ETC 2420/5242 Lab 9 2017

Di Cook

Week 9

Purpose

This lab is related to Bayesian thinking, to compute conditional probabilities and posterior probabilities.

Warmup

Watch the video "A visual guide to Bayesian thinking". And also the snippet from "Big Bang Theory".

Exercise 1

A situation where Bayes theorem is routinely used is the spam filter in your mail server. The message is scrutinized for the appearance of key words which make it likely that the message is spam. Let us describe how one of these filters might work. We imagine that the evidence for spam is that the subject message of the mail contains the word "easy". We define events *spam* (the message is spam) and *easy* (the subject line contains this word).

From previous examination of my mail I have learned that 70% of emails are spam, 2% of spam email have "free" in the subject line, and .1% of non-spam emails have this word in the subject line.

Use Bayes Theorem to tackle to find $P(spam|easy)$.

Let A=mail is spam, B=mail has "easy" in subject. What is

a. P(A)?
b. P(B|A)?
c. P(B)?
d. P(A|B)?

Exercise 2

For the dice problem explained in the lecture, write down the problem in tabular format, and specify `P(D)`.

H	P(H)	P(D|H)	P(H)P(D|H)	P(D)	P(H|D)
4					
6					
8					
12					
20					

Exercise 3

Suppose there were 3 20-side dice. Re-do the calculations for `P(H|D)` if a 6 is thrown. What is `P(D)`? Prior? Give an explanation for why the probability of the 20-sided dice increases.

Exercise 4

In the lecture we used a power law prior to dampen down the estimate of the number of locomotives. Instead fit a $U[\mu - 10, \mu + 10]$ prior. Instead fit a $N(\mu, 100)$ prior. Plot the posterior distributions.

Exercise 5

It helps to read the lecture notes to find the equations that you need for this question.

   a. Suppose we have a sample $X_1, \ldots, X_n \sim N(\theta, 9)$.

   - Write down the likelihood function.
   - What is the MLE estimate $\hat{\theta}_{\text{MLE}}$, theoretically? (If we actually had a sample, we could use the function `fitdistr` from the `MASS` library to estimate it.)
   - If we have reason to believe that θ is not a fixed value, but can vary according to $\theta \sim N(\mu, \tau^2)$, what is the posterior distribution, $\pi(\theta|x_1, \ldots, x_n)$?
   - What is the posterior mean $E[\theta|x_1, \ldots, x_n]$?

   b. Suppose the "true" value is $\theta = 2$. Consider (1) $\mu = 5$ and $\tau = 1$, and (2) $\mu = 2$ and $\tau = 2$. For $n \in \{1, 10, 20, 50, 100, 10000\}$. Use Nat's shiny app for this.

   - Simulate a data set consisting of n observations
   - Plot on the same graphic $\pi(\theta)$, $\pi(\theta|x_1, \ldots, x_n)$ and $\hat{\theta}_{\text{MLE}}$.

   c. Discuss the behavior of $\pi(\theta|x_1, \ldots, x_n)$ as n increases and the impact of the prior distribution.

Exercise 6

Suppose your prior distribution for θ, the proportion of Australians who will vote yes on the Marriage Equality Postal Survey, is beta with mean 0.5 and standard deviation 0.4.

   a. What are the parameters α and β of your prior distribution. Plot the prior density function.
   b. Why might it be reasonable to consider θ in a Bayesian way?
   c. A random sample of 1000 Australians is taken, and 65% say they will vote "yes". What are the posterior mean and variance for θ? Plot the posterior density function. (Note that the mean and variance of a beta distribution (`Beta(a, b)`), are given by $\frac{a}{a+b}$ and $\frac{ab}{(a+b)^2(a+b+1)}$. The posterior distribution for a binomial with beta prior is `Beta(a+x, b+n-x)` where `x` is the observed number of successes and `n` is the number of observations.)

TURN IN

   - Your `.Rmd` file
   - Your `html` file that results from knitting the Rmd.