The Bayesian Approach to Machine Learning

In the previous chapter, we saw how explicitly adding noise to our model allowed us to obtain more than just point predictions. In particular, we were able to quantify the uncertainty present in our parameter estimates and our subsequent predictions. Once content with the idea that there will be uncertainty in our parameter estimates, it is a small step towards considering our parameters themselves as random variables. Bayesian methods are becoming increasingly important within Machine Learning and we will devote the next two chapters to providing an introduction to an area that many people find challenging. In this chapter, we will cover some of the fundamental ideas of Bayesian statistics through two examples. Unfortunately, the calculations required to perform Bayesian inference are often not analytically tractable. In Chapter 4 we will introduce three approximation methods that are popular in the Machine Learning community.

3.1 A COIN GAME

Imagine you are walking around a fairground and come across a stall where customers are taking part in a coin tossing game. The stall owner tosses a coin ten times for each customer. If the coin lands heads on six or fewer occasions, the customer wins back their £1 stake plus an additional £1. Seven or more and the stall owner keeps their money. The binomial distribution (described in Section 2.3.2) describes the probability of a certain number of successes (heads) in N binary events. The probability of y heads from N tosses where each toss lands heads with probability r is given by

$$P(Y = y) = \binom{N}{y} r^y (1 - r)^{N-y}. \tag{3.1}$$

You assume that the coin is fair and therefore set $r = 0.5$. For $N = 10$ tosses, the probability distribution function can be seen in Figure 3.1, where the bars corresponding to $y \leq 6$ have been shaded. Using Equation 3.1, it is possible to calculate the probability of winning the game, i.e. the probability that Y is less than or equal to 6, $P(Y \leq 6)$:

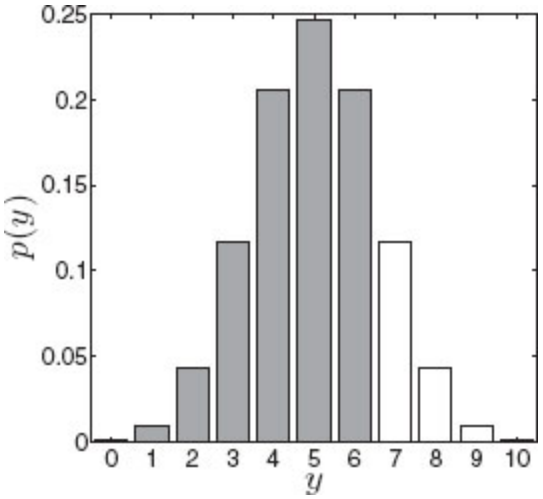

FIGURE 3.1 The binomial density function (Equation 3.1) when $N = 10$ and $r = 0.5$.

$$\begin{aligned} P(Y \leq 6) = 1 - P(Y > 6) &= 1 - [P(Y = 7) + P(Y = 8) + P(Y = 9) + P(Y = 10)] \\ &= 1 - [0.1172 + 0.0439 + 0.0098 + 0.0010] \\ &= 0.8281. \end{aligned}$$

This seems like a pretty good game – you'll double your money with probability 0.8281. It is also possible to compute the expected return from playing the game. The expected value of a function $f(X)$ of a random variable X is computed as (introduced in Section 2.2.8)

$$\mathbf{E}_{P(x)}\left\{f(X)\right\} = \sum_{x} f(x)P(x),$$

where the summation is over all possible values that the random variable can take. Let X be the random variable that takes a value of 1 if we win and a value of 0 if we lose: $P(X = 1) = P(Y \leq 6)$. If we win, $(X = 1)$, we get a return of £2 (our original stake plus an extra £1) so $f(1) = 2$. If we lose, we get a return of nothing so $f(0) = 0$. Hence our expected return is

$$f(1)P(X = 1) + f(0)P(X = 0)2 \times P(Y \leq 6) + 0 \times P(Y > 6) = 1.6562.$$

Given that it costs £1 to play, you win, on average, 1.6562 − 1 or approximately 66p per game. If you played 100 times, you'd expect to walk away with a profit of £65.62.

Given these odds of success, it seems sensible to play. However, whilst waiting you notice that the stall owner looks reasonably wealthy and very few customers seem to be winning. Perhaps the assumptions underlying the calculations are wrong. These assumptions are

1. The number of heads can be modelled as a random variable with a binomial distribution, and the probability of a head on any particular toss is r.
2. The coin is fair – the probability of heads is the same as the probability of tails, $r = 0.5$.

It seems hard to reject the binomial distribution – events are taking place with only two possible outcomes and the tosses do seem to be independent. This leaves r, the probability that the coin lands heads. Our assumption was that the coin was fair – the probability of heads was equal to the probability of tails. Maybe this is not the case? To investigate this, we can treat r as a parameter (like $\mathbf{w}$ and σ^2 in the previous chapter) and fit it to some data.

3.1.1 Counting heads

There are three people in the queue to play. The first one plays and gets the following sequence of heads and tails:

$$\mathbf{H, T, H, H, H, H, H, H, H, H,}$$

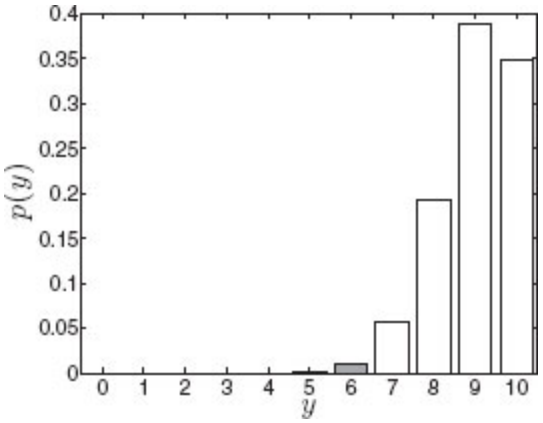

FIGURE 3.2 The binomial density function (Equation 3.1) when $N = 10$ and $r = 0.9$.

nine heads and one tail. It is possible to compute the maximum likelihood value of r as follows. The likelihood is given by the binomial distribution:

$$P(Y = y|r, N) = \binom{N}{9} r^y (1-r)^{N-y}. \tag{3.2}$$

Taking the natural logarithm gives

$$L = \log P(Y = y|r, N) = \log \binom{N}{9} + y \log r + (N - y) \log(1 - r).$$

As in Chapter 2, we can differentiate this expression, equate to zero and solve for the maximum likelihood estimate of the parameter:

$$\begin{aligned} \frac{\partial L}{\partial r} &= \frac{y}{r} - \frac{N-y}{1-r} = 0 \\ y(1-r) &= r(N-y) \\ y &= rN \\ r &= \frac{y}{N}. \end{aligned}$$

Substituting $y = 9$ and $N = 10$ gives $r = 0.9$. The corresponding distribution function is shown in Figure 3.2 and the recalculated probability of winning is $P(Y \leq 6) = 0.0128$. This is much lower than that for $r = 0.5$. The expected return is now

$$2 \times P(Y \leq 6) + 0 \times P(Y > 6) = 0.0256.$$

Given that it costs £1 to play, we expect to make 0.0256 − 1 = −0.9744 per game – a loss of approximately 97*p*. $P(Y \leq 6) = 0.0128$ suggests that only about 1 person in every 100 should win, but this does not seem to be reflected in the number of people who *are* winning. Although the evidence from this run of coin tosses suggests $r = 0.9$, it seems too biased given that several people *have* won.

3.1.2 The Bayesian way

The value of r computed in the previous section was based on just ten tosses. Given the random nature of the coin toss, if we observed several sequences of tosses it is likely that we would get a different r each time. Thought about this way, r feels a bit like a random variable, R. Maybe we can learn something about the distribution of R rather than try and find a particular value. We saw in the previous section that obtaining an exact value by counting is heavily influenced by the particular tosses in the short sequence. No matter how many such sequences we observe there will always be some uncertainty in r – considering it as a random variable with an associated distribution will help us measure and understand this uncertainty.

In particular, defining the random variable Y_N to be the number of heads obtained in N tosses, we would like the distribution of r conditioned on the value of Y_N:

$$p(r|y_N).$$

Given this distribution, it would be possible to compute the expected probability of winning by taking the expectation of $P(Y_{\text{new}} \leq 6|r)$ with respect to $p(r|y_N)$:

$$P(Y_{\text{new}} \leq 6\Big|y_N) = \int P(Y_{\text{new}} \leq 6|r)p(r|y_N)dr,$$

where Y_{new} is a random variable describing the number of heads in a future set of ten tosses.

In Section 2.2.7 we gave a brief introduction to Bayes' rule. Bayes' rule allows us to reverse the conditioning of two (or more) random variables, e.g. compute $p(a|b)$ from $p(b|a)$. Here we're interested in $p(r|y_N)$, which, if we reverse the conditioning, is $p(y_N|r)$ – the probability distribution function over the number of heads in N independent tosses where the probability of a head in a single toss is r. This is the binomial distribution function that we can easily compute for any y_N and r. In our context, Bayes' rule is (see also Equation 2.11)

$$p(r\Big|y_N) = \frac{P(y_N|r)p(r)}{P(y_N)}. \tag{3.3}$$

This equation is going to be very important for us in the following chapters so it is worth spending some time looking at each term in detail.

The likelihood, $P(y_N|r)$ We came across likelihood in Chapter 2. Here it has exactly the same meaning: how likely is it that we would observe our data (in this case, the data is y_N) for a particular value of r (our model)? For our example, this is the binomial distribution. This value will be high if r could have feasibly produced the result y_N and low if the result is very unlikely. For example, Figure 3.3 shows the likelihood $P(y_N|r)$ as a function of r for two different scenarios. In the first, the data consists of ten tosses ($N = 10$) of which six were heads. In the second, there were $N = 100$ tosses, of which 70 were heads.

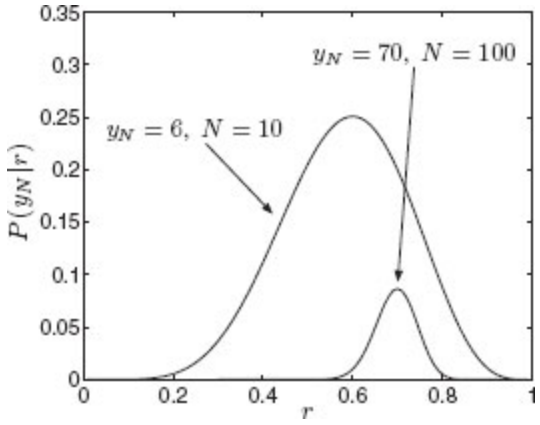

FIGURE 3.3 Examples of the likelihood $p(y_N|r)$ as a function of r for two scenarios.

This plot reveals two important properties of the likelihood. Firstly, it is not a probability density. If it were, the area under both curves would have to equal 1. We can see that this is not the case without working out the area because the two areas are completely different. Secondly, the two examples differ in how much they appear to tell us about r. In the first example, the likelihood has a non-zero value for a large range of possible r values (approximately $0.2 \leq r \leq 0.9$). In the second, this range is greatly reduced (approximately $0.6 \leq r \leq 0.8$). This is very intuitive: in the second example, we have much more data (the results of 100 tosses rather than 10) and so we *should* know more about r.

The prior distribution, $p(r)$ The prior distribution allows us to express any belief we have in the value of r *before* we see any data. To illustrate this, we shall consider the following three examples:

1. We do not know anything about tossing coins or the stall owner.
2. We think the coin (and hence the stall owner) is fair.
3. We think the coin (and hence the stall owner) is biased to give more heads than tails.

We can encode each of these beliefs as different prior distributions. r can take any value between 0 and 1 and therefore it must be modelled as a continuous random variable. Figure 3.4 shows three density functions that might be used to encode our three different prior beliefs.

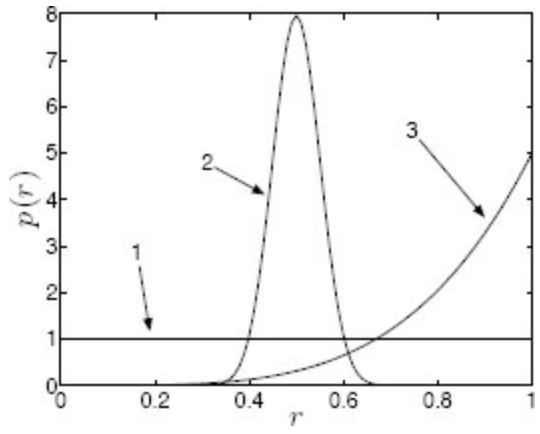

FIGURE 3.4 Examples of prior densities, $p(r)$, for r for three different scenarios.

Belief number 1 is represented as a uniform density between 0 and 1 and as such shows no preference for any particular r value. Number 2 is given a density function that is concentrated around r

= 0.5, the value we would expect for a fair coin. The density suggests that we do not expect much variance in r: it's almost certainly going to lie between 0.4 and 0.6. Most coins that any of us have tossed agree with this. Finally, number 3 encapsulates our belief that the coin (and therefore the stall owner) is biased. This density suggests that $r > 0.5$ and that there is a high level of variance. This is fine because our belief is just that the coin is biased: we don't really have any idea how biased at this stage.

We will not choose between our three scenarios at this stage, as it is interesting to see the effect these different beliefs will have on $p(r|y_N)$.

The three functions shown in Figure 3.4 have not been plucked from thin air. They are all examples of beta probability density functions (see Section 2.5.2). The beta density function is used for continuous random variables constrained to lie between 0 and 1 – perfect for our example. For a random variable R with parameters α and β, it is defined as

$$p(r) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r^{\alpha-1}(1-r)^{\beta-1}. \tag{3.4}$$

$\Gamma(\alpha)$ is known as the gamma function (see Section 2.5.2). In Equation 3.4 the gamma functions ensure that the density is normalised (that is, it integrates to 1 and is therefore a probability density function). In particular

$$\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} = \int_{r=0}^{r=1} r^{\alpha-1}(1-r)^{\beta-1} dr,$$

ensuring that

$$\int_{r=0}^{r=1} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r^{\alpha-1}(1-r)^{\beta-1} dr = 1.$$

The two parameters α and β control the shape of the resulting density function and must both be positive. Our three beliefs as plotted in Figure 3.4 correspond to the following pairs of parameter values:

1. Know nothing: $\alpha = 1$, $\beta = 1$.
2. Fair coin: $\alpha = 50$, $\beta = 50$.
3. Biased: $\alpha = 5$, $\beta = 1$.

The problem of choosing these values is a big one. For example, why should we choose $\alpha = 5$, $\beta = 1$ for a biased coin? There is no easy answer to this. We shall see later that, for the beta distribution, they can be interpreted as a number of previous, hypothetical coin tosses. For other distributions no such analogy is possible and we will also introduce the idea that maybe these too should be treated as random variables. In the mean time, we will assume that these values are sensible and move on.

The marginal distribution of y_N – $P(y_N)$ The third quantity in our equation, $P(y_N)$, acts as a normalising constant to ensure that $p(r|y_N)$ is a properly defined density. It is known as the marginal distribution of y_N because it is computed by integrating r out of the joint density $p(y_N, r)$:

$$P(y_N) = \int_{r=0}^{r=1} p(y_N, r) dr.$$

This joint density can be factorised to give

$$P(y_N) = \int_{r=0}^{r=1} P(y_N|r)p(r)\,dr,$$

which is the product of the prior and likelihood integrated over the range of values that r may take.

$p(y_N)$ is also known as the **marginal likelihood**, as it is the likelihood of the data, y_N, averaged over all parameter values. We shall see in Section 3.4.1 that it can be a useful quantity in model selection, but, unfortunately, in all but a small minority of cases, it is very difficult to calculate.

The posterior distribution – $p(r|y_N)$ This **posterior** is the distribution in which we are interested. It is the result of updating our prior belief $p(r)$ in light of new evidence y_N. The shape of the density is interesting – it tells us something about how much information we have about r after combining what we knew beforehand (the prior) and what we've seen (the likelihood). Three hypothetical examples are provided in Figure 3.5 (these are purely illustrative and do not correspond to the particular likelihood and prior examples shown in Figures 3.3 and 3.4). (a) is uniform – combining the likelihood and the prior together has left all values of r equally likely. (b) suggests that r is most likely to be low but could be high. This might be the result of starting with a uniform prior and then observing more tails than heads. Finally, (c) suggests the coin is biased to land heads more often. As it is a density, the posterior tells us not just which values are likely but also provides an indication of the level of uncertainty we still have in r having observed some data.

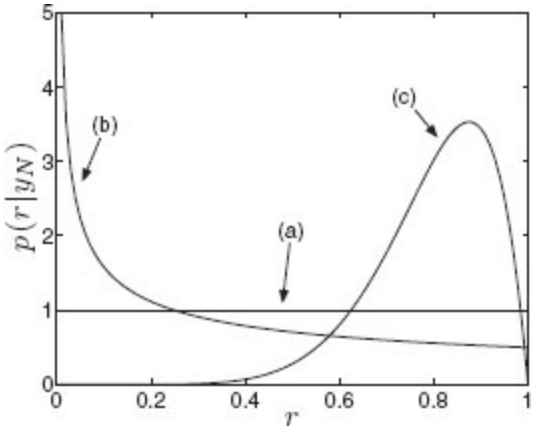

FIGURE 3.5 Examples of three possible posterior distributions $p(r|y_N)$.

As already mentioned, we can use the posterior density to compute expectations. For example, we could compute

$$\mathbf{E}_{p(r|y_N)}\left\{P(Y_{10} \leq 6) = \int_{r=0}^{r=1} P(Y_{10} \leq 6|r)p(r|y_N)dr,\right.$$

the expected value of the probability that we will win. This takes into account the data we have observed, our prior beliefs and the uncertainty that remains. It will be useful in helping to decide whether or not to play the game. We will return to this later, but first we will look at the kind of posterior densities we obtain in our coin example.

Comment 3.1 – Conjugate priors:

A likelihood-prior pair is said to be conjugate if they result in a posterior which is of the same form as the prior. This enables us to compute the posterior density analytically without having to worry about computing the denominator in Bayes' rule, the marginal likelihood. Some common conjugate pairs are listed in the table to the right.

Prior	Likelihood
Gaussian	Gaussian
Beta	Binomial
Gamma	Gaussian
Dirichlet	Multinomial

3.2 THE EXACT POSTERIOR

The beta distribution is a common choice of prior when the likelihood is a binomial distribution. This is because we can use some algebra to compute the posterior density exactly. In fact, the beta distribution is known as the **conjugate** prior to the binomial likelihood (see Comment 3.1). If the prior and likelihood are conjugate, the posterior will be of the same form as the prior. Specifically, $p(r|y_N)$ will give a beta distribution with parameters δ and γ, whose values will be computed from the prior and y_N. The beta and binomial are not the only conjugate pair of distributions and we will see an example of another conjugate prior and likelihood pair when we return to the Olympic data later in this chapter.

Using a conjugate prior makes things much easier from a mathematical point of view. However, as we mentioned in both our discussion on loss functions in Chapter 1 and noise distributions in Chapter 2, it is more important to base our choices on modelling assumptions than mathematical convenience. In the next chapter we will see some techniques we can use in the common scenario that the pair are non-conjugate.

Returning to our example, we can omit $p(y_N)$ from Equation 3.3, leaving

$$p(r|y_N) \propto P(y_N|r)p(r).$$

Replacing the terms on the right hand side with a binomial and beta distribution gives

$$p(r|y_N) \propto \left[\binom{N}{y_N} r^{y_N}(1-r)^{N-y_N}\right] \times \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r^{\alpha-1}(1-r)^{\beta-1}\right]. \quad 3.5$$

Because the prior and likelihood are conjugate, we know that $p(r|y_N)$ has to be a beta density. The beta density, with parameters δ and γ, has the following general form:

$$p(r) = Kr^{\delta-1}(1-r)^{\gamma-1},$$

where K is a constant. If we can arrange all of the terms, including r, on the right hand side of Equation 3.5 into something that looks like $r_{\delta-1}(1-r)_{\gamma-1}$, we can be sure that the constant must also be correct (it has to be $\Gamma(\delta+\gamma)/(\Gamma(\delta)\Gamma(\gamma))$ because we know that the posterior density is a beta density). In other words, we know what the normalising constant for a beta density is so we do not need to compute $p(y_N)$.

Rearranging Equation 3.5 gives us

$$
\begin{aligned}
p(r|y_N) &\propto \left[\binom{N}{y_N}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right] \times \left[r^{y_N} r^{\alpha-1}(1-r)^{N-y_N}(1-r)^{\beta-1}\right] \\
&\propto r^{y_N+\alpha-1}(1-r)^{N-y_N+\beta-1} \\
&\propto r^{\delta-1}(1-r)^{\gamma-1} \\
\textbf{where } \delta &= y_N+\alpha \textbf{ and } \gamma = N-y_N+\beta.
\end{aligned}
$$

Therefore

$$
p(r|y_N) = \frac{\Gamma(\alpha+\beta+N)}{\Gamma(\alpha+y_N)\Gamma(\beta+N-y_N)} r^{\alpha+y_N-1}(1-r)^{\beta+N-y_N-1} \tag{3.6}
$$

(note that when adding y and δ, the y_N terms cancel). This is the posterior density of r based on the prior $p(r)$ and the data y_N. Notice how the posterior parameters are computed by adding the number of heads (y_n) to the first prior parameter (α) and the number of tails ($N - y_N$) to the second (β). This allows us to gain some intuition about the prior parameters α and β – they can be thought of as the number of heads and tails in $\alpha + \beta$ previous tosses. For example, consider the second two scenarios discussed in the previous section. For the fair coin scenario, $\alpha = \beta = 50$. This is equivalent to tossing a coin 100 times and obtaining 50 heads and 50 tails. For the biased scenario, $\alpha = 5$, $\beta = 1$, corresponding to six tosses and five heads. Looking at Figure 3.4, this helps us explain the differing levels of variability suggested by the two densities: the fair coin density has much lower variability than the biased one because it is the result of many more hypothetical tosses. The more tosses, the more we should know about r.

The analogy is not perfect. For example, α and β don't have to be integers and can be less than 1 (0.3 heads doesn't make much sense). The analogy also breaks down when $\alpha = \beta = 1$. Observing one head and one tail means that values of $r = 0$ and $r = 1$ are impossible. However, density 1 in Figure 3.4), suggests that all values of r are equally likely. Despite these flaws, the analogy will be a useful one to bear in mind as we progress through our analysis (see Exercises 3.1, 3.2, 3.3 and 3.4)

3.3 THE THREE SCENARIOS

We will now investigate the posterior distribution $p(r|y_N)$ for the three different prior scenarios shown in Figure 3.4 – no prior knowledge, a fair coin and a biased coin.

3.3.1 No prior knowledge

In this scenario (MATLAB script: **coin_scenario1.m**), we assume that we know nothing of coin tossing or the stall holder. Our prior parameters are $\alpha = 1$, $\beta = 1$, shown in Figure 3.6(a).

To compare different scenarios we will use the expected value and variance of r under the prior. The expected value of a random variable from a beta distribution with parameters α and β (the density function of which we will henceforth denote as $\mathcal{B}(\alpha, \beta)$) is given as (see Exercise 3.5)

$$
\begin{aligned}
p(r) &= \mathcal{B}(\alpha,\beta) \\
\mathbf{E}_{p(r)}\{R\} &= \frac{\alpha}{\alpha+\beta}.
\end{aligned}
$$

For scenario 1:

$$
\mathbf{E}_{p(r)}\left\{R\right\} = \frac{\alpha}{\alpha+\beta} = \frac{1}{2}.
$$

The variance of a beta distributed random variable is given by (see Exercise 3.6)

$$\mathbf{var}\Big\{R\Big\} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}, \tag{3.7}$$

which for $\alpha = \beta = 1$ is

$$\mathbf{var}\Big\{R\Big\} = \tfrac{1}{12}.$$

Note that in our formulation of the posterior (Equation 3.6) we are not restricted to updating our distribution in blocks of ten – we can incorporate the results of any number of coin tosses. To illustrate the evolution of the posterior, we will look at how it changes toss by toss.

A new customer hands over £1 and the stall owner starts tossing the coin. The first toss results in a head. The posterior distribution after one toss is a beta distribution with parameters $\delta = \alpha + y_N$ and $\gamma = \beta + N - y_N$:

$$p(r|y_N) = \mathscr{B}(\delta, \gamma).$$

In this scenario, $\alpha = \beta = 1$, and as we have had $N = 1$ tosses and seen $y_N = 1$ heads,

$$\begin{aligned} \delta &= 1+1 = 2 \\ \gamma &= 1+1-1 = 1. \end{aligned}$$

This posterior distribution is shown as the solid line in Figure 3.6(b) (the prior is also shown as a dashed line). This single observation has had quite a large effect – the posterior is very different from the prior. In the prior, all values of r were equally likely. This has now changed – higher values are more likely than lower values with zero density at $r = 0$. This is consistent with the evidence – observing one head makes high values of r slightly more likely and low values slightly less likely. The density is still very broad, as we have observed only one toss. The expected value of r under the posterior is

$$\mathbf{E}_{p(r|y_N)}\Big\{R\Big\} = \tfrac{2}{3}$$

and we can see that observing a solitary head has increased the expected value of r from 1/2 to 2/3. The variance of the posterior is (using Equation 3.7)

$$\mathbf{var}\Big\{R\Big\} = \tfrac{1}{18}$$

which is lower than the prior variance (1/12). So, the reduction in variance tells us that we have less uncertainty about the value of r than we did (we have learnt something) and the increase in expected value tells us that what we've learnt is that heads are slightly more likely than tails.

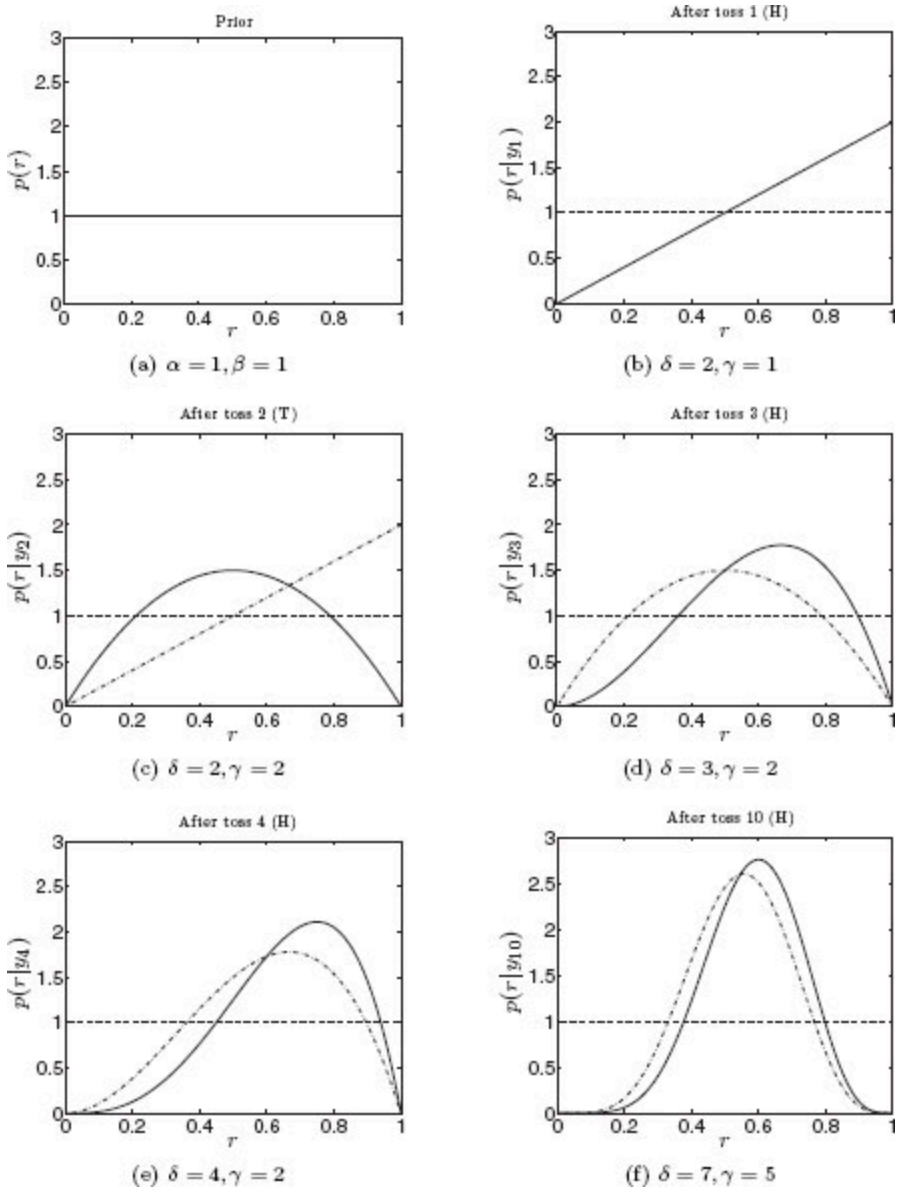

FIGURE 3.6 Evolution of $p(r|y_N)$ as the number of observed coin tosses increases.

The stall owner tosses the second coin and it lands tails. We have now seen one head and one tail and so $N = 2$, $y_N = 1$, resulting in

$$\begin{aligned} \delta &= 1+1=2 \\ \gamma &= 1+2-1=2. \end{aligned}$$

The posterior distribution is shown as the solid dark line in Figure 3.6(c). The lighter dash-dot line is the posterior we saw after one toss and the dashed line is the prior. The density has changed again to reflect the new evidence. As we have now observed a tail, the density at $r = 1$ should be zero and is ($r = 1$ would suggest that the coin always lands heads). The density is now curved rather than straight (as we have already mentioned, the beta density function is very flexible) and observing a tail has made lower values more likely. The expected value and variance are now

$$\mathbf{E}_{p(r|y_N)}\{R\} = \tfrac{1}{2}, \mathbf{var}\{R\} = \tfrac{1}{20}.$$

The expected value has decreased back to 1/2. Given that the expected value under the prior was also 1/2, you might conclude that we haven't learnt anything. However, the variance has decreased again (from 1/18 to 1/20) so we have less uncertainty in r and have learnt something. In fact, we've learnt that r is closer to 1/2 than we assumed under the prior.

The third toss results in another head. We now have $N = 3$ tosses, $y_N = 2$ heads and $N - y_N = 1$ tail. Our updated posterior parameters are

$$\begin{aligned}\delta &= \alpha + y_N = 1 + 2 = 3 \\ \gamma &= \beta + N - y_N = 1 + 3 - 2 = 2.\end{aligned}$$

This posterior is plotted in Figure 3.6(d). Once again, the posterior is the solid dark line, the previous posterior is the solid light line and the dashed line is the prior. We notice that the effect of observing this second head is to skew the density to the right, suggesting that heads are more likely than tails. Again, this is entirely consistent with the evidence – we have seen more heads than tails. We have only seen three coins though, so there is still a high level of uncertainty – the density suggests that r could potentially still be pretty much any value between 0 and 1. The new expected value and variance are

$$\mathbf{E}_{p(r|y_N)}\{R\} = \tfrac{3}{5}, \mathbf{var}\{R\} = \tfrac{1}{25}.$$

The variance has decreased again reflecting the decrease in uncertainty that we would expect as we see more data.

Toss 4 also comes up heads ($y_N = 3$, $N = 4$), resulting in $\delta = 1 + 3 = 4$ and $\gamma = 1 + 4 - 3 = 2$. Figure 3.6(e) shows the current and previous posteriors and prior in the now familiar format. The density has once again been skewed to the right – we've now seen three heads and only one tail so it seems likely that r is greater than 1/2. Also notice the difference between the $N = 3$ posterior and the $N = 4$ posterior for very low values of r – the extra head has left us pretty convinced that r is not 0.1 or lower. The expected value and variance are given by

$$\mathbf{E}_{p(r|y_N)}\{R\} = \tfrac{2}{3}, \mathbf{var}\{R\} = \tfrac{2}{63} = 0.0317,$$

where the expected value has increased and the variance has once again decreased. The remaining six tosses are made so that the complete sequence is

$$\mathbf{H}, \mathbf{T}, \mathbf{H}, \mathbf{H}, \mathbf{H}, \mathbf{H}, \mathbf{T}, \mathbf{T}, \mathbf{T}, \mathbf{H},$$

a total of six heads and four tails. The posterior distribution after $N = 10$ tosses ($y_N = 6$) has parameters $\delta = 1 + 6 = 7$ and $\gamma = 1 + 10 - 6 = 5$. This (along with the posterior for $N = 9$) is shown in Figure 3.6(f). The expected value and variance are

$$\mathbf{E}_{p(r|y_N)}\{R\} = \tfrac{7}{12} = 0.5833, \mathbf{var}\{R\} = 0.0187. \tag{3.8}$$

Our ten observations have increased the expected value from 0.5 to 0.5833 and decreased our variance from 1/12 = 0.0833 to 0.0187. However, this is not the full story. Examining Figure 3.6(f), we see that we can also be pretty sure that $r > 0.2$ and $r < 0.9$. The uncertainty in the value of r is still quite high because we have only observed ten tosses.

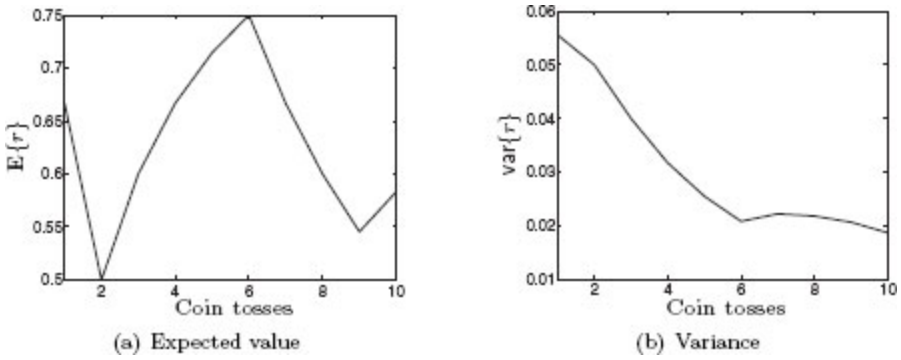

FIGURE 3.7 Evolution of expected value (a) and variance (b) of r as coin toss data is added to the posterior.

Figure 3.7 summarises how the expected value and variance change as the 10 observations are included. The expected value jumps around a bit, whereas the variance steadily decreases as more information becomes available. At the seventh toss, the variance increases. The first seven tosses are

$$\mathbf{H, T, H, H, H, H, T.}$$

The evidence up to and including toss 6 is that heads is much more likely than tails (5 out of 6). Tails on the seventh toss is therefore slightly unexpected. Figure 3.8 shows the posterior before and after the seventh toss. The arrival of the tail has forced the density to increase the likelihood of low values of r and, in doing so, increased the uncertainty.

The posterior density encapsulates all of the information we have about r. Shortly, we will use this to compute the expected probability of winning the game. Before we do so, we will revisit the idea of using point estimates by extracting a single value $\hat{r}$ of r from this density. We will then be able to compare the expected probability of winning with the probability of winning computed from a single value of r. A sensible choice would be to use $\mathbf{E}_{p(r|yN)}\{R\}$. With this value, we can compute the probability of winning – $-P(Y_{\text{new}} \leq 6|\hat{r})$. This quantity could be used to decide whether or not to play. Note that, to make the distinction between observed tosses and future tosses, we will use Y_{new} as a random variable that describes ten future tosses.

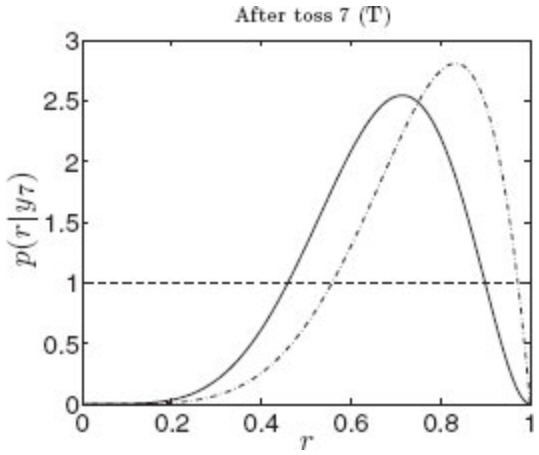

FIGURE 3.8 The posterior after six (light) and seven (dark) tosses.

After ten tosses, the posterior density is beta with parameters $\delta = 7$, $\gamma = 5$. $\hat{r}$ is therefore

$$\hat{r} = \frac{\delta}{\delta+\gamma} = \frac{7}{12}.$$

The probability of winning the game follows as

$$\begin{aligned} P(Y_{\text{new}} \leq 6|\hat{r}) &= 1 - \sum_{y_{\text{new}}=7}^{10} P(Y_{\text{new}} = y_{\text{new}}|\hat{r}) \\ &= 1 - 0.3414 \\ &= 0.6586, \end{aligned}$$

suggesting that we will win more often than lose.

Using all of the posterior information requires computing

$$\mathbf{E}_{p(r|y_N)}\left\{P(Y_{\text{new}} \leq 6|r)\right\}.$$

Rearranging and manipulating the expectation provides us with the following expression:

$$\begin{aligned} \mathbf{E}_{p(r|y_N)}\left\{P(Y_{\text{new}})\right\} &= \mathbf{E}_{p(r|y_N)}\left\{1 - P(Y_{\text{new}} \geq 7|r)\right\} \\ &= 1 - \mathbf{E}_{p(r|y_N)}\left\{P(Y_{\text{new}} \geq 7|r)\right\} \\ &= 1 - \mathbf{E}_{p(r|y_N)}\left\{\sum_{y_{\text{new}}=7}^{y_{\text{new}}=10} P(Y_{\text{new}} = y_{\text{new}}|r)\right\} \\ &= 1 - \sum_{y_{\text{new}}=7}^{y_{\text{new}}=10} \mathbf{E}_{p(r|y_N)}\left\{P(Y_{\text{new}} = y_{\text{new}}|r)\right\}. \end{aligned} \tag{3.9}$$

To evaluate this, we need to be able to compute $\mathbf{E}_{p(r|y_N)}\left\{P(Y_{\text{new}} = y_{\text{new}}|r)\right\}$. From the definition of expectations, this is given by

$$\begin{aligned} &\mathbf{E}_{p(r|y_N)}\left\{P(Y_{\text{new}} = y_{\text{new}}|r)\right\} = \int_{r=0}^{r=1} P(Y_{\text{new}} = y_{\text{new}}|r)p(r|y_N)dr \\ &= \int_{r=0}^{r=1} \left[\binom{N_{\text{new}}}{y_{\text{new}}} r^{y_{\text{new}}}(1-r)^{N_{\text{new}}-y_{\text{new}}}\right]\left[\frac{\Gamma(\delta+\gamma)}{\Gamma(\delta)\Gamma(\gamma)} r^{\delta-1}(1-r)^{\gamma-1}\right] dr \\ &= \binom{N_{\text{new}}}{y_{\text{new}}} \frac{\Gamma(\delta+\gamma)}{\Gamma(\delta)\Gamma(\gamma)} \int_{r=0}^{r=1} r^{y_{\text{new}}+\delta-1}(1-r)^{N_{\text{new}}-y_{\text{new}}+\gamma-1} dr. \end{aligned} \tag{3.10}$$

This integral looks a bit daunting. However, on closer inspection, the argument inside the integral is an unnormalised beta density with parameters $\delta + y_{\text{new}}$ and $y + N_{\text{new}} - y_{\text{new}}$. In general, for a beta density with parameters α and β, the following *must* be true:

$$\int_{r=0}^{r=1} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} r^{\alpha-1}(1-r)^{\beta-1} dr = 1,$$

and therefore

$$\int_{r=0}^{r=1} r^{\alpha-1}(1-r)^{\beta-1}dr = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Our desired expectation becomes

$$\mathbf{E}_{p(r\backslash y_N)}\left\{P(Y_{\text{new}} = y_{\text{new}}\Big|r)\right\} = \binom{N_{\text{new}}}{y_{\text{new}}} \frac{\Gamma(\delta+\gamma)}{\Gamma(\delta)\Gamma(\gamma)} \frac{\Gamma(\delta+y_{\text{new}})\Gamma(\gamma+N_{\text{new}}-y_{\text{new}})}{\Gamma(\delta+\gamma+N_{\text{new}})}$$

which we can easily compute for a particular posterior (i.e. values of y and δ) and values of N_{new} and y_{new}.

After ten tosses, we have $\delta = 7$, $y = 5$. Plugging these values in, we can compute the expected probability of success:

$$\begin{aligned}\mathbf{E}_{p(r|y_N)}\left\{P(Y_{\text{new}} \leq 6\big|r)\right\} &= 1 - \sum_{y_{\text{new}=7}}^{y_{\text{new}}=10} \mathbf{E}_{p(r|y_N)}\left\{P(Y_{\text{new}} = y_{\text{new}}\big|r)\right\} \\ &= 1 - 0.3945 \\ &= 0.6055.\end{aligned}$$

Comparing this with the value obtained using the point estimate, we can see that both predict we will win more often than not. This is in agreement with the evidence – the one person we have fully observed got six heads and four tails and hence won £2. The point estimate gives a higher probability – ignoring the posterior uncertainty makes it more likely that we will win.

Another customer plays the game. The sequence of tosses is

$$\mathbf{H}, \mathbf{H}, \mathbf{T}, \mathbf{T}, \mathbf{H}, \mathbf{H}, \mathbf{H}, \mathbf{H}, \mathbf{H}, \mathbf{H},$$

eight heads and two tails – the stall owner has won. Combining all 20 tosses that we have observed, we have $N = 20$, $y_N = 6 + 8 = 14$ heads and $N - y_N = 20 - 14 = 6$ tails. This gives $\delta = 15$ and $y = 7$. The posterior density is shown in Figure 3.9 where the light line shows the posterior we had after ten and the dashed line the prior. The expected value and variance are

$$\mathbf{E}_{p(r|y_N)}\{R\} = 0.6818, \mathbf{var}\{R\} = 0.0094.$$

The expected value has increased and the variance has decreased (c.f. Equation 3.8). Both behaviours are what we would expect – eight heads and two tails should increase the expected value of r and the increased data should decrease the variance. We can now recompute $\mathbf{E}_{p(r|y_N)}\left\{P(Y_{\text{new}} \leq 6\big|r)\right\}$ in light of the new evidence. Plugging in the appropriate values, this is

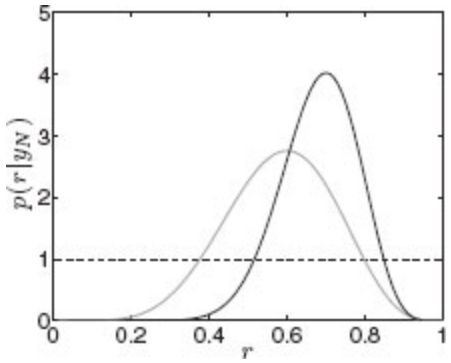

FIGURE 3.9 Posterior distribution after observing 10 tosses (light curve) and 20 tosses (dark curve). The dashed line corresponds to the prior density.

$$\mathbf{E}_{p(r|y_N)}\left\{P(Y_{\text{new}} \leq 6|r)\right\} = 0.4045.$$

The new evidence has pushed the density to the right, made high values of r (and hence the coin landing heads) more likely and reduced the probability of winning. For completeness, we can also compute $P(Y_{\text{new}} \leq 6|\hat{r}) = 0.3994$.

This corresponds to an expected return of:

$$2 \times 0.4045 - 1 = -0.1910,$$

equivalent to a loss of about 20p per go.

In this example we have now touched upon all of the important components of Bayesian Machine Learning – choosing priors, choosing likelihoods, computing posteriors and using expectations to make predictions. We will now repeat this process for the other two prior scenarios.

3.3.2 The fair coin scenario

For the fair coin scenario (MATLAB script: **coin_scenario2.m**), we assumed that $\alpha = \beta = 50$, which is analogous to assuming that we have already witnessed 100 tosses, half of which resulted in heads. The first thing to notice here is that 100 tosses corresponds to much more data than we are going to observe here (20 tosses). Should we expect our data to have the same effect as it did in the previous scenario?

Figure 3.10(a) shows the prior density and Figures 3.10(b), 3.10(c), 3.10(d), 3.10(e) and 3.10(f) show the posterior after 1, 5, 10, 15 and 20 tosses, respectively. For this scenario, we have not shown the previous posterior at each stage – it is too close to the current one. However, in most cases, the change in posterior is so small that the lines almost lie right on top of one another. In fact, it is only after about ten tosses that the posterior has moved significantly from the prior. Recalling our analogy for the beta prior, this prior includes the evidential equivalent of 100 tosses and so it is not surprising that adding another ten makes much difference.

The evolution of $\mathbf{E}_{p(r|y_N)}\{R\}$ and var{R} as the 20 tosses are observed can be seen in Figure 3.11. We see very little change in either as the data appear compared to the changes we observed in Figure 3.6. Such small changes are indicative of a very *strong* prior density. The prior will dominate over the data until we've observed many more tosses – i.e., $p(r)$ dominates $p(y_N|r)$ in Equation 3.3. We have created a model that is stuck in its ways and will require a lot of persuasion to believe otherwise.

Just as in the previous section, we can work out $\mathbf{E}_{p(r|y_N)}\left\{P(Y_{\text{new}} \leq 6|r)\right\}$. After all 20 tosses have been observed, we have $\delta = \alpha + y_N = 50 + 14 = 64$ and $\gamma = \beta + N - y_N = 50 + 20 - 14 = 56$. The expectation works out as

$$\mathbf{E}_{p(r|y_N)}\left\{P(Y_{\text{new}} \leq 6|r)\right\} = 0.7579. \tag{3.11}$$

As before, we can also see how much difference there is between this value and the value obtained using the point estimate $\hat{r}, P(Y_{\text{new}} \leq 6|\hat{r})$ (in this case, $\hat{r} = 64/(64+56) = 0.5333$):

$$P(Y_{\text{new}} \leq 6|\hat{r}) = 0.7680.$$

Both quantities predict that we will win more often than not. In light of what we've seen about the posterior, this should come as no surprise. The data has done little to overcome the prior assumption

that the coin is fair, and we already know that, if the coin is fair, we will tend to win (a fair coin will result in us winning, on average, 66p per game – see the start of Section 3.1).

As an aside, consider how accurate our approximation $P(Y_{\text{new}} \leq 6|\hat{r})$ is to the proper expectation in this scenario and the previous one. In the previous one, the difference between the two values was

$$\left|\mathbf{E}_{p(r|y_N)}\left\{P(Y_{\text{new}} \leq 6|r)\right\} - P(Y_{\text{new}} \leq 6|\hat{r})\right| = 0.0531.$$

In this example, the values are closer:

$$\left|\mathbf{E}_{p(r|y_N)}\left\{P(Y_{\text{new}} \leq 6|r)\right\} - P(Y_{\text{new}} \leq 6|\hat{r})\right| = 0.0101.$$

There is a good reason why this is the case – as the variance in the posterior decreases (the variance in scenario 2 is much lower than in scenario 1), the probability density becomes more and more condensed around one particular point. Imagine the variance