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 seen this in several forms so far:

- "What's the chance that a given rule of true error ε or more will have observed error of 0?"

  In this case, from first principles we calculated the chance to be $(1-\varepsilon)^m$, where m is the number of examples. We then said that if there are $|H|$ hypotheses under consideration, then we want this chance to be $\delta/|H|$, so that there is at most a δ chance there exists *any* hypothesis that fools us. Solving for m, we got:

  $$m \geq \frac{1}{\varepsilon}\left[\ln(|H|) + \ln(\frac{1}{\delta})\right].$$

- If we fix the desired confidence to 95%, then a convenient rule says:

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

  where σ is the standard deviation. For instance, for the case of a hypothesis with true error p, if we want with 95% confidence for our true and observed errors to differ by only ε, then we need to see $m \geq 4p(1-p)/\varepsilon^2$ examples (approximating 1.96^2 by 4).

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). 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. The Hoeffding bounds state that for any $\varepsilon \in [0,1]$,

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

The Chernoff bounds state that under the same conditions,

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

Let's do a quick example of using Hoeffding bounds. Suppose we'd like to say that with high probability $(1 - \delta)$, there will be no hypothesis in H with observed error more than ε *less than* its true error. To achieve this, we just need for each hypothesis individually to have confidence parameter $\delta/|H|$. So, we set $e^{-2m\varepsilon^2} \leq \delta/|H|$, or:

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

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 get:

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

This says roughly that we need to go $\sqrt{\ln(|H|/\delta)}$ standard deviations to get our desired confidence.

Solving for m we get:

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

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