14.4: Bayesian Methods

Taylor

University of Virginia

Introduction

This is the main difference between frequentist statistics and Bayesian statistics:

Frequentist statistics assumes that parameter values are fixed.

Bayesian statistics assumes that parameter values are random.

To emphasize this difference in notation, instead of writing $f(x_1, \ldots, x_n; \theta)$, we will write

$$f(x_1, \ldots, x_n | \theta)$$

to emphasize that we are conditioning on a random variable θ.

Introduction

You may have seen Bayes' rule in terms of probabilities of simple events A and B as

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}$$

Now we'll use Bayes' theorem for densities

$$h(\theta|x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_n|\theta)g(\theta)}{\int f(x_1, \ldots, x_n|\theta)g(\theta)d\theta}$$

Here's our main formula again:

$$h(\theta|x_1, \dots, x_n) = \frac{f(x_1, \dots, x_n|\theta)g(\theta)}{\int f(x_1, \dots, x_n|\theta)g(\theta)d\theta}$$

$f(x_1, \dots, x_n|\theta)$ is still called the likelihood. But $g(\theta)$ is called the **prior distribution**. The prior distribution is chosen to reflect prior knowledge we have about the parameter before we see our data.

After we see our data, we can compute the **posterior distribution** $h(\theta|x_1, \dots, x_n)$. This is the main thing we're after here. This is the distribution of our unknown quantity after we take into account all possible information.

Last thing before an example...

In the formula

$$h(\theta|x_1,\ldots,x_n) = \frac{f(x_1,\ldots,x_n|\theta)g(\theta)}{\int f(x_1,\ldots,x_n|\theta)g(\theta)d\theta}$$

The denominator $\int f(x_1,\ldots,x_n|\theta)g(\theta)d\theta$ is just a normalizing constant to make our density integrate to 1 (make sure you see why...it isn't a function in θ). Sometimes we don't care about it. Because of this, people doing Bayesian statistics will use "proportional to" symbol a lot like this:

$$h(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)g(\theta)$$

This means that the posterior distribution is proportional to the pointwise product (in θ) between f and g.

Example 14.7

Suppose we have data from a binomial distribution. The only parameter that we are uncertain about is p. Let's let p be a random variable now.

p needs to be in the interval $(0, 1)$. We might use the *Beta* distribution as a prior for p because its support is the interval $(0, 1)$. The parameters to this distribution will be α and β.

$$g(p; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1 - p)^{\beta-1}$$

Example 14.7 continued

Remember how we said we can use the $\propto$ symbol? If we take the product $f(\mathbf{x}|\theta)g(\theta)$ and we can "recognize" the density, then we already know the normalizing constant. This happens in our first example (but not always). And instead of writing θ, we'll write our parameter as p.

Let $X|p \sim \text{Binomial}(n, p)$ and $p \sim \text{Beta}(\alpha, \beta)$.

$$\begin{aligned}
h(p|\mathbf{x}) &\propto f(\mathbf{x}|p)g(p) \\
&= \left[\binom{n}{x} p^x (1-p)^{n-x}\right] \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}\right] \\
&\propto p^x (1-p)^{n-x} p^{\alpha-1}(1-p)^{\beta-1} \\
&= p^{\alpha+x-1}(1-p)^{n+\beta-x-1}
\end{aligned}$$

And this looks like a beta distribution, so

$$h(p|\mathbf{x}) = \text{Beta}(\alpha + x, n + \beta - x)$$

Example 14.7 continued

$$h(p|\mathbf{x}) = \text{Beta}(\alpha + x, n + \beta - x)$$

Using quick formulas for a Beta distribution

$$\begin{aligned}
E[p|x_1, \ldots, x_n] &= \frac{\alpha + x}{(\alpha + x) + (n + \beta - x)} \\
&= \frac{\alpha + x}{\alpha + n + \beta} \\
&= \frac{\alpha}{\alpha + n + \beta} \frac{\alpha + \beta}{\alpha + \beta} + \frac{x}{\alpha + n + \beta} \frac{n}{n} \\
&= \frac{\alpha}{\alpha + \beta} w_1 + \frac{x}{n} w_2
\end{aligned}$$

Where $w_1 = \frac{\alpha+\beta}{\alpha+n+\beta}$ and $w_2 = \frac{n}{\alpha+n+\beta}$. So the posterior average is a weighted average of the prior and likelihood averages.

Example 14.7 continued

Earlier when we were doing frequentist statistics, we did point estimation, confidence intervals, and hypothesis testing. We only used the likelihood function.

Now we have a probability distribution for our parameter. A few things worth mentioning:

1. we can calculate the mode of this distribution (kind of like MLE)
2. we can calculate the mean of this distribution (kind of like MLE too)
3. answer questions like "What's the probability our parameter was less than a half"
4. we don't have to be careful with the distinction between probability and likelihood
5. we can predict new data like this $p(x_{new}|x_{old}) = \int p(x_{new}|\theta)p(\theta|x_{old})d\theta$
6. we took into account a subjective prior distribution to reflect what we already know about p

Example 14.7 continued

Find a Bayesian *credible interval* for p.

$$h(p|\mathbf{x}) = \text{Beta}(\alpha + x, n + \beta - x)$$

Answer...just take appropriate quantiles from the posterior distribution.

Example 14.8

$X_1, \ldots, X_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma_1^2)$ and $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$. Check that we have
$f(x_1, \ldots, x_n|\mu) = (2\pi\sigma_1^2)^{-n/2} \exp\left[-\frac{\sum(x_1-\mu)^2}{2\sigma_1^2}\right]$

$$\begin{aligned}
h(\mu|x_1, \ldots, x_n) &\propto f(x_1, \ldots, x_n|\mu)g(\mu) \\
&= (2\pi\sigma_1^2)^{-n/2} \exp\left[-\frac{\sum(x_1-\mu)^2}{2\sigma_1^2}\right] \times \\
&(2\pi\sigma_0^2)^{-1/2} \exp\left[-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right] \\
&\propto \exp\left[-\frac{1}{2}\left\{\frac{\sum(x_1-\mu)^2}{\sigma_1^2} + \frac{(\mu-\mu_0)^2}{\sigma_0^2}\right\}\right] \\
&= \exp\left[-\frac{1}{2}\left\{\frac{\sum x_1^2 - 2\mu\sum x_i + n\mu^2}{\sigma_1^2} + \frac{(\mu-\mu_0)^2}{\sigma_0^2}\right\}\right]
\end{aligned}$$

Example 14.8 continued

$$
\begin{aligned}
\ldots &= \exp\left[-\frac{1}{2}\left\{\frac{\sum x_1^2 - 2\mu\sum x_i + n\mu^2}{\sigma_1^2} + \frac{(\mu-\mu_0)^2}{\sigma_0^2}\right\}\right] \\
&= \exp\left[-\frac{1}{2}\left\{\frac{\sum x_1^2 - 2\mu\sum x_i + n\mu^2}{\sigma_1^2} + \frac{\mu^2 - 2\mu\mu_0 + \mu_0^2}{\sigma_0^2}\right\}\right] \\
&\propto \exp\left[-\frac{1}{2}\left\{\frac{-2\mu\sum x_i + n\mu^2}{\sigma_1^2} + \frac{\mu^2 - 2\mu\mu_0}{\sigma_0^2}\right\}\right] \\
&= \exp\left[-\frac{1}{2}\left\{\mu^2\left(\frac{n}{\sigma_1^2} + \frac{1}{\sigma_0^2}\right) - 2\mu\left(\frac{\sum x_i}{\sigma_1^2} + \frac{\mu_0}{\sigma_0^2}\right)\right\}\right] \\
&= \exp\left[-\frac{1}{2}\left(\frac{n}{\sigma_1^2} + \frac{1}{\sigma_0^2}\right)\left\{\mu^2 - 2\mu\left(\frac{n}{\sigma_1^2} + \frac{1}{\sigma_0^2}\right)^{-1}\left(\frac{\sum x_i}{\sigma_1^2} + \frac{\mu_0}{\sigma_0^2}\right)\right\}\right] \\
&\propto \exp\left[-\frac{1}{2}\left(\frac{n}{\sigma_1^2} + \frac{1}{\sigma_0^2}\right)\left\{\mu - \left(\frac{n}{\sigma_1^2} + \frac{1}{\sigma_0^2}\right)^{-1}\left(\frac{\sum x_i}{\sigma_1^2} + \frac{\mu_0}{\sigma_0^2}\right)\right\}^2\right]
\end{aligned}
$$

Example 14.8 continued

So $h(\mu|x_1, \ldots, x_n)$ is *still* normally distributed with mean

$$\left(\frac{n}{\sigma_1^2} + \frac{1}{\sigma_0^2}\right)^{-1} \left(\frac{\sum x_i}{\sigma_1^2} + \frac{\mu_0}{\sigma_0^2}\right)$$

and variance

$$\left(\frac{n}{\sigma_1^2} + \frac{1}{\sigma_0^2}\right)^{-1}$$

Summary of Examples

In the examples we did, the posterior distribution is from the sample family as the prior distribution (but it usually has different parameters). This is a very special case; we say here that the prior is **conjugate** to the data distribution. Said differently, the prior is a **conjugate prior** for the likelihood.