{\ln(2|H|/\delta)}{2m}}.$$

This says roughly that it suffices to go $\sqrt{2\ln(2|H|/\delta)}$ standard deviations to get our desired confidence. Actually, this is a nice way of thinking about the meaning of Hoeffding bounds: they say that if we want to replace the "0.05" in the earlier equation with some arbitrary small δ', then it suffices to go about $\sqrt{2\ln(2/\delta')}$ standard deviations.

Solving for m we get:

$$m \geq \frac{1}{2\epsilon^2}\ln(2|H|/\delta) = \frac{1}{2\epsilon^2}[\ln(|H|) + \ln(2/\delta)].$$

This is a lot like the first PAC bound, but we are now quadratic in $1/\epsilon$.

What about that pesky ϵ^2? If we believe the best $h \in H$ has low error, say $\leq \epsilon/2$, we can get rid of the ϵ^2 by using Chernoff bounds instead. In particular we can ask: how many examples do we need such that with high probability all hypotheses of true error $> 2\epsilon$ have empirical error $> \epsilon$, and all of true error $\leq \epsilon/2$ have empirical error $\leq \epsilon$? (So, this implies that the hypothesis of minimum empirical error will have true error at most 2ϵ).

Using Chernoff bounds, we calculate as follows. If $err(h) = p \geq 2\epsilon$ and we want empirical error at least $p/2 \geq \epsilon$ with confidence δ', we can solve $e^{-mp/8} \leq \delta'$ to get that it suffices to have $m \geq \frac{8}{p}\ln(1/\delta')$ (and we can replace p with 2ϵ). On the other hand, if $err(h) = p \leq \epsilon/2$, we can write our goal as saying that we want the observed error to be no more than $\frac{\epsilon}{2}(1+1)$, which by Chernoff bounds implies that $m \geq \frac{6}{\epsilon}\ln(1/\delta')$ examples suffice. So, setting $\delta' = \delta/|H|$ we get that:

$$m \geq \frac{6}{\epsilon}[\ln(|H|) + \ln(1/\delta)]$$

examples suffice so that if the target has true error $\leq \epsilon/2$ then with high probability the hypothesis of minimum empirical error will have true error less than 2ϵ.