Ec 5043 Prof. Le Wang
Managerial Economics II

Homework #3

February 19, 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. [Classification and Bayes Classifier]

Use the **Default** data in the ISLR package used in class.

1. Discretize the **balance** variable into four groups and called it **balance.group**: 1 if balance is less than 481.7, 2 if balance is betwen 481.7 and 823.6, 3 if balance is between 823.6 and 1166.3, and 4 if balance is greater than 1166.3.
2. Calculate the following conditional distribution

$$\Pr[\text{default} \mid \text{student} = \text{Yes}, \text{balance} = 3]$$

   (a) Use the subset approach
   (b) Obtain conditional distribution using the joint distribution of all three variables and the marginal distribution of the predictors.

3. What is your classification or prediction for default when someone is a student with the balance level of 3?
4. Let's use **naive_bayes()** command and **predict()** command to generate a prediction. Note that here we have more than two predictors, the Naive Bayes classifier imposes additional assumptions. But it is okay. I just want to see if you know how to run the command and create a new dataset with information that someone is a student with the balance level of 3.

2. Question 2. [Prediction with more than one discrete predictors]

Use the data **titanic.csv** on Github for this question. The data contain all the information on Titanic passengers:

1. **Survived** Survival (0 = No; 1 = Yes)
2. **pclass** Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
3. **name** Name
4. **sex** Sex
5. **age** Age
6. **sibsp** Number of Siblings/Spouses Aboard
7. **parch** Number of Parents/Children Aboard

8. **ticket** Ticket Number
9. **fare** Passenger Fare
10. **cabin** Cabin
11. **embarked** Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

Answer the following questions:

1. The prediction using these predictors, Pr[Survival|Pclass, Sex], is simply a function of Pclass,Sex. Use R to calculate survival predictions for each combination of Pclass,Sex values.
   (a) Write out this function using your answer in (1) Pr[Survival=1|Pclass, Sex]] and Pr[Survival = 0|Pclass, Sex]for every possible combination of predictors.
   (b) Write out this function in ONE equation using your answer in 1.a with the help of an indicator function, as we did in class.
   (c) Use the **-aggregate()-** function to calculate the conditional distribution.

3. Question 3. [Prediction with more than one discrete predictors]

Use the **airbnb.csv** data on Canvas. This dataset contains (fake) historical data on various hosts' decisions for requests with different check-in gaps (as discussed in class). Answer the following questions:

1. If a customer requests a room with a check-in gap of 11 days, is it likely for this person to get his request accepted? Why?

2. Suppose that a customer submitted his request. Coincidentally, every host in your dataset has a check-in gap of 2 days for this particular request. Which host(s) should you recommend to this customer in order to maximize the chance that this recommended host will accept this request?

4. Question 4. [Conditional Distribution, Transition Matrix, and Income Mobility]

As discussed above, policymakers and economists are often interested in measuring and understanding income inequality in society. However, as discussed in class, a snapshot of income dispersion in a given time period t, which ignores potential movement over time, does not necessarily give us a good sense of income inequality. A better measure would be to take into account income dynamics for each individual and capture how his/her income moves up and down over his/her life cycle. Transition matrix serves this purpose. Specifically, transition matrix is nothing but a variant of conditional distribution, with each cell being

$$\Pr[Y_t|Y_{t-1}]$$

where Y_t is one's income class in t (defined as lower, middle, and upper classes;); and Y_{t-1} is one's income class in $t-1$. Let's use the following simulated data (remember to set the seed as instructed above). `y.tm1` and `y.t` are incomes in time periods $t-1$ and t, respectively. **This question is corresponding to the definition of various concepts of joint, marginal and conditional distributions. It is also corresponding to the application to income mobility, as discussed in class.**

```
library(MASS)
Sigma <- matrix(c(10,3,3,2),2,2)
data<-mvrnorm(n=1000, rep(0, 2), Sigma)
y.tm1<- data[,1]
y.t<- data[,2]
```

1. Discretize the continuous income variables into three categories. Lower class is defined as those with income less than or equal to the first quartile of the income distribution in each period. Middle class is defined as those with income more than the first quartile but less than or equal to the third quartile of the income distribution. Upper class is the rest of the population. **Here I would like you to learn how you can generate groups based on continuous variables. This provides a very convenient way to approximate continuous variables in pratice.**
2. Use R to calculate the joint distribution
3. Use R to calculate the marginal distribution
4. Use R to calculate the conditional distribution, that is, the transition matrix.
5. Is there any income mobility in this society?