Introduction to Bayesian Data Analysis

Assignment 2

Linus Hof

Info

The assignment comprises of 10 tasks, each giving 2 points (correct approaches and form give points). In total, 20 points can be obtained.

By the end of the deadline (1 June, 8pm), upload one `R` script that produces the answers to Moodle:

- Indicate which task the code belongs to by adding a line `#tasknumber` over the code for the respecitve task.
- `# comment` your code by describe what the code is supposed to do.
- At the beginning of the script, specify which packages you used (`library(package)`)

Bayesian Updating

1. Suppose there are 3 companies, Company A to C. Company A has a customer satisfaction rate of .70, Company B of .50, and Company C of .80. Suppose that you receive 10 customer reviews on the same company, but you don't know on which company. However, based on your past experience, you think it is twice as likely that the reviews belong to Company B compared to Company A and also compared to Company C. 6 reviews state that the customer is satisfied, 4 state that they are not. Show that the posterior probability that the reviews are on Company A, conditional on seeing 6 positive and 4 negative ratings, is 0.29.

   2 points

2. Continuing the previous example, suppose that you received 10 more reviews, 9 positive and 1 negative. Show that the posterior probability that Company C received the reviews increases by approximately 33 percentage points, when considering all 20 rather than only the first 10 reviews. To obtain the updated posterior, compute the likelihood of the 10 most recent reviews only (i.e., don't include the first 10 reviews again).

2 points

3. Suppose there are two factories, Factory A and Factory B, producing the same product. The company C receives equally many shipments from both factories. Even though the machines, processes, and standards are virtually identical, the factories differ in their defect rates. A shipment from Factory A entails defective products with 10% of the time, while a shipment from Factory B entails defective products 20% of the time. Assume these defect rates are known with certainty based on historical company data. Now, you receive a shipment from one of the factories, and upon inspection, you find that the shipment contains defective products. What is the probability that the next shipment from the company will also contain defective products?

   2 points

5. Bayesian inference makes it easy to use all of the data, even if the data are of different types. So suppose now that the R&D department came up with a Machine Learning algorithm that can identify the Factory a shipment comes from based on product features. But the classification algorithm is imperfect. This is the information you have about the algorithm:

   - The probability it correctly identifies a Factory A product is 0.8.
   - The probability it correctly identifies a Factory B product is 0.65.

   An R&D employee administers the algorithm to the shipped products and tells you that the test is positive for Factory A. First ignore the information on defected products and compute the posterior probability that the shipment is from Factory A given the output of the algorithm. Redo your calculation, using both information, the fact that there are defected products and the fact that the algorithm is positive for Factory A.

   2 points

Bayesian Workflow

Suppose you want to estimate the proportions of land on the earth's surface.

6. Define and plot (visualize) the prior distribution over all possible proportions of land. Justify (briefly explain) your choice of the prior based on what you know about the earth's surface.

   Note: When using a continuous distribution, the densities can be larger than 1. Don't be bothered by this. The area under the curve is still 1.

   2 points

7. Produce 10,000 random samples from your prior distribution and store them in a vector `sample`.

Suppose I throw a globe repeatedly into the air and catch it. After each catch, I note down whether my right index finger landed on land or water. After 100 tosses, the index finger landed on land 26 times.

8. Run the following code chunk that uses your object `sample` to obtain prior probabilities for the possible proportions of land 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1that approximate your prior distribution. Use these priors to compute the posterior probabilities, given the 100 globe tosses.

   2 points

```
prop <- seq(0, 1, length.out = 12)
priors <- vector("numeric", length(prop))
for (i in seq_along(prop)){

 priors[i] <- round( sum(sample >= prop[i] & sample < prop[i+1]) / 1e4 , 2)

}

poss <- tibble(prop_L = seq(0, 1, .1),
                  prior = priors[1:11])
```

9. Take 1,000 samples from the computed posterior distribution using the `sample()` function.

   2 points

10. Use each sampled value to predict the outcomes of 100 globe tosses (the number of times the index finger lands on land in 100 tosses). Plot the posterior predictive distribution for the number of lands.

    2 points