The bounds that tie

Danny Heap

Past and future

In machine learning we try to use the classification of examples we've seen (informally, the past) in order to predict the classification of examples we haven't seen (the future). A shrewd adversary can skew the past so that any hypothesis based on it tends to do very badly in the future (how badly is the subject of the Mistake Bound model). In order to limit the amount of mischief such an adversary is capable of, we can insist that past and future be connected: examples of both sorts are from the same distribution, randomly chosen according to a probability function that is fixed in advance.

Distributions form a strong, though elastic, connection between past and future. The error rate of a given hypothesis, $\text{err}(h)$, on the entire distribution is generally not identical to h's observed error on an initial sequence from the distribution, $\text{freq}(h)$. However, the probability that $\text{freq}(h)$ is much greater or smaller than $\text{err}(h)$ tails off as $\text{freq}(h)/\text{err}(h)$ gets large or small. I derive some of the techniques we've used to bound this probability, since I think some of the methods in their derivations can be recycled in other situations, and a technique derived from (close to) scratch is easier to remember.

Lazy bounds

In general you'd like to say something predictive while making a few assumptions as possible about a distribution. For any random non-negative variable X with expected value $E(X)$ something can be said about the probability that X exceeds some positive multiple of its expected value, $\kappa E(X)$, even though the connection between a point and average is pretty weak. Probabilities are in the interval $[0, 1]$, so it's easy to make an informal underestimate of $E(X)$ as follows:

$$E(X) \geq \Pr[X > \kappa E(X)] \cdot \kappa E(X) + \Pr[X \leq \kappa E(X)] \cdot 0 = \Pr[X > \kappa E(X)] \cdot E(X)\kappa$$

Dividing by $E(X)\kappa$ gives $\Pr[X > \kappa E(X)] \leq 1/\kappa$, the Markov Bound. This gives you no information if $\kappa \leq 1$, and much less information than other bounds if $\kappa > 1$. But it's an easy derivation to remember (hence attractive when it does the job), and it's tempting to take the same approach for bounding the bottom tail. Suppose you knew that X takes values in $[0, 1]$ (perhaps X is the distribution of some probability function), then you might want to bound how far X gets from its expected value:

$$E(X) \leq \Pr[X > \kappa E(X)] \cdot 1 + \Pr[X \leq \kappa E(X)] \cdot \kappa E(X) \leq \Pr[X > \kappa E(X)] + \kappa E(X)$$

...so $\Pr[X > \kappa E(X)] \geq (1-\kappa)E(X)$, which is only helpful if $\kappa < 1$, and even then is truncated at $E(X)$. You can do a little better if X is capped by some multiple of $E(X)$, say $X \leq \lambda E(X)$, denoting $\Pr[X > \kappa E(X)]$ as p:

$$E(X) \leq p\lambda E(X) + (1-p)\kappa E(X) \Longrightarrow p \geq \frac{1-\kappa}{1-\lambda}.$$

Additive Chernoff bounds

In exchange for very little knowledge of the distribution of X, Markov bounds give you very weak bounds on the tails of the distribution. In most cases you know a little bit more: membership in a target concept is boolean (an example is either in or out), the labels are either $+$ or $-$, and hypotheses are either consistent with an example or not. This bit of structure means that, for a given hypothesis h, the examples that are consistent and inconsistent with h form a distribution of their own, a sequence of independent Bernoulli trials that either succeed ($+$) or fail ($-$). This distribution has tails that drop off sharply (exponentially), and we can bound them much closer than we can with a Markov bound.

You can improve on Markov bounds by transforming the reciprocal bound, $\Pr[X \geq \kappa] \leq E(X)/\kappa$, into a reciprocal exponential bound, $\Pr[e^X \geq e^\kappa] \leq E[e^X]/e^\kappa$. Since $\exp(x)$ is monotonic increasing, it preserves inequalities (so the two probabilities are the same), and the new bound has an exponential, rather than linear, denominator.[1] The inequalities are still preserved if you add an extra non-negative factor t (for tuning, I guess) to the exponent, yielding $\Pr[e^{tX} \geq e^{t\kappa}] \leq E[e^{tX}]/e^{t\kappa}$. The extra parameter t will allow you to optimize the inequality. Now your bound is

$$\Pr[X \geq \kappa] = \Pr[e^{tX} \geq e^{t\kappa}] \leq \frac{E[e^{tX}]}{e^{t\kappa}} = e^{-(t\kappa - \ln E[e^{tX}])}.$$

The bound is minimized exactly when $t\kappa - \ln E[e^{tX}]$ is maximized. However, to evaluate this you need some information about $E[X]$, so to make things concrete set $X_m = Y_1 + \cdots + Y_m$, a sequence of Bernoulli trials where $\Pr[Y_i = 1] = p$ and $\Pr[Y_i = 0] = (1-p)$. The expectation $E[X_m]$ is pm, and it's reasonable to try to bound $\Pr[X_m \geq xm] \leq e^{-(txm - \ln E[e^{tX_m}])}$. You can use the fact that, for independent random variables, the expectation of the product is the product of the expectations:

$$\ln E[e^{tX_m}] = \ln E[e^{tY_1 + \cdots + tY_m}] = \ln E[e^{tY_1} \cdots e^{tY_m}] = \ln E[e^{tY_i}]^m = m \ln E[e^{tY_i}].$$

Expectation is a weighted sum, so $E[e^{tY_i}] = pe^t + 1 - p$, and $\ln E[e^{tX_m}] = m \ln(pe^t + 1 - p)$. To make your bound as small as possible, you'd like to maximize $tx - \ln(pe^t + 1 - p)$, so you can take derivatives with respect to t (that's why that spurious t was stuck in there in the first place):

$$\frac{d}{dt}(tx - \ln(pe^t + 1 - p)) = \left(x - \frac{pe^t}{pe^t + 1 - p}\right) \qquad \frac{d^2}{dt^2}(tx - \ln(pe^t + 1 - p)) = \frac{pe^t(1-p)}{(pe^t + 1 - p)^2}.$$

The second derivative is positive for all $p \in (0,1)$, so the critical point where $t = \ln([x(1-p)]/[(1-x)p])$ is a maximum. Plug in this optimal value of t so $e^{-m(tx - \ln(pe^t + 1 - p))}$ will be as small as possible, and you have the following bound for $\Pr[X_m \geq xm]$:

$$\Pr[X_m \geq xm] \leq e^{-m(x \ln(x/p) + (1-x)\ln([1-x]/[1-p])}.$$

This is the fundamental bound,[2] but exponent looks pretty daunting — although it is plausible that you'll know something about the ratio x/p, it is not so clear that you'll have easy access to $(1-x)$ and $(1-p)$. Thankfully, it turns out that this bound can be be approximated with a polynomial expression in the exponent. To verify this, fix p and let $H(x,p) = x \ln(x/p) + (1-x)\ln([1-x]/[1-p])$, and show that this is always greater than the polynomial $2(x-p)^2$ by taking derivatives:

$$\frac{d}{dx}\left(H(x,p) - 2(x-p)^2\right) = \ln\left(\frac{x(1-p)}{p(1-x)}\right) - 4(x-p) \qquad \frac{d^2}{dx^2}\left(H(x,p) - 2(x-p)^2\right) = \frac{1}{x(1-x)} - 4.$$

The second derivative is always non-negative, and the critical point corresponding to the first derivative is a minimum. The critical value occurs when $x = p$, and plugging this in gives $H(x,p) - 2(x-p)^2 = 0$,

[1] I (mostly) follow the approach used in the notes on Chernoff and Hoeffding bounds at `www.math.rutgers.edu/courses/591/chern.ps`. See also *The probabilistic Method*, Alon, Spencer, and Erdös , Wiley 1992.

[2] The exponent is $-m$ times the relative entropy of $x, (1-x)$ and $p, (1-p)$, also known as Kullback-Liebler distance $D(x||p)$.

verifying that $H(x,p)$ is never less than $2(x-p)^2$. Assume that $x \geq p$ and let $\epsilon = x - p$, and this gives you the additive Chernoff, or Hoeffding, bound:

$$\Pr[X_m \geq (p+\epsilon)m] \leq e^{-2m\epsilon^2}.$$

Since X_m is the sum of the "successes" (the number of X_i that are 1), $m - X_m$ sums the "failures" (the number of X_i that are 0), and these occur with probability $(1-p)$. The expectation $E[m - X_m]$ is $(1-p)m$, so symmetry gets you another Chernoff bound for (almost) free, by substituting $m - X_m$ for X_m and $1-p$ for p in the first Chernoff bound:

$$\Pr[X_m \leq (p-\epsilon)m] = \Pr[m - X_m \geq (1-(p-\epsilon))m] = \Pr[m - X_m \geq (1-p+\epsilon)m] \leq e^{-2m\epsilon^2}.$$

Multiplicative Chernoff bounds

Another (multiplicative) flavour of Chernoff bound can be found by deriving a slightly looser fundamental bound. These bounds are useful for answering questions about the probability that X_m exceeds or falls short of its expected value by some factor.[3] You can start with bounding the probability that X_m doesn't exceed some multiple (less than 1) of the expected value $E[X_m]$ $(= pm)$, in other words $\Pr[X_m \leq (1-\epsilon)pm]$ (for $0 \leq \epsilon \leq 1$). Reciprocals of exponentials reverse inequalities, so (using the same idea as in the previous section), you find

$$\Pr[X_m \leq (1-\delta)pm] = \Pr[e^{-tX_m} \geq e^{-t(1-\delta)pm}] \leq \frac{E[e^{-tX_m}]}{e^{-t(1-\delta)pm}}.$$

Once again, you can use the fact that the expected value of a product of independent variables is the same as product of their expected values, or in symbols $E[e^{-tX_m}] = E[e^{-tY_1} \cdots e^{-tY_m}] = E[e^{-tY_1}] \cdots E[e^{-tY_m}]$ $= E[e^{-tY_1}]^m$. The expected value $E[e^{-tY_1}]$ is the weighted sum $pe^{-t} + 1 - p$, which can be re-written as $1 - p(1-e^{-t})$, which allows you to use the inequality $1 - x \leq e^{-x}$, to simplify the result so far (note that this inequality loosens the bound a bit):

$$\Pr[X_m \leq (1-\delta)pm] \leq \frac{E[e^{-tX_m}]}{e^{-t(1-\delta)pm}} \leq \frac{e^{pm(e^{-t}-1)}}{e^{-t(1-\delta)pm}} = e^{pm(e^{-t}+t-t\delta-1)}.$$

To make this bound as tight (small) as possible means minimizing $f(t) = e^{-t} + t - t\delta - 1$. The first derivative, $df(t)/dt = -e^{-t} + 1 - \delta$, and the second derivative, $d^2f(t)/dt^2 = e^{-t}$ is positive, so the critical point, when $t = \ln(1/(1-\delta))$ is a minimum. Substitute this back into our current estimate of the probability to find:

$$\Pr[X_m \leq (1-\delta)pm] \leq e^{pm((1-\delta)+(1-\delta)\ln(1/(1-\delta))-1)} = \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{pm}.$$

The denominator of the bound is (again) daunting, since you wouldn't expect to manipulate $(1-\delta)^{(1-\delta)}$ easily, but (again) there is an approximation using a polynomial in δ in the exponent. Since δ has absolute value at most 1, you can expand the Taylor series around 1 for $\ln(1-\delta) = -\delta - \delta^2/2 - \delta^3/3 - \cdots$, and now you can multiply this by $(1-\delta)$ to get $(1-\delta)\ln(1-\delta) = -\delta + \delta^2/2 + (\text{positive terms}) \geq -\delta + \delta^2/2$, which gives a simpler bound (but again loses some precision), for $0 \leq \delta < 1$:

$$\Pr[X_m \leq (1-\delta)pm] \leq \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{pm} = \left(\frac{e^{-\delta}}{(e^{(1-\delta)\ln(1-\delta)}}\right)^{pm} \leq \left(\frac{e^{-\delta}}{(e^{-\delta+\delta^2/2}}\right)^{pm} = e^{-pm\delta^2/2}.$$

Things are a little messier for the upper tail, $\Pr[X_m \geq (1+\delta)pm]$, for $0 \leq \delta \leq 1$. Proceed as above (using t instead of $-t$), and you'll get an eerily familiar bound:

$$\Pr[X_m \geq (1+\delta)pm] \leq \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{pm}.$$

[3] I (mostly) follow the approach in John Canny's CS174 lecture notes on "Chernoff Bounds" at `www.cs.berkeley.edu/~jfc/cs174/lecs/lec10`. See also *A guided tour of chernoff bounds*, Torben Hagerup and Christine Rüb, Information Processing Letters, 33:305–308, 1990.

Again the denominator looks daunting, and the Taylor series expansion this time isn't quite so kind,[4] giving you $(1+\delta)\ln(1+\delta) = \delta + \delta^2/2 - \delta^3/6 + \delta^4/12 - \delta^5/20 + \cdots \geq \delta + \delta^2/2 - \delta^3/6 + (\delta^4/12 - \delta^5/20) + \cdots + (\delta^{2n}/((2n-1)2n) - \delta^{2n+1}/((2n)(2n+1))) \geq \delta + \delta^2/3$.and (substituting this back into the bound), you get the multiplicative Chernoff bound for the upper tail, for $0 \leq \delta \leq 1$:

$$\Pr[X_m \geq (1+\delta)pm] \leq \left(\frac{e^\delta}{(1+\delta)^{(1+\delta)}}\right)^{pm} \leq e^{-pm\delta^2/3}.$$

Comparing the bounds

You derived additive and multiplicative bounds using independent routes, and it's reasonable to wonder which bound is tighter (although in practice convenience often outweighs finding the absolutely tightest bound). Since there are a couple of steps in deriving the multiplicative bounds where looser estimates are made, it's not too surprising that the additive bounds are tighter if p is large enough. However, if you recall that the polynomial factor in the additive bound is $2(x-p)^2$, and if $x = (1 \pm \delta)p$, the corresponding factor in the multiplicative factor is $(x-p)^2/2p$ (if $x \leq p$) or $(x-p)^2/3p$ (if $x \geq p$), then it is clear that the multiplicative bounds become tighter when $p = 1/4$ (if $x \leq p$) and $p = 1/6$ (if $x \geq p$). Below are some figures demonstrating the relative strengths of the bounds, to finish up this look at Chernoff bounds.

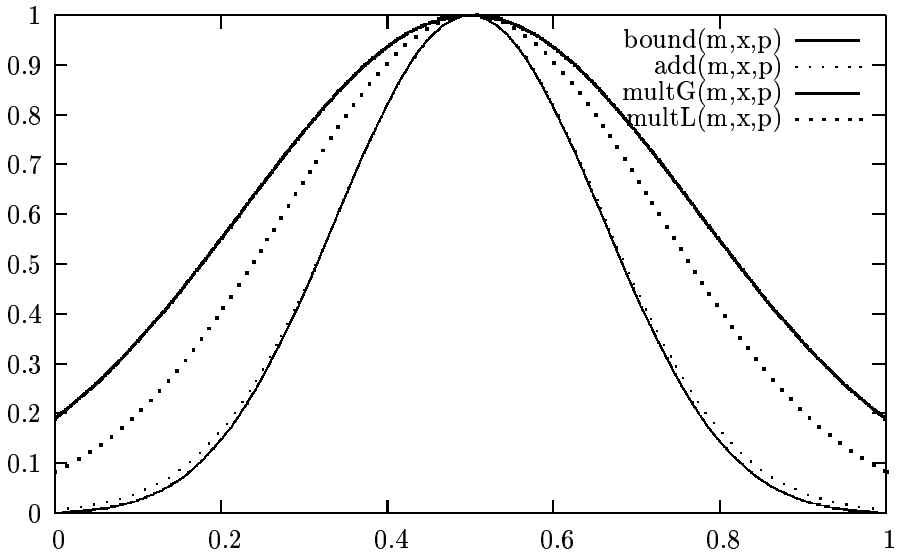

Figure 1: Here is a comparison of the bounds when $p = 0.5$ and $m = 10$. The first bound derived is "bound," the additive Chernoff bound is "add,", the multiplicative bound when $x \geq p$ is "multG," and the multiplicative bound when $x \leq p$ is "multL."

[4]This approximation is my own (so far as I know), so you might want to double-check the details.

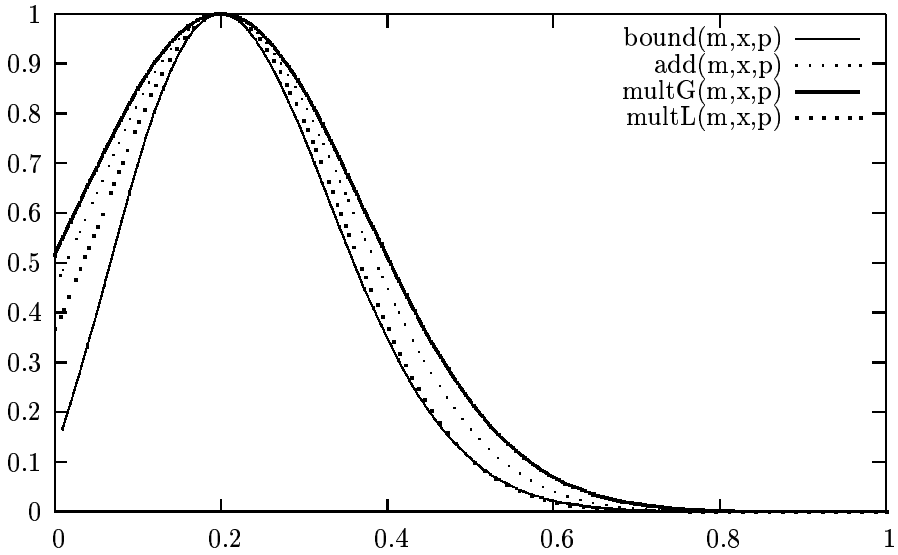

Figure 2: Here's the situation when $p = 0.2$, so you'd expect "add" to a little worse than "multL."