Notes on Bayesian Estimators for Bernoulli Distributions

K.S. Ng

August 13, 2010

1 Basic Definitions

Definition 1. The beta function $B(x, y)$ and gamma function $\Gamma(x)$ are defined as follows:

$$B(x, y) := \int_0^1 \theta^{x-1}(1-\theta)^{y-1} d\theta \quad \text{for } x > 0, y > 0 \tag{1}$$

$$\Gamma(x) := \int_0^\infty \theta^{x-1} e^{-\theta} d\theta. \tag{2}$$

Two useful identities associated with the beta and gamma functions that we will need later are

$$\Gamma(x+1) = x\Gamma(x) \tag{3}$$

$$B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}. \tag{4}$$

These results can be found in most textbooks; see e.g. Problem 1.2.6-41 in [Knu97].

Definition 2. The beta(x,y) distribution has the following probability density function:

$$b(\theta; x, y) := \frac{\theta^{x-1}(1-\theta)^{y-1}}{\int_0^1 \theta^{x-1}(1-\theta)^{y-1} d\theta} = \frac{1}{B(x, y)} \theta^{x-1}(1-\theta)^{y-1}.$$

The beta distribution captures several interesting special cases: beta(1,1) gives a uniform distribution over $\theta \in [0, 1]$; beta(0.5,0.5) gives a distribution favouring θ's close to 0 or 1 (i.e. a bowl shape distribution over $[0, 1]$).

2 Bayesian Estimators

Using the beta distribution as a prior on the parameter of an unknown Bernoulli distribution, we have the following formula for the probability of a bit string $x_{1:n} \in \{0,1\}^n$ with n_0 zeros and $n_1 = n - n_0$ ones.

$$\begin{aligned}
\Pr(x_{1:n} \,|\, b(\theta;x,y)) &= \int_0^1 \theta^{n_1}(1-\theta)^{n_0} b(\theta;x,y) d\theta && (5)\\
&= \frac{1}{B(x,y)} \int_0^1 \theta^{n_1+x-1}(1-\theta)^{n_0+y-1} d\theta && (6)\\
&= \frac{B(n_1+x, n_0+y)}{B(x,y)} && (7)\\
&= \frac{\Gamma(x+y)}{\Gamma(x)\Gamma(y)} \frac{\Gamma(n_1+x)\Gamma(n_0+y)}{\Gamma(n+x+y)}. && (8)
\end{aligned}$$

The above, in turn, gives us the following expression for the conditional probability $\Pr(x_{n+1} \,|\, x_{1:n}, b(\theta;x,y))$.

$$\begin{aligned}
&\Pr(x_{n+1} = 1 \,|\, x_{1:n}, b(\theta;x,y)) && (9)\\
&= \frac{\Pr(x_{1:n}1 \,|\, b(\theta;x,y))}{\Pr(x_{1:n} \,|\, b(\theta;x,y))} && (10)\\
&= \frac{\Gamma(n_1+x+1)\Gamma(n_0+y)/\Gamma(n+1+x+y)}{\Gamma(n_1+x)\Gamma(n_0+y)/\Gamma(n+x+y)} && (11)\\
&= \frac{(n_1+x)\Gamma(n_1+x)/(n+x+y)\Gamma(n+x+y)}{\Gamma(n_1+x)/\Gamma(n+x+y)} && (12)\\
&= \frac{n_1+x}{n+x+y}. && (13)
\end{aligned}$$

Since $\Pr(x_{1:n} \,|\, b(\theta;x,y))$ is only a function of the number of zeros n_0 and ones n_1 in $x_{1:n}$, or in other words,

$$\Pr(x_{1:n} \,|\, b(\theta;x,y)) = \Pr(0^{n_0}1^{n_1} \,|\, b(\theta;x,y)),$$

we can write $\Pr(n_0, n_1 \,|\, b(\theta;x,y))$ in place of $\Pr(x_{1:n} \,|\, b(\theta;x,y))$. It is easily seen that

$$\begin{aligned}
&\Pr(0,0 \,|\, b(\theta;x,y)) = 1 && (14)\\
&\Pr(n_0+1, n_1 \,|\, b(\theta;x,y)) = \Pr(n_0, n_1 \,|\, b(\theta;x,y)) \frac{n_0+y}{n_0+n_1+x+y} && (15)\\
&\Pr(n_0, n_1+1 \,|\, b(\theta;x,y)) = \Pr(n_0, n_1 \,|\, b(\theta;x,y)) \frac{n_1+x}{n_0+n_1+x+y}. && (16)
\end{aligned}$$

The last equation follows from the fact that

$$\Pr(n_0, n_1 + 1 \mid b(\theta; x, y)) = \Pr(0^{n_0} 1^{n_1} \mid b(\theta; x, y)) \Pr(1 \mid 0^{n_0} 1^{n_1}, b(\theta; x, y)).$$

Similarly for $\Pr(n_0 + 1, n_1 \mid b(\theta; x, y))$.

Special Cases The Krichevsky-Trofimov estimator [KT81, WST95] is obtained for the case of x = y = 0.5:

$$kt(x_{n+1} = 1 \mid x_{1:n}) := \frac{n_1 + 0.5}{n + 1}.$$

The Laplace estimator (see e.g. [Mac03, §3.2]) is obtained for the case of x = y = 1:

$$lp(x_{n+1} = 1 \mid x_{1:n}) := \frac{n_1 + 1}{n + 2}.$$

The Laplace estimator is also the subject matter of Rev Thomas Bayes' famous 1763 paper (see e.g. [GCSR03, §2.1]).

MML Estimators As a side note, the Minimum Message Length [WF87] estimator $\hat{\theta}$ for the parameter of a Bernoulli distribution using the beta(x,y) prior is given by

$$\hat{\theta} = \frac{n_1 + x - 0.5}{n + x + y - 1}.$$

This means the MML estimator with the beta(1,1) prior is the KT estimator, and that for beta(0.5,0.5) is the empirical estimate n_1/n.

3 A Redundancy Bound

We now state a redundancy bound by adapting the argument used in [WST95].

Lemma 1. *For all $x \geq 0, y \geq 0, a + b \geq 1$, we have*

1. *if $x \geq 1/2$ and $x + y \leq 1$, then*

$$\Pr(a, b \mid b(\theta; x, y)) \geq \frac{y}{x + y} \frac{1}{a + b} \left(\frac{a}{a + b}\right)^a \left(\frac{b}{a + b}\right)^b.$$

2. if $y \geq 1/2$ and $x + y \leq 1$, then

$$\Pr(a, b \,|\, b(\theta; x, y)) \geq \frac{x}{x+y} \frac{1}{a+b} \left(\frac{a}{a+b}\right)^a \left(\frac{b}{a+b}\right)^b.$$

Proof. We give a proof of part (1) of the Lemma. The second part is similar. We start by defining the function

$$\Delta(a, b) := \frac{\Pr(a, b \,|\, b(\theta; x, y))}{\frac{1}{a+b}(\frac{a}{a+b})^a(\frac{b}{a+b})^b}.$$

Consider the ratio $\Delta(a, b+1)/\Delta(a, b)$. We have

$$\begin{aligned}
\frac{\Delta(a, b+1)}{\Delta(a, b)} &= \frac{b^b}{(b+1)^{b+1}} \frac{(a+b+1)^{a+b+2}}{(a+b)^{a+b+1}} \frac{\Pr(a, b+1 \,|\, b(\theta; x, y))}{\Pr(a, b \,|\, b(\theta; x, y))} \\
&= \frac{b^b(b+x)}{(b+1)^{b+1}} \left(\frac{a+b+1}{a+b}\right)^{a+b+1} \frac{a+b+1}{a+b+x+y} \\
&\geq \frac{b^b(b+0.5)}{(b+1)^{b+1}} \left(\frac{a+b+1}{a+b}\right)^{a+b+1}.
\end{aligned}$$

Define $f(b)$ and $g(t)$ as follows:

$$f(b) := \frac{b^b(b+0.5)}{(b+1)^{b+1}} \quad \text{and} \quad g(t) := \left(\frac{t+1}{t}\right)^{t+1}.$$

One can show that

1. $\forall b \geq 0$, $f(b) \geq f(b+1)$ and $\lim_{b\to\infty} f(b) = 1/e$.
2. $\forall t \geq 1$, $g(t) \geq g(t+1)$ and $\lim_{t\to\infty} g(t) = e$.

Together, the above implies that if $a + b \geq 1$, then

$$\frac{\Delta(a, b+1)}{\Delta(a, b)} \geq \frac{1}{e} e = 1. \tag{17}$$

Consider now the ratio $\Delta(a+1, 0)/\Delta(a, 0)$. We have

$$\begin{aligned}
\frac{\Delta(a+1, 0)}{\Delta(a, 0)} &= \frac{a+1}{a} \frac{\Pr(a+1, 0 \,|\, b(\theta; x, y))}{\Pr(a, 0 \,|\, b(\theta; x, y))} \\
&= \frac{a+1}{a} \frac{a+y}{a+x+y} \geq 1.
\end{aligned} \tag{18}$$

Combining (17) and (18), we have, if $a + b \geq 1$,

1. case of $a \neq 0$: $\Delta(a, b) \geq \Delta(a, 0) \geq \Delta(1, 0) = y/(x + y)$;
2. case of $a = 0$: $\Delta(a, b) = \Delta(0, b) \geq \Delta(0, 1) = x/(x + y) \geq y/(x + y)$.

thus giving our bound. $\square$

We can now state the redundancy result. In the following, log is base 2. For all $\theta \in [0, 1]$, if $x \geq 1/2$, $x + y \leq 1$ and $a + b \geq 1$, we have

$$\begin{aligned}
\log \frac{(1-\theta)^a \theta^b}{\Pr(a, b \mid b(\theta; x, y))} &\leq \log \frac{(1-\theta)^a \theta^b}{\frac{y}{x+y} \frac{1}{a+b} (\frac{a}{a+b})^a (\frac{b}{a+b})^b} \\
&= \log \frac{x+y}{y} + \log(a+b) + \log \frac{(1-\theta)^a \theta^b}{(\frac{a}{a+b})^a (\frac{b}{a+b})^b} \\
&\leq \log \frac{x+y}{y} + \log(a+b). \qquad (19)
\end{aligned}$$

The last step follows from the fact that given a and b, the maximum likelihood estimate for θ is $b/(a + b)$. We have a similar result for the case when $y \geq 1/2$ and $x + y \leq 1$.

In [WST95], the authors gave, using essentially the same proof as above, the following tighter bound for the KT estimator

$$\log \frac{(1-\theta)^a \theta^b}{\Pr(a, b \mid b(\theta; 0.5, 0.5))} \leq \frac{1}{2} \log(a+b) + 1$$

and stated that it is "impossible to prove such a uniform bound for the Laplace estimator." I find that comment slightly puzzling. The conditions of (19) is clearly violated for the case of $x = y = 1$; but numerical experiments appear to suggest that bound

$$\log \frac{(1-\theta)^a \theta^b}{\Pr(a, b \mid b(\theta; 1, 1))} \leq \log(a+b) + 1$$

holds for all $\theta \in [0, 1]$.

References

[GCSR03] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis.* Chapman & Hall, 2nd edition, 2003.

[Knu97] Donald E. Knuth. *The Art of Computer Programming, Volume 1, Fundamental Algorithms.* Addison-Wesley, 3rd edition, 1997.

[KT81] Raphail E. Krichevsky and Victor K. Trofimov. The performance of universal coding. *IEEE Transactions on Information Theory*, IT-27(2):199–207, 1981.

[Mac03] David J.C. MacKay. *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press, 2003.

[WF87] C.S. Wallace and P.R. Freeman. Estimation and inference by compact coding. *Journal of the Royal Statistical Society*, 49(3):240–265, 1987.

[WST95] Frans M.J. Willems, Yuri M. Shtarkov, and Tjalling J. Tjalkens. The context tree weighting method: Basic properties. *IEEE Transactions on Information Theory*, 41:653–664, 1995.