1: Probability and inference

Taylor

University of Virginia

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. $X = (x_1, \ldots, x_n)$, 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. $X = (x_1, \ldots, x_n)$, 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: $p(\tilde{y} \mid y)$: (e.g. forecasting, prediction, etc.)
2. unobservable quantities: $p(\theta \mid y)$

Bayes' rule

Bayes' rule:

$$
\begin{aligned}
p(\theta \mid y) &= \frac{p(y \mid \theta)p(\theta)}{p(y)} \\
&\propto p(y \mid \theta)p(\theta)
\end{aligned}
$$

Bayes' rule

Bayes' rule:

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

or perhaps

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

Bayes' rule

Bayes' rule:

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

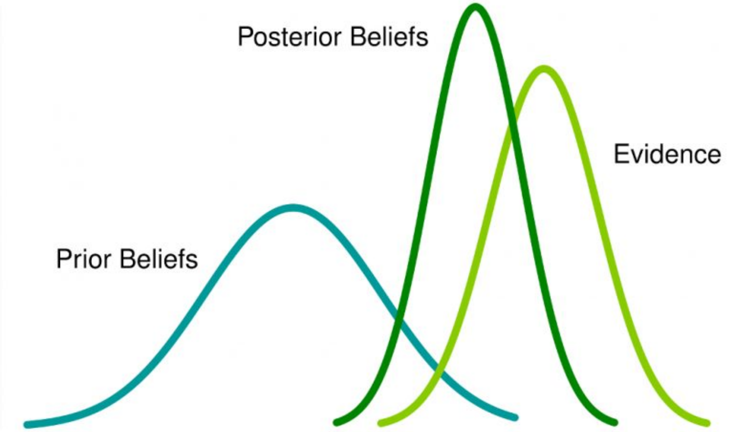
or perhaps

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

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

Bayes' Rule

google's best image:



Prediction

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

$$p(y) = \int p(y \mid \theta)p(\theta)\mathrm{d}\theta$$

Prediction

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

$$p(y) = \int p(y \mid \theta) p(\theta) \mathrm{d}\theta$$

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

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

Both are averages but with different distributions for θ

Exchangeability

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

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

where σ is any permutation of the indexes.

Exchangeability

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

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

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 $p(y) = \int p(y \mid \theta) p(\theta) \mathrm{d}\theta$

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

but $p(y)$ does not factor

LTE and LTV

Apply the law of total expectation:

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

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

LTE and LTV

Apply the law of total variance:

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

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

LTE and LTV

You can also switch things around:

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

and

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

Conditional Independence

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

$$p(x, y \mid z) = p(x \mid z)p(y \mid z).$$

This is equivalent to a more useful form:

$$p(x \mid y, z) = p(x \mid z).$$

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

Computation

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/