Stat238: Lab 1

August 25, 2016

Comments:

- A writeup of the first problem will be due as part of the first problem set. I suggest you write it up during the lab period or shortly afterwards.
- Labs are intended to allow for hands-on work with a partner or in groups to work through problems. For problems on problem sets that were presented in lab, you may present a group answer.
- Note: I quite like the spelling correction example in BDA Section 1.4 as another example of the use of Bayes theorem in the real world, with both the likelihood and prior playing important roles.

Problems

1. Consider a problem akin to the Bayes theorem diagnostic testing example I alluded to in Class 1 on Wednesday. Most of you will have seen and possibly taught an example like this. The goal here is to revisit such a problem in the context of explicitly thinking about prior probabilities and the likelihood. Here Y is the data (a 0/1 random variable) and D is the state of the world/system (also a 0/1 random variable).

   Suppose the police have DNA evidence recovered from a crime scene. The police test a group of people and find one who matches. What is the posterior probability that the DNA is from that person, $P(D = 1|Y = 1)$? Suppose the sensitivity of the test is $P(Y = 1|D = 1) = 0.999$, and one minus the specificity is $P(Y = 1|D = 0) = 0.0001$, i.e, the probability of false positive matches is 0.999 and that of false negative matches is 0.0001. I.e., the test is very accurate.

   $$P(D = 1 \mid Y = 1) =$$

   $$\frac{\overbrace{P(Y = 1|D = 1)}^{\text{Sensitivity}} P(D = 1)}{P(Y = 1|D = 1)P(D = 1) + \underbrace{P(Y = 1|D = 0)}_{\text{1-Specificity}} P(D = 0)} = \frac{P(Y = 1|D = 1)P(D = 1)}{P(Y = 1)}$$

   (a) Suppose the crime takes place on a small island (or if you prefer, in a train car on the Orient Express as in the Agatha Christie mystery) with a small population of people, say 100, all of whom are tested. Alternatively suppose the crime takes place in a large city such as New York City and there is DNA testing of a million people. How would your priors differ and how does that affect the answer?

(b) Suppose you want to robustify your analysis and consider the possibility that there is 50% chance the culprit is someone not from the island or not in the city. How do you change the analysis and how does the result change? How much would your prior probability on the culprit not being one of the islanders have to change for the final conclusion to change in that scenario?

(c) Write a short paragraph explaining in lay language (but potentially with some numbers as part of an example) why a jury or judge in the New York City case should be wary of interpreting the accuracy of the test as an indication of guilt of the person with the match.

2. (if you have time) Suppose we wanted to write the problem out in terms of (binary) random variables and find the posterior distribution, $p(D|Y)$? We'll use the lower case $p()$ to indicate the probability mass function whereas we used $P()$ above to indicate probabilities.

   (a) Write out the prior for $D \in \{0,1\}$ as a Bernoulli distribution.

   (b) Write the likelihood for $Y \in \{0,1\}$ as a Bernoulli distribution. You'll probably need to have the probability for the Bernoulli be a sum of two terms.

   (c) Use Bayes theorem for distributions,

   $$p(D|Y) = \frac{p(Y|D)p(D)}{p(Y)},$$

   and find the posterior for this problem. Notice that the answer for $p(D = 1|Y = 1)$ is the same as when using Bayes theorem with probabilities, $P(D = 1|Y = 1)$. That's reassuring; otherwise we should all probably spend our time on other pursuits this semester.