

Introduction to Bayesian learning

What is Bayesian learning?

1. A formal model of uncertainty
2. A method for expressing prior beliefs
3. A methodology for making inference about data
4. A paradigm for decision making

The central difference *on the learning side* between Bayesian and non-Bayesian learning (aka frequentist learning or learning theory) is the Bayesian treat parameters as true unknowns—i.e., as random variables.

Let's take an example from statistics.

Let's say we have a coin that may be biased. It has probability $\pi \in [0, 1]$ of coming up heads. Suppose we flip it once and it comes up tails. How do we infer π ?

Frequentist answer: $\pi = 0$, because this is the maximum likelihood solution.

Sort-of frequentist answer: $\pi = \frac{1}{3}$ because I'll "smooth" and compute $\pi = (\# \text{ heads} + 1)/(\text{total flips} + 2)$.

These are derived because we assume that we want to find π which maximizes the likelihood of the data, $p(D | \pi)$. Several people have complained that conditioning on π is weird and it is! Only random variables should be conditioned on, and in frequentist land, a parameter is definitely *not* a random variable.

Let's say that we know $\pi \in \{0, 0.25, 0.5, 0.75, 1\}$. Still, the ML solution would give 0.

The Bayesian solution is quite different. We don't actually try to "find" a single value of π , but rather compute a distribution over possible π . This comes from a simple application of Bayes rule:

$$p(\pi | D) = \frac{p(\pi)p(D | \pi)}{p(D)} = \frac{p(\pi)p(D | \pi)}{\sum_{\pi'} p(\pi')p(D | \pi')}$$

Here, $p(\pi)$ is called the *prior*, $p(D | \pi)$ is the *likelihood* and $p(D)$ is the *marginal* (or *evidence* or *partition function*).

In our coin flipping example, our likelihood is just $\pi^h(1 - \pi)^t$, where h and t are the counts of heads and tails.

If we think about the frequentist perspective, what happens is that they effectively put a uniform prior over π and "approximate" the posterior by a point distribution centered at the maximum. (We will soon see how to justify smoothing in a similar manner.)

But this entails two weird approximations: maybe we don't want a uniform prior and maybe we don't want to make this approximation.

Let's say that *a priori*, we believe the five values of π have probability 0.1, 0.2, 0.4, 0.2, 0.1, respectively. This basically means that we expect the coin is likely to not be severely biased. This is a valid prior because it sums to one over the range of π .

Now, let's revisit the case where we flip once and it comes up tails. This gives us the following unnormalized posterior:

$$\begin{aligned}
p(\pi = 0 \mid D) &\propto 0.1 \times 0^0 \times 1^1 = 0.1 \\
p(\pi = 0.25 \mid D) &\propto 0.2 \times 0.25^0 \times 0.75^1 = 0.15 \\
p(\pi = 0.5 \mid D) &\propto 0.4 \times 0.5^0 \times 0.5^1 = 0.2 \\
p(\pi = 0.75 \mid D) &\propto 0.2 \times 0.75^0 \times 0.25^1 = 0.05 \\
p(\pi = 1 \mid D) &\propto 0.1 \times 1^0 \times 0^1 = 0
\end{aligned}$$

After normalizing, we get: 0.2, 0.3, 0.4, 0.1, 0 as the posterior. Note that the posterior at $\pi = 0.5$ hasn't changed, but the probability of $\pi > 0.5$ has significantly decreased (and we *know* that $\pi = 1$ is impossible).

Suppose we flip again and get another tails. This gives:

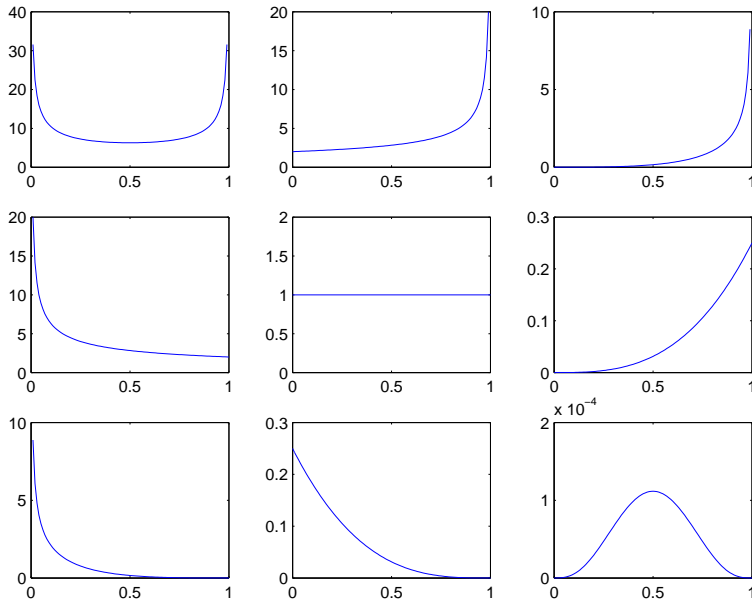
$$\begin{aligned}
p(\pi = 0 \mid D) &\propto 0.2 \times 0^0 \times 1^1 = 0.2 \\
p(\pi = 0.25 \mid D) &\propto 0.3 \times 0.25^0 \times 0.75^1 = 0.225 \\
p(\pi = 0.5 \mid D) &\propto 0.4 \times 0.5^0 \times 0.5^1 = 0.2 \\
p(\pi = 0.75 \mid D) &\propto 0.1 \times 0.75^0 \times 0.25^1 = 0.025 \\
p(\pi = 1 \mid D) &\propto 0.0 \times 1^0 \times 0^1 = 0
\end{aligned}$$

Here we have used a technique known as *posterior updating*. We take the posterior from the first example and treat it as the prior for the second example. After normalizing, we get approximately: 0.31, 0.35, 0.31, 0.03, 0. Now, we are more sure that π should be 0.25, but only by a little. We can repeat this process indefinitely. If we observe an infinite number of flips, we will converge to the true value (this is known as *consistency*).

Now, let's say that we don't want to confine π to one of five values but want to allow it to range continuously. That is, we need a probability distribution p with domain $[0, 1]$. One could cook up many such distributions (with a bit of integration to ensure proper normalization). However, there is a standard such distribution known as the *beta* distribution:

$$\mathcal{B}et(\pi \mid a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \pi^{a-1} (1-\pi)^{b-1}$$

Here, a and b are parameters of the prior, or *hyperparameters* of the model. Ignore the fraction term for a second (it serves to normalize the beta). Nine beta distributions are shown below with $a \in \{0, 1, 5\}$ in the columns and $b \in \{0, 1, 5\}$ in the rows:



Now, let's consider our posterior updating after observing a single tails. We have:

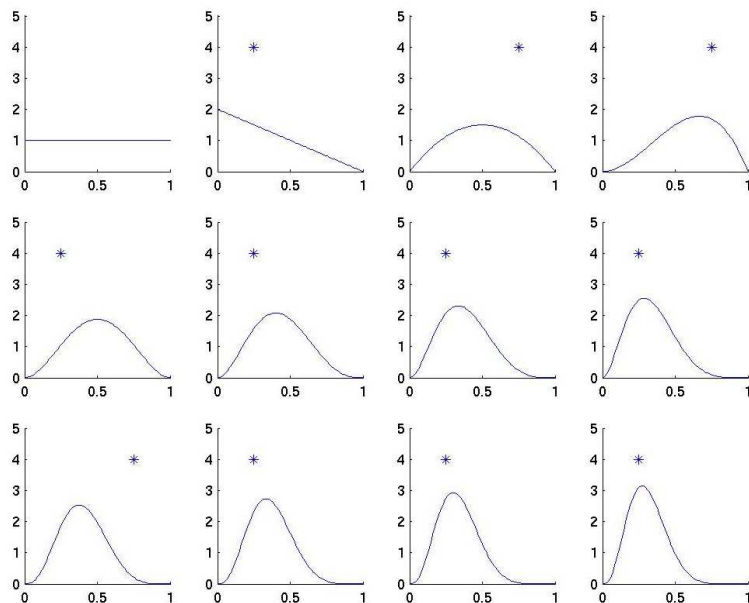
$$\begin{aligned}
 p(\pi \mid D) &= \frac{1}{Z} p(\pi) p(D \mid \pi) \\
 &= \frac{1}{Z} \text{Bet}(\pi \mid a, b) p(D \mid \pi) \\
 &= \frac{1}{Z} [\pi^{a-1} (1-\pi)^{b-1}] [\pi^0 (1-\pi)^1] \\
 &= \frac{1}{Z} \pi^{a-1} (1-\pi)^{b+1-1} \\
 &= \text{Bet}(\pi \mid a, b+1)
 \end{aligned}$$

In general, this works for any amount of heads and tails. Suppose we have h heads and t tails, then we get:

$$\begin{aligned}
 p(\pi \mid D) &= \frac{1}{Z} p(\pi) p(D \mid \pi) \\
 &= \frac{1}{Z} \text{Bet}(\pi \mid a, b) p(D \mid \pi) \\
 &= \frac{1}{Z} [\pi^{a-1} (1-\pi)^{b-1}] [\pi^h (1-\pi)^t] \\
 &= \frac{1}{Z} \pi^{a+h-1} (1-\pi)^{b+t-1} \\
 &= \text{Bet}(\pi \mid a+h, b+t)
 \end{aligned}$$

We don't need to worry about computing Z because we know that the beta distribution is properly normalized!

An example of how this works is shown below:



This shown results after eleven flips with a uniform ($a = b = 1$) Beta prior. The flips are: THHTTTTHTTT (the stars indicate whether it was tails or heads). You can see that over time, the distribution tends toward $\pi < 0.5$ and becomes more and more peaked.

It's easy to verify that the value of π that maximizes $\text{Bet}(\pi | a, b)$ is exactly $a/(a+b)$. This (somewhat) justifies smoothing: to obtain “add one” smoothing, we pretend that we start with a beta prior with $a = b = 1$. To get “add alpha” smoothing, we set $a = b = \lambda$. Then we do “maximum likelihood.” Technically, this is called *maximum a posteriori* or *MAP*, since we’re choosing a value that maximizes the posterior, rather than one that maximizes the likelihood.

Now, back to the $\Gamma(\cdot)$ function. This thing appears all the time in normalization terms, and is defined by:

$$\Gamma(z) = \int_0^{\infty} dt t^{z-1} \exp[-t]$$

This integral has no closed form solution. However, it can be computed by standard techniques, available in matlab and many other languages. It has the nice property that it extends the factorial function to the real line: if n is a positive integer, then $\Gamma(n) = (n-1)!$. Given this, we often compute $\log \Gamma(\cdot)$, since it grows too quickly. The functions in matlab are `gamma` and `gamma1n`.

There are two other distributions you’ll need to know about (other than the standard Normal, Multinomial, Binomial, etc.). These are the gamma and the Dirichlet.

We’ll do Dirichlet first, since it’s basically an extension of the beta. Remember that a multinomial is just like a more complicated binomial. Instead of having a coin with a single parameter π , we have a die with a parameter vector $\theta_1, \dots, \theta_K$, all positive and sum to one. We would like a prior on this. The Dirichlet is a multivariate version of the beta. It is parameterized by a vector $\alpha_1, \dots, \alpha_K$, all positive but need *not* sum to one:

$$\text{Dir}(\theta | \alpha) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_k^{\alpha_k - 1}$$

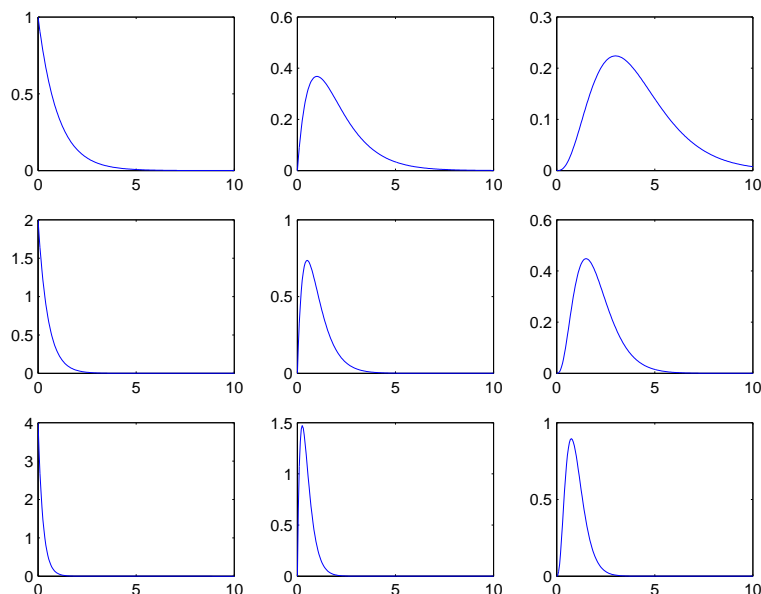
In the two parameter case, this is exactly a beta distribution. We also get the same posterior updating. If the prior was $\text{Dir}(\theta | \alpha)$, then after observing x_1 rolls of a 1, x_2 rolls of a 2 and so on, the *posterior*

hyperparameters becomes $(\alpha_1 + x_1, \dots, \alpha_K + x_K) = \alpha + x$. Again, we can think of smoothing as MAP inference with a Dirichlet prior.

Finally, we need a gamma distribution. This is a distribution over positive reals. This will be useful as a prior for the inverse variance of a normal distribution (i.e., $p(1/\sigma^2)$), but for now, just think of it as some distribution over $(0, \infty)$:

$$\text{Gam}(\lambda \mid a, b) = \frac{b^a}{\Gamma(a)} \lambda^{-a-1} \exp[-\lambda/b]$$

Note that several definitions of *Gam* actually exist – sometimes people use $-a$ instead of a , which puts the b^a in the denominator and replaces λ^{-a-1} with λ^{a-1} . Examples of gamma priors are shown below:



Now, suppose we have a posterior $p(\theta \mid D)$ (here, θ is just an arbitrary symbol for “parameters”). What do we do with it? Well, at the end of the day, sometimes the posterior is of interest in its own right. Often, however, we probably want to make predictions. That is, we may want to predict how many of 100 coin flips will land tails. In general, if there’s a quantity $f(\theta)$ that we want to predict, we want to compute:

$$\begin{aligned} \mathbb{E}_{\theta \sim p(\cdot \mid D)} [f(\theta)] &= \int_{\Theta} d\theta p(\theta \mid D) f(\theta) \\ &= \frac{1}{p(D)} \int_{\Theta} d\theta p(\theta) p(D \mid \theta) f(\theta) \end{aligned}$$

(If Θ is a discrete space, replace the integrals by sums.)

For instance, take the coin flipping example. Suppose we have a $\text{Bet}(1, 1)$ prior, then observe 9 heads and 19 tails. This gives us a $\text{Bet}(10, 20)$ posterior. Let’s say for simplicity that we want to know the probability that the next two flips will come up tails. In this case (writing π for θ), we have $f(\pi) = (1 - \pi)^2$. Thus, we want to compute:

$$\begin{aligned}
\int d\pi \mathcal{B}et(\pi | 10, 20)(1 - \pi)^2 &= \int d\pi \frac{\Gamma(30)}{\Gamma(10)\Gamma(20)} \pi^9 (1 - \pi)^{19} (1 - \pi)^2 \\
&= \int d\pi \frac{\Gamma(30)}{\Gamma(10)\Gamma(20)} \pi^9 (1 - \pi)^{21} \\
&= \int d\pi \frac{\Gamma(30)}{\Gamma(10)\Gamma(20)} \frac{\Gamma(10)\Gamma(22)}{\Gamma(32)} \frac{\Gamma(32)}{\Gamma(10)\Gamma(22)} \pi^9 (1 - \pi)^{21} \\
&= \frac{\Gamma(30)}{\Gamma(10)\Gamma(20)} \frac{\Gamma(10)\Gamma(22)}{\Gamma(32)} \int d\pi \frac{\Gamma(32)}{\Gamma(10)\Gamma(22)} \pi^9 (1 - \pi)^{21} \\
&= \frac{\Gamma(30)}{\Gamma(20)} \frac{\Gamma(22)}{\Gamma(32)} \int d\pi \mathcal{B}et(\pi | 10, 22) \\
&= \frac{\Gamma(30)}{\Gamma(20)} \frac{\Gamma(22)}{\Gamma(32)} = \frac{29!21!}{31!19!} = \frac{21 \times 20}{31 \times 30} = 0.45
\end{aligned}$$