From Bayes' Rule to Bayesian Statistics

Bayesian Data Analysis

Steve Buyske

How Bayes' Rule leads to Bayesian statistics

- If we think of parameters as being *random variables* rather than *fixed constants*, we can use Bayes' Rule with parameters.
- Suppose that we have a statistical model for our data, so that we can calculate

$$\text{Prob}(\text{data} \mid \text{parameters}),$$

- and suppose—since the parameters are random variables—that we also know a distribution for the parameters, which we call the *prior distribution*, or *prior*:

$$\text{Prob}(\text{parameters}).$$

- If we also can calculate the probability of the data, without reference to the parameters (more on this later),

$$\text{Prob}(\text{data}),$$

- then Bayes' Rule can be rewritten from

$$\text{Prob}(A \mid B) = \frac{\text{Prob}(B \mid A)\text{Prob}(A)}{\text{Prob}(B)}$$

  to

$$\text{Prob}(\text{parameters} \mid \text{data}) = \frac{\text{Prob}(\text{data} \mid \text{parameters})\text{Prob}(\text{parameters})}{\text{Prob}(\text{data})}.$$

- The probability on the left is the *posterior probability*, or *posterior* for short.
- This is how Bayes' Rule enables us to reason about the values of the parameter given data, assumptions about the model of how the data is generated, and a prior probability about the values of the parameters.
- So much flows from this!

- A little vocabulary
  - We've already mentioned the *posterior* and the *prior*.
  - The term $\text{Prob}(\text{data} \mid \text{parameters})$ is called the *likelihood*.
  - The bottom of the denominator, $\text{Prob}(\text{data})$, doesn't really have a standard name, but can be called the
  - *evidence*,
  - the *marginal likelihood*,
  - the *probability of the data*, or
  - the *average likelihood*.

More notation

- Let's write
  - θ ("theta") to denote the parameters
  - D for data
- Then Bayes' Rule looks rather mathematical in the form

$$\text{Prob}(\theta \mid D) = \frac{\text{Prob}(D \mid \theta)\text{Prob}(\theta)}{\text{Prob}(D)}.$$

- If there are θ takes discrete values, then by the Law of Total Probability we can rewrite this as

$$\text{Prob}(\theta \mid D) = \frac{\text{Prob}(D \mid \theta)\text{Prob}(\theta)}{\sum\limits_{\text{all possible }\theta^*} \text{Prob}(D \mid \theta^*)\text{Prob}(\theta^*)}.$$

Continuous parameters

- The version of Bayes' Rule on the previous slide,

$$\text{Prob}(\theta \mid D) = \frac{\text{Prob}(D \mid \theta)\text{Prob}(\theta)}{\sum\limits_{\text{all possible } \theta^*} \text{Prob}(D \mid \theta^*)\text{Prob}(\theta^*)},$$

  only makes sense for a discrete parameter (like the unknown number of red marbles in a bag).
- Most parameters (think population mean or standard deviation) are continuous variables.
- In that case, we just replace the sum in the expression below with an integral:

$$\text{Prob}(\theta \mid D) = \frac{\text{Prob}(D \mid \theta)\text{Prob}(\theta)}{\int \text{Prob}(D \mid \theta^*)\text{Prob}(\theta^*)\, d\theta^*},$$

The problem with Bayesian statistics

- It turns out the right hand side can be **very** difficult to calculate, but there are different approaches:

1. Analytically, that is, using formal math, or
2. Numerical integration over a grid, or
3. Quadratic approximation, or
4. Markov chain Monte Carlo (MCMC) **This is what we will mostly do.**

Analytically

- Once done it's very fast, but it often cannot be done.
- It requires integrating $\int \mathrm{Prob}(D \mid \theta^*)\mathrm{Prob}(\theta^*)\, d\theta^*$.

Numerical integration over a grid

- Imagine defining a grid over the possible values of the parameters.
- Compute the value of the prior at each grid point.
- Compute the value of the likelihood at each grid point.
- Multiply them together to get a numerical approximation of the numerator of Bayes' Rule, sometimes called the "unstandardized posterior."
- Add the unstandardized posterior up (that approximates the denominator of Bayes' Rule), and then divide the unstandardized posterior by that sum.
- That's a numerical approximation of the posterior.
- This approach won't work if there are too many parameters—the grid will just have too many points.

Quadratic approximation

- Use numerical optimization to find the maximimum of the unweighted posterior.
- Find the curvature there and fit a quadratic.
- The result is a normal approximation to the likelihood.
- We won't use this approach, but you can find a nice description here and in the first section of this blog post
- The approximation gets worse the further away you move away from the maximum.

Markov chain Monte Carlo (MCMC)

- We will spend a lot of time on this.
- for now, the idea is to develop a way to draw samples from the posterior:
- regions with high probability will be sampled a lot
- regions with low probability will be sampled only a little
- If we sample enough times—think tens of thousands—then the empirical distribution of our sample will look a lot like the actual distribution of the posterior.
- The difficult is in figuring out *how* to draw samples from the posterior.