Chernoff Bounds

May 1, 2006

The Chernoff bounds family of inequalities upper bound the probability that the fraction of time we observe an event in a random sample differs significantly from the true probability of the event. There are many variants. We will derive a few simple bounds for the binary case in which there are only two possible events. For example, one such case is the outcome of a coin flip.

Think of events as the values of a sequence of independent random variables (irv) $X_1, X_2, \ldots, X_m$ such that (st)

$$X_i = \begin{cases} 1, & \text{with probability } p \\ 0, & \text{with probability } 1-p \end{cases}$$

The fraction of m times in which event "1" occurs is equal to $\hat{p} = \frac{1}{m}\sum_{i=1}^{m} X_i$, this is the *empirical* probability of p. Is is easy to show that the *expected value* of $\hat{p}$ is p:

$$E[\hat{p}] = (1 \times p) + (0 \times (1-p)) = p$$

This means that $\hat{p}$ is a *fair* or *unbiased* estimator of p.

Our goal is to show that, with high probability, this estimate ($\hat{p}$) is pretty good. Specifically, we want to upper bound the probability $P[\hat{p} \geq q]$ for any $q > p$.

The main idea of the derivation is to combine Markov's inequality with the fact that the expected value of a product of *independent* random variables $Z_1, \ldots, Z_n$ is equal to the product of the expectations, $E[\prod_i Z_i] = \prod_i E[Z_i]$.

We start by transforming the sum of random variables into a product. To do that we exponentiate the sum $\hat{p} = \frac{1}{m}\sum_{i=1}^{m} X_i$. Let $\lambda > 0$ be a parameter whose value will be chosen later. For any $\lambda > 0$ the function $e^{m\lambda q}$ is monotone increasing in x. Thus

$$P[\hat{p} \geq q] = P\left[e^{m\lambda\hat{p}} \geq e^{m\lambda q}\right]$$

As $e^{m\lambda\hat{p}}$ is always larger than zero, we can use Markov's inequality to get

$$E\left[e^{m\lambda\hat{p}}\right] \geq e^{m\lambda q} P\left[e^{m\lambda\hat{p}} \geq e^{m\lambda q}\right],$$

and thus

$$P[\hat{p} \geq q] \leq e^{-m\lambda q} E\left[e^{m\lambda\hat{p}}\right]. \tag{1}$$

Plugging the definition of $\hat{p}$ into the expression of the expectation, we get that

$$E\left[e^{m\lambda\hat{p}}\right] = E\left[e^{\lambda\sum_{i=1}^{m} x_i}\right] = E\left[\prod_{i=1}^{m} e^{\lambda x_i}\right].$$

As x_i are independent of each other, so are $e^{\lambda x_i}$. We are now ready to use the fact that the product of independent random variables is equal to the product of the expected values.

$$E\left[e^{m\lambda\hat{p}}\right] = \prod_{i=1}^{m} E\left[e^{\lambda x_i}\right] = \prod_{i=1}^{m}\left[pe^{\lambda} + (1-p)e^{0}\right] = \left[pe^{\lambda} + (1-p)\right]^{m}$$

We have thus expressed $E\left[e^{m\lambda\hat{p}}\right]$ as a product of expectations, each of which is expressed as a sum over the two possible states of x_i. As all the x_i are identically distributed, the expressions are identical for all m random variables. Combining this expression with Equation (1) we get

$$\begin{aligned}
P\left[\hat{p} > q\right] &\leq e^{-m\lambda q}[pe^{\lambda} + (1-p)]^{m} \\
\text{apply ln to both sides} & \\
\ln P\left[\hat{p} \geq q\right] &\leq \ln \frac{[pe^{\lambda} + (1-p)]^{m}}{e^{m\lambda q}} \\
&= m\ln[pe^{\lambda} + (1-p)] - m\lambda q \\
\text{exponentiate both sides} & \\
P\left[\hat{p} \geq q\right]) &\leq e^{-m\lambda q + m\ln[pe^{\lambda}+(1-p)]} \qquad (2)
\end{aligned}$$

We define $f_\lambda(q,p) \doteq \lambda q - \ln[pe^{\lambda} + (1-p)]$ which simplifies Equation (2) to

$$P\left[\hat{p} > q\right] \leq e^{-mf_\lambda(q,p)} \qquad (3)$$

Note that $f_\lambda(q,p)$ does not depend on m, the number of samples; this means that if we can find λ so that that $f_\lambda(q,p) > 0$ then we get an upper bound on the probability that decreases exponentially as m increases. To get the tightest possible bound, we wish to find the value of λ that maximizes $f_\lambda(q,p)$. We do that by requiring that the partial derivative of $f_\lambda(q,p)$ with respect to λ is equal zero:

$$\begin{aligned}
f_\lambda(q,p) &= \lambda q - \ln[pe^{\lambda} + (1-p)] \\
0 = \frac{\partial f}{\partial \lambda} &= q - \frac{pe^{\lambda}}{pe^{\lambda} + (1-p)} \\
e^{\lambda}(p - qp) &= q - qp \\
e^{\lambda} &= \frac{q(1-p)}{p(1-q)} \\
\lambda &= \ln\frac{q(1-p)}{p(1-q)} \qquad (4)
\end{aligned}$$

Plugging equation (4) in the definition of $f_\lambda(q,p)$ gives us (after some ugly algebraic manipulation) that

$$f_\lambda(q,p) = q \ln \frac{q}{p} + (1-q) \ln \frac{1-q}{1-p} \tag{5}$$

This expression we derived for $\max_\lambda f_\lambda(q,p)$ is the well known Kullback-Leibler (KL) divergence between the distributions $(p, 1-p)$ and $(q, 1-q)$ which we denote by denoted $D_{\mathrm{KL}}(q||p)$. It is easy to show that $D_{\mathrm{KL}}(q||p) \geq 0$ which equality iff $p = q$. Plugging $f_\lambda(q,p) = D_{\mathrm{KL}}(q||p)$ into Equation 3 we get

$$P\left[\hat{p} > q\right] \leq e^{-mD_{\mathrm{KL}}(q||p)} \tag{6}$$

which is provably the best bound of this form on the binomial tail.

Notice that while we have stated the bounds only for the case $\hat{p} > q > p$ we can consider the case $\hat{p} < q < p$ by exchanging p and $1-p$, q and $1-q$, and $\hat{p}$ and $1-\hat{p}$.

However, it is often easier to have looser bounds in which the expression inside the exponent are simper and easier to manipulate. These bounds can be derived from various bounds on the KL-divergence. Three important examples of such bounds are

1. $D_{\mathrm{KL}}(q||p) \geq 2(p-q)^2$ implies that

   $$P\left[\hat{p} > q\right] \leq e^{-2m(p-q)^2}$$

   . Taking the union of the two cases: $\hat{p} > q > p$ and $\hat{p} < q < p$ we get that

   $$P\left[|\hat{p} - p| > \epsilon\right] \leq 2e^{-2m\epsilon^2}$$

   This is the Hoeffding bound.

2. if $q \geq p$,

   $$D_{\mathrm{KL}}(q||p) \geq \frac{1}{3}\frac{(p-q)^2}{p} \Rightarrow P\left[\hat{p} > q\right] \leq e^{-\frac{1}{3}m\frac{(p-q)^2}{p}}$$

3. if $q \leq p$,

   $$D_{\mathrm{KL}}(q||p) \geq \frac{1}{2}\frac{(p-q)^2}{p} \Rightarrow P\left[\hat{p} < q\right] \leq e^{-\frac{1}{2}m\frac{(p-q)^2}{p}}$$

Hoeffding's bound is, in general, the most useful. However if p is close to zero then we can derive better bounds from inequalities (2) and (3). For example, suppose that $(p-q) = \epsilon$, then Hoeffding's bound gives $e^{-2m\epsilon^2}$. However, if we assume $p = \epsilon$ and $q = 2\epsilon$ then bound (2) gives $e^{-(1/3)m\epsilon}$ The general rule of thumb we can derive from this is if p is close to 0 or to 1 then the number of examples to achieve accuracy ϵ is $O(1/\epsilon)$. While, if p is far from both 0 or 1 then the number of examples that are needed to achieve accuracy ϵ is $O(1/\epsilon^2)$.

At this point I suggest you fire up matlab, and write a little program that compares the bound to the actual value of the tail. You can use `LargeVsSmall.m` as a starting point.

A few suggestions:

1. Calculating the tail exactly can be done by summing over the appropriate range of entries of the vector output by `binopdf`.

2. Vary the sample size m to see how the exact tail varies, compare that to the way the bound varies.

3. As both the bound and the tail yield very small numbers, it is useful to use `semilogy` instead of `plot` to plot the bound (or exact value) as a function of m.

4. The bound has to always be above the exact value, if not, then you have a bug in your code.

5. What is the ratio between the bound and the exact value, how does this ratio change as a function of m?

6. Cosider $p = 0.7, q = 0.8$ and plot the value you get for the different bounds, which forms of the bounds are tighter?

7. Consider $p = 10/m$ and $q = 20/m$, which bounds are tighter now?

The Kullback-Liebler divergence generalizes to multinomial distributions and is defined as:

$$D_{\mathrm{KL}}(q||p) = \sum_i q_i \log \frac{q_i}{p_i}$$

The exponential bound on the tail generalized too, in this form it is called Sanov's bound. A recommended place to read about Sanov's theorem is section 12.4 of the book "Elements of Information Theory" by Cover and Thomas.

1 Implications of the strong law of large numbers

Chernoff's bound and it's variants are one of the manifestation of the "law of large numbers" which, simply put, states that a system consisting of many independent elements behaves, with very high probability, as would be predicted using the expected behaviour of any one of the elements.

Sanov's bound is one of the foundations of Shannon's information theory. One of it's incarnations is the so-called "Asymptotic Equipartition Property" (AEP). The AEP property for IID binomial variables, also called "memoryless sources" states that the probability of sequences in which the fraction of the different letters is significantly different than their probabilities is extremely small. A little more precisely, it states that the probability of the set of all sequences for which the KL-divergence between the empirical distirbution and the true probability is larger than ϵ decreases exponentially as with the length of the sequence.

This result has profound implication on the optimal way for compressing sequences that are generated by memoryless sources. A subject which we will come back to later in the course.

Another important manifestation of the law of large numbers is in statistical mechanics, which studies the relationship between the microscopic world of atoms and the macroscopic world which is the world at the human scale.

The first and simplest example of statistical mechanics is the behaviour of an ideal gas. An ideal gas is a state of matter in which the individual atoms fly through space without interacting with each other, their only physical interaction is with the walls of the chamber in which they are placed. The number of atoms in any human-scale chamber is vast, and is measured in units called "Avogadro numbers" which is equal approximately 6×10^{23} and is the number of hydrogen atoms in one gram of hydrogen. As a result of this fact, the macroscopic properties of an ideal gas can be derived using the law of large numbers by applying it to the sum of the effect of many independent particles.