The Bayesian Approach to Parameter Learning

Peter Lucas

April 4, 2017

1 Introduction

Parameters of probability distributions are frequently learnt using a Bayesian method, where an a priori distribution is updated based on newly observed data. Here we shall investigate the simple case where a single discrete random variable is updated according to this scheme under the assumption that the variable is distributed according to a binomial distribution. This simply means that in an experimental trial we measure the number of successes and failures, and based these results compute the probability of success (failure).

2 Background

The entire theory is based on two probability distributions, the binomial and the beta distribution. These are linked to each other by means of an augmented Bayesian network, to be discussed below.

2.1 Binomial distribution

Let Y be a variable that measures the number of successes of n experimental (Bernoulli) trials, each with two possible outcomes: success and failure. Then it holds that

$$P(Y = k) = f(k)$$

where k is the number of successes and f the probability mass function. This function is defined as follows:

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k},$$

where $p \in [0, 1]$ is the probability of success in a trial. This formula can be easily understood as after n trials the probability of the observed result, k successes and $(n-k)$ failures, is equal to $p^k(1-p)^{n-k}$, where, of course, $(1-p)$ is the probability of failure, as a result is always success or failure.

However, when we say that a binary random variable X is distributed according to a binomial distribution, we mean that by *sampling* from the distribution we obtain a probability of success, which will be

$$P(X = yes) = p.$$

We write $X \sim \text{bin}(p)$, i.e., p is the parameter of the distribution.

2.2 Beta distribution

The beta function is of interest as it has the property that if a prior probability distribution is a beta distribution and updated by sampling from a binomial distribution, the posterior distribution is also a beta distribution. It is said that the beta distribution is the *conjugate prior* of the binomial distribution. As expected, the beta distribution looks very similar to the binomial distribution, but, different from the binomial distribution, it is a *continuous* distribution. That it should be continuous follows from the fact that as a prior it acts as a probability distribution of the probabilistic parameter p, which is continuous. It is defined in terms of a *gamma function*, which is a generalisation of the well-known factorial:

$$\Gamma(x) = \int_0^\infty e^{-t} t^{x-1} \, \mathrm{d}t,$$

$x \in \mathbb{R}^+$. For this function it holds that

$$\Gamma(x+1) = x\Gamma(x).$$

Note that $\Gamma(1) = 1$. For $x \in \mathbb{N}_0$, this property yields, in fact, the factorial function, i.e., $\Gamma(n+1) = n!$ for $n \in \mathbb{N}_0$.

Using the definition of the gamma function allows us to define the beta distribution:

$$f(p) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1},$$

where f is a probability density function, $a, b \in \mathbb{R}$ are the parameters of the distribution, and $p \in [0,1]$. If a random variable X is distributed according to a beta function with parameters a and b we write: $X \sim B(a,b)$.

To make a distinction between the random variable P and the probability distribution P we shall write θ for p and Θ for the random variable P, otherwise the notation becomes confusing.

It appears to holds for Θ that the expectation or mean of Θ is equal to:

$$E(\Theta) = \frac{a}{a+b}.$$

The proof goes as follows:

$$\begin{aligned} E(\Theta) &= \int_0^1 \theta f(\theta) \, \mathrm{d}\theta \\ &= \int_0^1 \theta \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1} \, \mathrm{d}\theta \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \theta\theta^{a-1}(1-\theta)^{b-1} \, \mathrm{d}\theta \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} \\ &= \frac{a}{a+b} \end{aligned} \tag{1}$$

Here, we make use of the fact that

$$\int_0^1 \theta^a (1-\theta)^b \, \mathrm{d}\theta = \frac{\Gamma(a+1)\Gamma(b+1)}{\Gamma(a+b+2)}$$

Figure 1: Augmented Bayesian network.

Remark: recall that for a binomial distribution we have that the mean, or expectation, is equal to $E(X) = p$, with p the success probability. If $a + b = n$ is the total number of observed cases, a the number of successes, and b the number of failures, then we see that the two approaches coincide.

2.3 Augmented Bayesian networks

If we assume that there is uncertainty with respect to the parameter θ of a binomially distributed variable X we can write

$$P(X = yes \mid \theta) = \theta,$$

i.e., the probability of success is based on the parameter θ. Uncertainty with respect to this parameter can now be expressed by defining a probability distribution for this θ, which introduces another random variable Θ. We can use a so-called *augmented Bayesian network* to represent the relationship between X and Θ, as shown in Figure 2.3. Thus, this figure shows that

$$f(X, \theta) = P(X \mid \theta) f(\theta)$$

we have to use a mass and density functions f in this expression, as Θ (modelling a probability distribution) is a continuous variables, and thus $P(\Theta = \theta) = 0$.

We now have the following important property:

$$P(X = yes) = E(\Theta),$$

i.e., the prior probability of $X = yes$ is equal to the expectation or mean of Θ. The proof is as follows:

$$\begin{aligned} P(X = yes) &= \int_0^1 P(X = yes \mid \Theta = \theta) f(\theta)\, \mathrm{d}\theta \\ &= \int_0^1 \theta f(\theta)\, \mathrm{d}\theta \\ &= E(\Theta) \\ &= \frac{a}{a+b} \end{aligned}$$

using Equation (1).

3 Bayesian updating

The result that will be derived in a number of steps is that if (binary) data D is available with s the number of successes (yes) and f the number of failures (no) in D, then the data

will affect the probability density on Θ as follows given that we have the prior probability $f'(\theta) = f(\theta; a, b)$ (i.e., $\Theta \sim B(a, b)$):

$$f(\theta \mid D) = f(\theta; a + s, b + f),$$

or equivalently $\Theta \mid D \sim B(a+s, b+f)$. This means that updating the probability distribution Θ is just a matter of updating the parameters a and b of a beta distribution.

Let D be a binary sample (dataset) of size N with θ the underlying success probability and s the number of successes, and f the number of failures. It then holds that the joint probability of the data D is equal to:

$$P(D) = E(\Theta^s(1-\Theta)^f)$$

The proof goes as follows:

$$\begin{aligned} P(D) &= \int_0^1 P(D \mid \theta) f(\theta)\, \mathrm{d}\theta \\ &= \int_0^1 \prod_{k=1}^{N} P(r_k \mid \theta) f(\theta)\, \mathrm{d}\theta \\ &= \int_0^1 \theta^s (1-\theta)^f f(\theta)\, \mathrm{d}\theta \\ &= E(\Theta^s(1-\Theta)^f) \end{aligned} \tag{2}$$

where r_k is a record in the dataset (in this case just one variable with value) and E is the expectation.

The next step is to reformulate $E(\Theta^s(1-\Theta)^f)$ in terms of the gamma function. Let $\Theta \sim B(a, b)$, then for two integers s, f it holds that:

$$E(\Theta^s(1-\Theta)^f) = \frac{\Gamma(a+b)}{\Gamma(a+b+s+f)} \frac{\Gamma(a+s)\Gamma(b+f)}{\Gamma(a)\Gamma(b)}$$

The proof goes as follows:

$$\begin{aligned} E(\Theta^s(1-\Theta)^f) &= \int_0^1 \theta^s (1-\theta)^f f(\theta)\, \mathrm{d}\theta \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \theta^{a+s-1}(1-\theta)^{b+f-1}\, \mathrm{d}\theta \\ &= \frac{\Gamma(a+b)}{\Gamma(a+b+s+f)} \frac{\Gamma(a+s)\Gamma(b+f)}{\Gamma(a)\Gamma(b)} \end{aligned}$$

Again, we make use of the fact that

$$\int_0^1 \theta^a (1-\theta)^b\, \mathrm{d}\theta = \frac{\Gamma(a+1)\Gamma(b+1)}{\Gamma(a+b+2)}$$

Thus, we now know from Equation (2) that

$$P(D) = \frac{\Gamma(a+b)}{\Gamma(a+b+s+f)} \frac{\Gamma(a+s)\Gamma(b+f)}{\Gamma(a)\Gamma(b)}$$

The final step is to consider $f(\theta \mid D)$, which, according to Bayes' theorem, can be written as:

$$f(\theta \mid D) = \frac{P(D \mid \theta) f(\theta)}{P(D)}$$

The only thing we have to do is to fill in the details, which gives us:

$$\begin{aligned} f(\theta \mid D) &= \frac{\prod_{k=1}^{N} P(r_k \mid \theta) f(\theta)}{P(D)} \\ &= \frac{\theta^s (1-\theta)^f f(\theta)}{P(D)} \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\theta^{a+s-1}(1-\theta)^{b+f-1}}{P(D)} \\ &= \frac{\Gamma(a+b+s+f)}{\Gamma(a+s)\Gamma(b+f)} \theta^{a+s-1}(1-\theta)^{b+f-1} \\ &= f(\theta; a+s, b+f) \end{aligned}$$

where $N = s + f$ is the total number of records r_k. Thus, $\Theta \mid D \sim B(a+s, b+f)$, whereas $\Theta \sim B(a, b)$.