

1: Probability and inference

August 26, 2019

Introduction

First, some notation:

- ① y : observed data (could be vector- or matrix-valued)
- ② θ : parameter (usually a greek letter)
- ③ \tilde{y} : unknown, potentially observable (future?) data
- ④ $X = (x_1, \dots, x_n)$, random or nonrandom covariate or predictor

Introduction

First, some notation:

- ① y : observed data (could be vector- or matrix-valued)
- ② θ : parameter (usually a greek letter)
- ③ \tilde{y} : unknown, potentially observable (future?) data
- ④ $X = (x_1, \dots, x_n)$, random or nonrandom covariate or predictor

Distributions

- ① $p(\theta)$: prior distribution
- ② $p(y \mid \theta)$ sampling/data distribution

Introduction

Goal of statistical inference: estimate unobservable quantities!

- ① potentially observables: $p(\tilde{y} | y)$: (e.g. forecasting, prediction, etc.)
- ② unobservable quantities: $p(\theta | y)$

Bayes' rule

Bayes' rule:

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y)}$$
$$\propto p(y \mid \theta)p(\theta)$$

Bayes' rule

Bayes' rule:

$$p(\theta | y) = \frac{p(y | \theta)p(\theta)}{p(y)} \\ \propto p(y | \theta)p(\theta)$$

or perhaps

$$p(\theta | y, x) = \frac{p(y | x, \theta)p(\theta | x)}{p(y | x)} \\ \propto p(y | x, \theta)p(\theta | x)$$

Bayes' rule

Bayes' rule:

$$p(\theta | y) = \frac{p(y | \theta)p(\theta)}{p(y)} \\ \propto p(y | \theta)p(\theta)$$

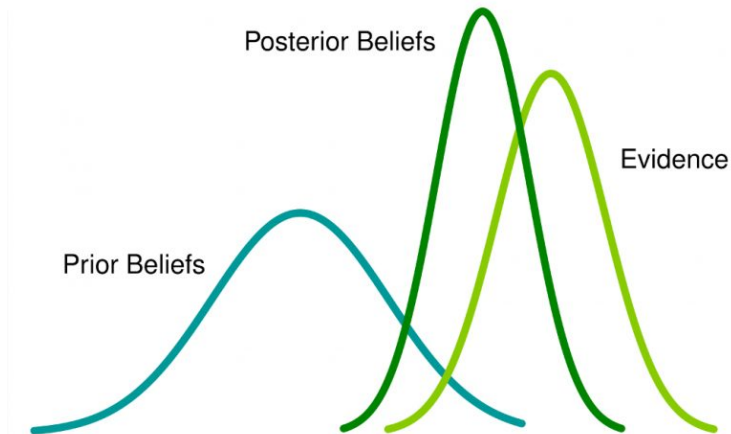
or perhaps

$$p(\theta | y, x) = \frac{p(y | x, \theta)p(\theta | x)}{p(y | x)} \\ \propto p(y | x, \theta)p(\theta | x)$$

- 1 switch/invert order of conditioning!
- 2 think of $p(y | \theta)$, $p(y | x, \theta)$ as a function of θ
- 3 in practice, the normalizing constant is often the most problematic

Bayes' Rule

google's best image:



The **prior predictive distribution**: when you haven't seen any data yet:

$$p(y) = \int p(y \mid \theta) p(\theta) d\theta$$

Prediction

The **prior predictive distribution**: when you haven't seen any data yet:

$$p(y) = \int p(y | \theta) p(\theta) d\theta$$

The **posterior predictive distribution**: when you've seen data

$$\begin{aligned} p(\tilde{y} | y) &= \int p(\tilde{y}, \theta | y) d\theta \\ &= \int p(\tilde{y} | \theta, y) p(\theta | y) d\theta \\ &= \int p(\tilde{y} | \theta) p(\theta | y) d\theta \quad (\text{cond. indep.}) \end{aligned}$$

Both are averages but with different distributions for θ

Posterior odds: $p(\theta_1|y)/p(\theta_2|y)$

Bayes' Rule in terms of posterior odds:

$$\frac{p(\theta_1|y)}{p(\theta_2|y)} = \frac{p(\theta_1)p(y|\theta_1)/p(y)}{p(\theta_2)p(y|\theta_2)/p(y)} = \frac{p(\theta_1)}{p(\theta_2)} \frac{p(y|\theta_1)}{p(y|\theta_2)}$$

Exchangeability

Often $y = (y_1, \dots, y_n)$ are assumed to be **exchangeable**, or

$$p_{Y_1, \dots, Y_n}(y_1, \dots, y_n) = p_{Y_{\sigma(1)}, \dots, Y_{\sigma(n)}}(y_1, \dots, y_n)$$

where σ is any permutation of the indexes.

Exchangeability

Often $y = (y_1, \dots, y_n)$ are assumed to be **exchangeable**, or

$$p_{Y_1, \dots, Y_n}(y_1, \dots, y_n) = p_{Y_{\sigma(1)}, \dots, Y_{\sigma(n)}}(y_1, \dots, y_n)$$

where σ is any permutation of the indexes.

For example, assume Y_1, Y_2 are discrete. Then
 $p(Y_1 = a, Y_2 = b) = p(Y_2 = a, Y_1 = b)$.

Exchangeability

The iid condition implies exchangeability:

$$\begin{aligned} p_{Y_1, \dots, Y_n}(y_1, \dots, y_n) &= \prod_{i=1}^n p_{Y_i}(y_i) && \text{(indep.)} \\ &= \prod_{i=1}^n p_{Y_{\sigma(i)}}(y_i) && \text{(ident.)} \\ &= p_{Y_{\sigma(1)}, \dots, Y_{\sigma(n)}}(y_1, \dots, y_n) \end{aligned}$$

Exchangeability

However, it isn't the other way around. We will often take

$$p(y) = \int p(y \mid \theta) p(\theta) d\theta$$

$$\begin{aligned} p(y) &= p(y_1, \dots, y_n) \\ &= \int p(y_1, \dots, y_n \mid \theta) p(\theta) d\theta \\ &= \int p(y_{\sigma(1)}, \dots, y_{\sigma(n)} \mid \theta) p(\theta) d\theta \\ &= p(y_{\sigma(1)}, \dots, y_{\sigma(n)}) \end{aligned}$$

but $p(y)$ does not factor

Apply the law of total expectation:

$$\underbrace{E[\theta]}_{\text{prior mean}} = E\left[\underbrace{E(\theta | y)}_{\text{posterior mean}} \right]$$

outer expectation on the rhs is taken with respect to $p(y)$.

Apply the law of total variance:

$$\underbrace{\text{var}[\theta]}_{\text{prior variance}} = E[\underbrace{\text{var}(\theta \mid y)}_{\text{posterior var}}] + \underbrace{\text{var}[E(\theta \mid y)]}_{\text{dispersion of post. mean}}$$

outer expectation on the rhs is taken with respect to $p(y)$.

You can also switch things around:

$$E[y] = E[E(y \mid \theta)]$$

and

$$\text{var}(y) = \text{var}[E(y \mid \theta)] + E[\text{var}(y \mid \theta)]$$

Conditional Independence

Conditional independence will be used extensively. X and Y are **conditionally independent given Z** if

$$p(x, y | z) = p(x | z)p(y | z).$$

This is equivalent to a more useful form:

$$p(x | y, z) = p(x | z).$$

Knowing when you are conditioning on redundant variables will help derive a lot of things.

We will be using R

Some bookmarks:

- 1 https://github.com/tbrown122387/stat_6440_slides
- 2 <http://www.stat.columbia.edu/~gelman/book/>
- 3 https://github.com/avehtari/BDA_R_demos
- 4 <http://www.stat.columbia.edu/~gelman/book/data/>