Bayesian Stuff. Where To Start

Paul E. Johnson[1] [2]

[1]Department of Political Science

[2]Center for Research Methods and Data Analysis, University of Kansas

2011

Overview

- Study Distributions
- Study Updating
- Study Computing

Bayes Rule. I always forget

- Multiplication rule. The chance that both A and B occur

$$\begin{aligned} P(A \cap B) &= P(A|B)P(B) \\ &= P(B|A)P(A) \end{aligned} \tag{1}$$

- $P(A|B)$ is the probability that A will happen, given B happens.
- Re-arrange expression (1):

$$Bayes\ Rule: \ P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{2}$$

- Note assumes that $P(A)$ is not 0. If $P(A)$ is actually zero, then $P(A \cap B) = 0$

Some Bayes Rule stories

- Game Trees and information revelation
- "Monte Hall Game" story

How Is This Statistical

$$P(\textit{hypothesis}|\textit{obs data}) = \frac{P(\textit{obs data}|\textit{hypothesis})P(\textit{hypothesis})}{P(\textit{obs data})} \quad (3)$$

- $P(\textit{hypothesis})$ is the "prior" probability model.
- $P(\textit{obs data}|\textit{hypothesis})$ is the likelihood of observing a data set if *hypothesis* is equal to a particular value.
- $P(\textit{hypothesis}|\textit{obs data})$ is the "posterior conditional probability" that a particular *hypothesis* is correct, if *obs data* is observed.
- Very common in practice to consider the "maximum likelihood estimate" (particular value of *hypothesis* that maximizes $P(\textit{obs data}|\textit{hypothesis})$ as a reference point

Throw away the denominator

$$P(hypothesis|obs\ data) = \frac{P(obs\ data|hypothesis)P(hypothesis)}{P(obs\ data)} \quad (4)$$

- $P(obs\ data)$ is a proportionality constant, we usually don't worry about it
- The posterior is proportional to

$$P(hypothesis|obs\ data) \propto P(obs\ data|hypothesis)P(hypothesis) \quad (5)$$

- That boils down to

$$\text{posterior distribution} \propto \quad \text{likelihood} \quad \text{times prior}$$

Updating

- You begin project with a "prior"
- You observe some data, make some calculations
- After that, you should be able to adjust your prior

What's so tough about that?

- In ML analysis, you estimate the coefficients $\beta = (\beta_1, \beta_2, \ldots, \beta_m)$. You arrive at estimates $\hat{\beta}$ and estimated variances $Var[\hat{\beta}]$.
- In order to "make that Bayesian", you must
  - specify priors for $\hat{\beta}$ (usually that means you specify a mean and a variance of your belief)
  - specify priors for $Var[\hat{\beta}]$
- After that, you do the work of calculating your posterior beliefs

The problem is "tractability"

Call the likelihood HorriblyUglyFormula!

$$p(y|\theta) = \mathit{HorriblyUglyFormula}(y|\theta)$$

The HorriblyUglyFormula depends on θ, but if you know that value, you can calculate the likelihood.
The posterior is

$$p(\theta|y) = \mathit{prior}(\theta)\mathit{HorriblyUglyFormula}(y|\theta)$$

Just think, you multiply together HorriblyUglyFormula with $p(\theta)$ and you may get something worse than Frankenstein.

Sneaky Math trick to the rescue (conjugate prior)

- Try to choose a formula for the prior $p(\theta)$ So that the Calculation is easier.
- Look for a way to choose $p(\theta)$ so that $p(\theta|y)$ has the same functional form as $p(\theta)$.
- Dialog. You say, "but please, I don't get it. How could it possibly be that HorriblyUglyFormula is tamed by multiplying it against a magical $p(\theta)$?"
- The Conjugate Prior usually ends up "blending together" the sample information with the prior belief parameres.

Conjugate example: Poisson

- Suppose your likelihood is the Poisson,

$$p(y|\lambda) = \frac{e^{-\lambda}\lambda^{y}}{y!} \tag{6}$$

- And suppose you "guess" that a nice conjugate prior for λ would be the Gamma:

$$p(\lambda) = \frac{\lambda^{\alpha-1}e^{-\lambda/\beta}}{\Gamma(\alpha)\beta^{\alpha}} \tag{7}$$

  Or, $\lambda \sim \mathit{Gamma}(\alpha, \beta)$.
- Recall $E[\lambda]$ according to $\mathit{Gamma}(\alpha, \beta) = \alpha\beta$ and the variance is $\alpha\beta^2$

Insert Into Bayes Formula

$$p(\lambda|y) \propto \left(e^{-\lambda}\lambda^{y}\right)\left(\lambda^{\alpha-1}e^{-\lambda/\beta}\right)$$

$$= \lambda^{y+\alpha-1}e^{-\lambda(1+1/\beta)}$$

- Simplify: All those constants of proportionality are "washed out",
- The formula here for the posterior probability of λ is essentially the same formula as the prior, 7. In fact, if you translate the coefficients, it is apparent that

$$\lambda|y \sim G(\alpha + \sum y, \frac{\beta}{N\beta + 1})$$

- The posterior belief you have about the probability is just a "rescaling" of the prior belief you had.

That is Easy Because...

- The Poisson is a one-parameter distribution. No separate Variance Parameter floating about
- Somebody figured out that the Gamma belief on λ is conjugate. Vital!

An "Easy" Normal Example

- Your data is drawn from $N(\mu, \sigma^2)$
- You want to estimate μ.
- Suppose somehow (I don't know how) σ^2 is known

You believe...

- Your Prior belief about μ is characterized

$$\mu \sim Normal(\mu_o, \sigma_o^2) \tag{8}$$

- Here μ_o and σ_0^2 are numbers you "just know"
- You collect a sample of N cases, from which you calculate a mean $\bar{x}$
- The posterior belief about the mean of x will be Normal (Jackman, p. 516)

$$posterior\ \mu|x \sim Normal\left[\frac{\mu_0\sigma_0^{-2} + \bar{x}\frac{N}{\sigma^2}}{\sigma_0^{-2} + \frac{N}{\sigma^2}}, \left(\sigma_0^{-2} + \frac{n}{\sigma^2}\right)^{-1}\right] \tag{9}$$

Insert some "for instance" values

- $\sigma_0^2 = 1$. This simplified the posterior to

$$\textit{posterior}\ \mu|x \sim \textit{Normal}\left[\frac{\mu_0 + \bar{x}\frac{N}{\sigma^2}}{\frac{N}{\sigma^2}}, \left(1 + \frac{N}{\sigma^2}\right)^{-1}\right] \tag{10}$$

$$\textit{updated mean of belief}\ :\ \textit{posterior}\ \mu_0 = \bar{x} + \mu_0 \cdot \frac{\sigma^2}{N} \tag{11}$$

- Recall the variance of the mean by frequentist standards is σ^2/N
- So if N is huge or σ^2 is very small, then almost all of your posterior belief will be based on the data $\bar{x}$

And your posterior uncertainty

$$\left(\sigma_{\mathbf{0}}^{-\mathbf{2}} + \frac{\mathbf{N}}{\sigma^{\mathbf{2}}}\right)^{-\mathbf{1}} \tag{12}$$

- Recall σ^2/N is the variance of the mean of a sample of size N:

$$Var[\bar{x}] = \sigma^2/N \tag{13}$$

- I bet there is a more direct derivation: The posterior uncertainty can be written

$$\begin{aligned}\left(\frac{1}{\sigma_{\mathbf{0}}^{\mathbf{2}}} + \frac{1}{Var[\bar{x}]}\right)^{-\mathbf{1}} &= \frac{1}{\frac{1}{\sigma_0^2} + \frac{1}{Var[\bar{x}]}} = \frac{1}{\frac{Var[\bar{x}]+\sigma_0^2}{\sigma_0^2 Var[\bar{x}]}} \\ &= \frac{\sigma_0^2 Var[\bar{x}]}{\sigma_0^2 + Var[\bar{x}]} = \frac{\sigma_0^2}{1 + \sigma_0^2/Var[\bar{x}]}\end{aligned}$$

And the Posterior Uncertainty "means"...

$$\frac{\sigma_0^2}{1 + \sigma_0^2 / Var[\bar{x}]} \tag{14}$$

- Think of your prior σ_0^2 as fixed, and allow $Var[\bar{x}]$ to vary.
- As $Var[\bar{x}]$ shrinks to 0, the denominator explodes toward infinity, and posterior uncertainty collapses to 0.
  - Meaning: As the sample makes you more and more sure of the location of μ, your uncertainty about your estimate of μ is correspondingly reduced (duh!)
- As $Var[\bar{x}]$ grows toward ∞, the denominator shrinks to 1, and the whole thing tends to σ_0^2
- The main point is that your posterior uncertainty is in $[0, \sigma_0^2]$. It never can be greater than σ_0^2
- If $Var[\bar{x}] = \sigma_0^2$, then the posterior uncertainty is $\sigma_0^2/2$.

If you estimate the variance, well...

- You want to collect data to estimate the mean and the variance of data drawn from $N(\mu, \sigma^2)$
- Prior (conjugate)
- Prior for variance σ^2 is "inverse Gamma"
- Conditional on σ_o^2, the belief about the mean is $Normal(\mu_o, \sigma^2/n_o)$
- for μ, prior is $Normal(\mu_o$