Ec 5043 Prof. Le Wang
Managerial Economics II

Homework #4

March 3, 2020

Instruction: Do all the following empirical exercises using R. Turn in your answer with tables and graphs, if any, (along with your program and output files appended at the end of document). Refer to the output file whenever appropriate when discussing your results.

Note that for all simulation exercises, set the seed number to 123456 to ensure the reproducibility of your results.

1. Question 1. [Statistical Language Model]

Use the **austen_books()** command as in class to download the full texts of Jane Austen's 6 completed, published novels.

1. Generate the trigram dataset using the books downloaded. **And do not filter any stop words.** Save the dataset as **austen.trigram**.
2. Using your trigram data, evaluate the probability of observing the sentence "**The late owner of this estate was a single man**". Unlike the first-order Markov Chain assumption in class, invoke the assumption of second order Markov Chain. In other words

$$\Pr[W_n \mid W_{n-1}, W_{n-2}, \dots] = \Pr[W_n \mid W_{n-1}, W_{n-2}]$$

3. Using the trigram data, suppose that the first two words given to you are "John Dashwood". Generate next two words. Show your code.

2. Question 2. [Interpreting the Test Results (I)]

Say you wake up with spots all over your face. You rush to the doctor and he says that 90 percent of the people (!) who have smallpox have the symptoms you have. Since smallpox is often fatal, your first inclination may be to panic. But your doctor would tell you that the probability you have smallpox is only 1.1 percent (or 0.011), so you do not have to panic. Your doctor's statements seem contradictory to each other. How to reconcile it? Is your doctoring just comforting you? Or, is there any scientific reason to explain the seemingly contradictory statements that you doctor made? This is sort-of an essay question, but you can definitely give some numerical examples to support your argument. **This question corresponds to Bayes' Rule in our slides.**

3. Question 3. [Interpreting the Test Results (II)]

A well-known surprising result is called ***Prosecutor's Fallacy***: Suppose that we know what is the probability a person is telling the truth given the results of a Polygraph test. Let a positive reading on the Polygraph be denoted by $+$, and a negative reading be denoted by $-$. T denotes the person is telling truth and L denotes the person is lying. If we believe that $\Pr[T] = 0.99$, say, and know that

Pr[Positive Reading | L] = 0.88 and Pr[Positive Reading | T] = .14 from lab work. Suppose that we get a positive readout, what is the probability of someone telling the truth?

$$\Pr[T \mid \text{Positive Reading}] =?$$

4. Question 4. [Bayesian Estimation]

Let's practice what we learned about Bayesian estimation. Suppose that we would like to know whether or not a coin is biased, $\theta = \Pr[H]$. We observed the following data (**Note that the data are different from what we used in class**)

$$D = (9H, 3T)$$

1. Step 1. Let's assume the following prior ($f(\theta)$): we now believe that four parameters are possible $\theta = .15, .25, .5, , .75$ and $p(\theta = .15) = .3, p(\theta = .25) = .2, p(\theta = .5) = .3, p(\theta = .75) = .2$
2. Step 2. Now let's calculate the likelihood (the probability of coming up with a H is θ, then that of T is $1 - \theta$):
   (a) What is the general formula for $f(x|\theta)$?
   (b) Now calculate $f(x|0.15), f(x|0.25), f(x|0.5), f(x|0.75)$
3. Step 3. Now let's calculate $f(x)$ using the law of total probability
4. Step 4. Calculate the posterior distribution:
   (a) $f(\theta_1 = .15|x) = \frac{f(x|\theta_1) \cdot f(\theta_1)}{f(x)} =$
   (b) $f(\theta_2 = .25|x) = \frac{f(x|\theta_2) \cdot f(\theta_2)}{f(x)} =$
   (c) $f(\theta_3 = .50|x) = \frac{f(x|\theta_3) \cdot f(\theta_3)}{f(x)} =$
   (d) $f(\theta_4 = .75|x) = \frac{f(x|\theta_4) \cdot f(\theta_4)}{f(x)} =$
5. Now, can you intuitively explain why the posterior distribution changes the way it does? In other words, why, after observing the data we have, we update the prior distribution as we do? **This question is corresponding to the slides on Bayesian Estimation/Inference (I have posted on Canvas as well).**
6. Verify your results in R.

Below are practice questions and you do not need to submit the answers to these two questions.

5. Question 5. [Bayes' Rule and Hiring Decisions]

Hiring decisions are an important part of the managerial decisions. Despite its importance, people often rely on their intuition to hire, but not necessarily the past data to evaluate the optimality of the decision. For example, one may often rely heavily on the interview performance of a candidate. However, the value of an interview is often overstated. Consider the following data from one company of 1000 employees; 800 of them are qualified workers, while 200 of them are not. Among qualified workers, 90 percent interviewed well. Among unqualified workers, 20 percent interviewed well. Now, based on this data, what is the probability of a job candidate who interviewed well being a qualified worker? Note that this is an incredibly naive example of actual practices, and that many assumptions are imposed on the data, but it at least gives us some sense of how things should be done. **This question corresponds to Bayes' Rule in our slides.**

6. Question 3. [Interpreting the Medical Test (II)]

A well-known story of the *Phyysicists's Twins:* Thanks to onograms, a physicist found out she was going to have **twin boys**. "What is the probability my twins will be ***Identical***, rather than ***Fraternal?***" she asked. The doctor answered that one-third of twin births were Identicals, and two-thirds Fraternals. A crucial fact is that identical twins are always same-sex while fraternals have probability 0.5 of same or different. Apply Bayes' rule to answer the physicist's question. **This question corresponds to Bayes' Rule in our slides.**