Bayesian inference of a uniform distribution

Thomas P. Minka

January 25, 2001

Abstract

This note derives the posterior, the evidence, and the predictive density for a uniform distribution, given a conjugate parameter prior. These provide various Bayesian answers to the "taxicab" problem: viewing a city from the train, you see a taxi numbered x. Assuming taxis are consecutively numbered, how many taxis are in the city?

1 Introduction

The uniform distribution from $0..a$ is

$$p(x|a) \sim \mathcal{U}(0,a) = \begin{cases} 1/a & \text{if } 0 \leq x \leq a \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Given data, the maximum-likelihood estimate for a is precisely the maximal point, which is unsuitable for predicting future data. This paper derives an alternative estimate based on Bayesian theory which has better properties. The result is similar to estimating the variance parameter of a Gaussian: the Bayesian estimate is slightly bigger than the maximum likelihood estimate, to better explain future observations. Tenenbaum (1998) showed that humans exhibit similar behavior in concept learning.

For the Bayesian analysis, we use a conjugate prior. The conjugate prior to the uniform distribution is the Pareto distribution (figure 1):

$$p(a) \sim Pa(b,K) = \begin{cases} \frac{Kb^K}{a^{K+1}} & \text{if } a \geq b \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

This density asserts that a must be greater than some constant b but not too much greater, where K controls what is "too much." The prior becomes noninformative as we take $K \to 0$ and $b \to 0$. The mean of this distribution is

$$E[a] = \frac{K}{K-1}b \tag{3}$$

Given a Pareto prior, the joint distribution of a and a dataset $D = \{x_1, ..., x_N\}$ is

$$p(D,a) = \frac{Kb^K}{a^{N+K+1}} \qquad a \geq \max(D) \tag{4}$$

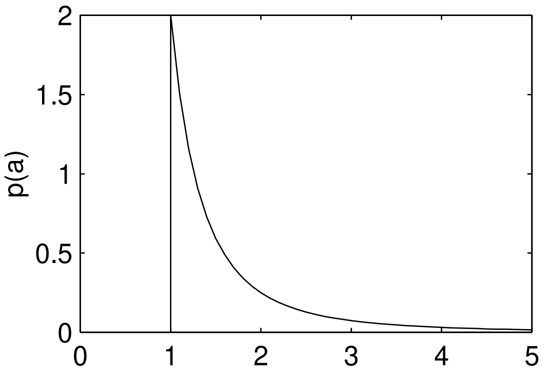

Figure 1: The $Pa(1,2)$ density function.

Let m denote the maximum value in D. Then the evidence is

$$
\begin{aligned}
p(D) &= \int_m^\infty \frac{Kb^K}{a^{N+K+1}} \mathrm{d}a & (5)\\
&= \begin{cases} \frac{K}{(N+K)b^N} & \text{if } m \le b \\ \frac{Kb^K}{(N+K)m^{N+K}} & \text{if } m > b \end{cases} & (6)
\end{aligned}
$$

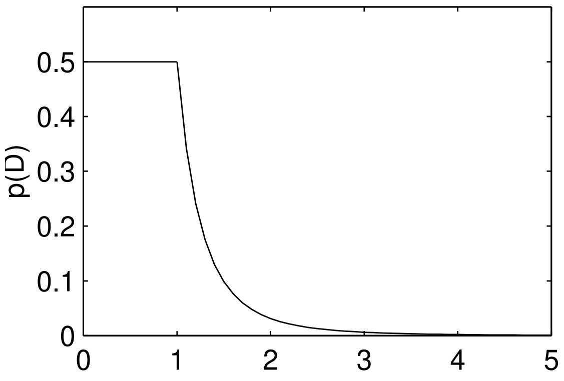

Figure 2: The evidence function for $N = 2$ with a $Pa(1,2)$ prior.

This function is the probability that all N samples came from the same uniform distribution. The posterior for a is

$$
\begin{aligned}
p(a|D) &= p(D,a)/p(D) \sim Pa(c, N+K) & (7)\\
\text{where } c &= \max(m,b) & (8)
\end{aligned}
$$

As we get more samples, we become more sure that a is not much greater than c.

This analysis provides us with various answers to the "taxicab problem," where we must provide an estimate for a given D but no other information. Ignoring the fact that these are taxicabs, we can use a noninformative prior to get $p(a|D) \sim Pa(m, N)$. The mode of the posterior is m, however this always underestimates the true value. The mean of the posterior, which minimizes the expected squared error, is $\frac{Nm}{N-1}$, which doesn't exist for $N = 1$. The median of the posterior, which minimizes the expected absolute error, is the point y such that $\int_y^\infty p(a|D)\mathrm{d}a = 1/2$, which happens at $y = 2^{1/N}m$. The density of m is given by

$$\begin{aligned}
Pr(\max \leq m|a) &= Pr(\text{all } N \text{ samples} \leq m|a) && (9)\\
&= \left(\frac{m}{a}\right)^N && (10)\\
p(m|a) &= \frac{\mathrm{d}}{\mathrm{d}m} Pr(\max \leq m|a) && (11)\\
&= \frac{N}{a}\left(\frac{m}{a}\right)^{N-1} && (12)\\
E[m] &= \int_0^a N\left(\frac{m}{a}\right)^N \mathrm{d}m && (13)\\
&= \frac{Na}{N+1} && (14)
\end{aligned}$$

so an unbiased estimate of a is $\frac{N+1}{N}m$. For $N = 1$, the median and unbiased estimator agree on $\hat{a} = 2m$. With a different prior, e.g. with $K = 1$, we get different answers. The mode is still m but the posterior mean is $\frac{N+K}{N+K-1}m = \frac{N+1}{N}m$, matching the unbiased estimator. The posterior median is $2^{1/(N+K)}m$.

However, if we want to know about what cab numbers we might see next, then we should avoid estimating a and instead use the predictive density, which is

$$p(x|D) = \int_a p(x|a)p(a|D) = \begin{cases} \frac{N+K}{(N+K+1)c} & \text{if } x \leq c \\ \frac{(N+K)c^{N+K}}{(N+K+1)x^{N+K+1}} & \text{if } x > c \end{cases} \tag{15}$$

It is the same as the evidence for one sample when the prior is $Pa(c, N + K)$. For the taxicab problem with $N = 1$ we have

$$p(x|m) = \begin{cases} \frac{1}{2m} \text{ if } x \leq m \\ \frac{m}{2x^2} \text{ if } x > m \end{cases} \tag{16}$$

The density has the curious property that both x and the transformation $y = m^2/x$ have the same density. 10

References

[1] J. Tenenbaum. Bayesian modeling of human concept learning. Advances in Neural Information Processing Systems (NIPS) 11, 1998. http://sucia.stanford.edu/~jbt/.