Illapani_Homework_HW9

Srini Illapani

December 10, 2015

Bayesian Data Analysis

2.1

Exercise 2.1. [Purpose: To get you actively manipulating mathematical models of probabilities.] Suppose we have a four-sided die from a board game. On a tetrahedral die, each face is an equilateral triangle. When you roll the die, it lands with one face down and the other three faces visible as a three-sided pyramid. The faces are numbered 1-4, with the value of the bottom face printed (as clustered dots) at the bottom edges of all three visible faces. Denote the value of the bottom face as x. Consider the following three mathematical descriptions of the probabilities of x. Model A: $p(x) = 1/4$. Model B: $p(x) = x/10$. Model C: $p(x) = 12/(25x)$. For each model, determine the value of $p(x)$ for each value of x. Describe in words what kind of bias (or lack of bias) is expressed by each model.

Model A:

p(x=1) = 1/4, p(x=2) = 1/4, p(x=3) = 1/4, p(x=4) = 1/4

This model has no bias, all values of x are equally likely.

Model B:

p(x=1) = 1/10. p(x=2) = 2/10, p(x=3) = 3/10, p(x=4) = 4/10

This model is biased toward higher values of x.

Model C:

p(x=1) = 12/25 = 144/300, p(x=2) = 12/50 = 72/300, p(x=3) = 12/75 = 48/300, p(x=4) = 12/100 = 36/300

This model is biased toward lower values of x.

5.1

Exercise 5.1. [Purpose: Iterative application of Bayes' rule, and seeing how posterior probabilities change with inclusion of more data.] This exercise extends the ideas of Table 5.4, so at this time, please review Table 5.4 and its discussion in the text. Suppose that the same randomly selected person as in Table 5.4 gets re-tested after the first test result was positive, and on the re-test, the result is negative. When taking into account the results of both tests, what is the probability that the person has the disease? *Hint:* For the prior probability of the re-test, use the posterior computed from the Table 5.4. Retain as many decimal places as possible, as rounding can have a surprisingly big effect on the results. One way to avoid unnecessary rounding is to do the calculations in R.

```
# Specify hit rate of test:
pPositiveGivenDisease = 0.99
# Specify false alarm rate of test:
pPositiveGivenNoDisease = 0.05

# Specify the original prior:
pDisease = 0.001

# Bayes rule for first, positive test:
pDiseaseGivenPositive = ( pPositiveGivenDisease * pDisease /
+ ( pPositiveGivenDisease * pDisease
+ + pPositiveGivenNoDisease * (1.0-pDisease) ) )

show(pDiseaseGivenPositive)
```

```
## [1] 0.01943463
```

```
# Set the prior to the new probability of having the disease:
pDisease = pDiseaseGivenPositive

# Bayes rule for second, negative test:
pDiseaseGivenNegative = ( (1.0-pPositiveGivenDisease) * pDisease /
+ ( (1.0-pPositiveGivenDisease) * pDisease
+ + (1.0-pPositiveGivenNoDisease) * (1.0-pDisease) ) )

show(pDiseaseGivenNegative)
```

```
## [1] 0.0002085862
```

5.2

Exercise 5.2. [Purpose: Getting an intuition for the previous results by using "natural frequency" and "Markov" representations]

(A) Suppose that the population consists of 100,000 people. Compute how many people would be expected to fall into each cell of Table 5.4. To compute the expected frequency of people in a cell, just multiply the cell probability by the size of the population. To get you started, a few of the cells of the frequency table are filled in here:

	$\theta = \ddot{\frown}$	$\theta = \ddot{\smile}$	
$D = +$	$\text{freq}(D=+, \theta=\ddot{\frown})$ $= p(D=+, \theta=\ddot{\frown})\,N$ $= p(D=+ \mid \theta=\ddot{\frown})\,p(\theta=\ddot{\frown})\,N$ $= 99$	$\text{freq}(D=+, \theta=\ddot{\smile})$ $= p(D=+, \theta=\ddot{\smile})\,N$ $= p(D=+ \mid \theta=\ddot{\smile})\,p(\theta=\ddot{\smile})\,N$ $=$	$\text{freq}(D=+)$ $= p(D=+)\,N$ $=$
$D = -$	$\text{freq}(D=-, \theta=\ddot{\frown})$ $= p(D=-, \theta=\ddot{\frown})\,N$ $= p(D=- \mid \theta=\ddot{\frown})\,p(\theta=\ddot{\frown})\,N$ $= 1$	$\text{freq}(D=-, \theta=\ddot{\smile})$ $= p(D=-, \theta=\ddot{\smile})\,N$ $= p(D=- \mid \theta=\ddot{\smile})\,p(\theta=\ddot{\smile})\,N$ $=$	$\text{freq}(D=-)$ $= p(D=-)\,N$ $=$
	$\text{freq}(\theta = \ddot{\frown})$ $= p(\theta = \ddot{\frown})\,N$ $= 100$	$\text{freq}(\theta = \ddot{\smile})$ $= p(\theta = \ddot{\smile})\,N$ $= 99,900$	N $= 100,000$

Notice the frequencies on the lower margin of the table. They indicate that out of 100,000 people, only 100 have the disease, while 99,900 do not have the disease. These marginal frequencies instantiate the prior probability that $p(\theta = \ddot{\frown}) = 0.001$. Notice also the cell frequencies in the column $\theta = \ddot{\frown}$, which indicate that of 100 people with the disease, 99 have a positive test result and 1 has a negative test result. These cell frequencies instantiate the hit rate of 0.99. Your job for this part of the exercise is to fill in the frequencies of the remaining cells of the table.

D=+|0.99 * 0.001 * 100,000 = 99 |0.05 * (1.0-0.001) * 100,000 = 4,995 | 5,094
D=-|(1.0-0.99) * 0.001 * 100,000 = 1 |(1.0-0.05) * (1.0-0.001) * 100,000 = 94,905 | 94,906
0.001 * 100,000 = 100 |(1.0-0.001) * 100,000 = 99,900 | 100,000

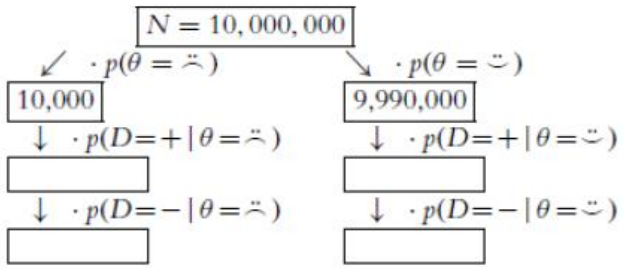
(B) Take a good look at the frequencies in the table you just computed for the previous part. These are the so-called "natural frequencies" of the events, as opposed to the somewhat unintuitive expression in terms of conditional probabilities (Gigerenzer & Hoffrage, 1995). From the cell frequencies alone, determine the proportion of people who have the disease, given that their test result is positive. Before computing the exact answer arithmetically, first give a rough intuitive answer merely by looking at the relative frequencies in the row $D = +$. Does your intuitive answer match the intuitive answer you provided when originally reading about Table 5.4? Probably not. Your intuitive answer here is probably much closer to the correct answer. Now compute the exact answer arithmetically. It should match the result from applying Bayes' rule in Table 5.4.

```
d <- 99/5094
d

## [1] 0.01943463
```

This matches what Bayes' rule provided.

(C) Now we'll consider a related representation of the probabilities in terms of natural frequencies, which is especially useful when we accumulate more data. This type of representation is called a "Markov" representation by Krauss, Martignon, and Hoffrage (1999). Suppose now we start with a population of $N = 10,000,000$ people. We expect 99.9% of them (i.e., 9,990,000) not to have the disease, and just 0.1% (i.e., 10,000) to have the disease. Now consider how many people we expect to test positive. Of the 10,000 people who have the disease, 99%, (i.e., 9,900) will be expected to test positive. Of the 9,990,000 people who do not have the disease, 5% (i.e., 499,500) will be expected to test positive. Now consider re-testing everyone who has tested positive on the first test. How many of them are expected to show a negative result on the retest? Use this diagram to compute your answer:



When computing the frequencies for the empty boxes above, be careful to use the proper conditional probabilities!

Left Branch of the tree:

10,000 * 0.99 = 9,900
9,900 * (1.0-0.99) = 99

Right Branch of the tree:

9,990,000 * 0.05 = 499,500
499,500 * (1.0-0.05) = 474,525

(D) Use the diagram in the previous part to answer this: What proportion of people, who test positive at first and then negative on retest, actually have the disease? In other words, of the total number of people at the bottom of the diagram in the previous part (those are the people who tested positive then negative), what proportion of them are in the left branch of the tree? *How does the result compare with your answer to Exercise 5.1?*

Result:

The proportion in the left branch of the tree is 99/(99 + 474,525) = 0.0002085862, this matches the result of Exercise 5.1.