MTH225 Fall 2017 Notes

E. Quinn

Contents

CHAPTER 1

Generative Models

Components of a Generative Model

An important feature of the Bayesian paradigm for statistics is that Bayesian models are usually **generative**, which means they specify a way to simulate underlying data generation process.

This allows us to validate the model (remember George Box's famous dictum that all models are wrong) by generating data using the model and comparing it to actual data.

A Bayesian model requires two components:

- The **likelihood**: A collection of probability models for the outcome, indexed by a parameter
- The **prior**: A probability model for our current state of knowledge about the value of the parameter

Characteristics of the generative model for a Bernoulli trial

A generative model for the outcome of a baseball game (say, between the Red Sox and Yankees) can be modeled as:

- A Bernoulli model for the likelihood, that is, a family of Bernoulli distributions indexed by a single parameter θ, with likelihood function $\theta^x(1-\theta)^{1-x}$
- A uniform distribution as the model for our initial state of knowledge about θ (the prior)

The likelihood component. The probability distribution associated with the likelihood, the Bernoulli distribution, corresponds to a probability experiment with two outcomes. In this setting we will denote them as 'R' and 'Y' corresponding to a win by the Red Sox and a win by the Yankees, respectively.

When the experiment is performed (i.e., the game is played),

- With probability θ, the outcome 'R' is observed. The random variable x is assigned a value of 1.
- With probability $1-\theta$, the outcome 'Y' is observed. The random variable x is assigned a value of 0.

The prior component. The prior component is a probability distribution representing our initial state of knowledge about the value of the parameter θ. In this case we are using the uniform distribution as the prior.

The uniform distribution corresponds to the experiment of randomly choosing a number between 0 and 1 in such a way that the probability of the number falling in any subinterval $[a, b]$ is $b-a$.

The data generation process

Once the likelihood and prior(s) are defined, the generative model can produce data by repeating a two-stage process:

- Step 1: Draw a value of θ (a sample of size 1 from the prior distribution)
- Step 2: Draw a value of x from the distribution associated with the likelihood (the Bernoulli distribution)

The R code for simulating a single game:

```
theta = runif(1)
y = rbinom(1,1,theta)
```

To model the behavior over time, we examine a large number of simulated games.

The code to simulate 4,000 games is:

```
theta = runif(4000)
y = rbinom(4000,1,theta)
```

CHAPTER 2

In-class Assignments

Appendix 1: In-class Assignments

Week 3 In-class exercises.

- Exercise 1: Test Stan software `Week3_IC1.Rnw`
- Exercise 2: Estimate the shape parameter of a beta distribution `Week3_IC2.Rnw`
- Exercise 3: Estimate the shape parameter of a beta distribution (part 2) `Week3_IC3.Rnw`

Week 4 In-class exercises.

- Exercise 1: Generative model for Bernoulli distribution `Week4_IC1.Rnw`
- Exercise 2: Generative model for Bernoulli distribution (part 2) `Week4_IC2.Rnw`
- Exercise 3: Computing the posterior distribution fo Bernoulli distribution `Week4_IC3.Rnw`