

Homework 3

BUAN 6356

Read the instructions below before you start your analysis.

1. Create an R Markdown document to prepare your answers. You should upload **two (2)** files on eLearning: (i) an **.RMD** file; and (ii) a **.PDF** file that is generated using “knit” in the .RMD file. Both of these files should contain the required R code, R tables and charts, and all the required explanations and answers to the questions in the homework.
2. Write the answers outside of the code chunks in the markdown file.
3. Include your last name followed by your first initial in the name of the file you upload. For example, if your name is Elon Mask, name the file BUAN6357_Homework1_MuskE.
4. **DO NOT** use an absolute directory path. I should be able to “knit” your R Markdown document to an .html/.pdf document without trying to find the input data in another directory. Test the “knit” process before uploading files on eLearning.
5. Import the relevant data set directly from the website. **DO NOT** change the dataset name before importing it into R. If you rename the dataset or any variable(s), use your R script to do that.
6. Use **set.seed(42)** anywhere you need to set a seed.
7. Label the charts and/or tables appropriately. Your reader should be able to figure out what information a chart is providing by looking at the chart title and its labels.
8. Any assignment submitted after the deadline will be considered late and will not be graded.

Homework 3

A team collected data on email messages to create a classifier that can separate spam from non-spam email messages.

The dataset and short descriptions about the variables in the dataset are available from the following data archive: <https://archive.ics.uci.edu/ml/datasets/spambase>.

Before running LDA, partition the data into training and validation sets by allocating 80 percent of the observations to the training dataset and 20 percent of the observations to the validation dataset. Also, standardize the data set using the *preProcess()* function from the *Caret* package.

1. Examine how each predictor differs between the spam and non-spam e-mails by comparing the spam-class average and non-spam-class average. Identify 10 predictors for which the difference between the spam-class average and the non-spam class average is the highest. Standardize the data set prior to running this analysis.
2. Perform a linear discriminant analysis using the training dataset. Include only 10 predictors identified in the question above in the model.
3. What are the prior probabilities?
4. What are the coefficients of linear discriminants? Explain your answer.
5. Generate linear discriminants using your analysis. How are they used in classifying spams and non-spams?
6. How many linear discriminants are in the model? Why?
7. Generate LDA plot using the training and validation data. What information is presented in these plots? How are they different?
8. Generate the relevant confusion matrix. Calculate and report precision and recall of this model. Explain what those two numbers imply about the performance of this model.
9. Generate lift and decile charts for the validation dataset. Discuss the effectiveness of the model in identifying spams.
10. Does a LDA model that uses a probability threshold of 0.3 to classify spam perform better than the model in Question 2? Explain your answer.