BIF524/CSC498H Data Mining Fall 2021

Midterm Exam -1-

General guidelines: This is the practical part of the midterm exam.

You have three hours to complete it. **The deadline is strictly on Saturday, October 16th, at 12:01pm.**

Use **R Markdown** to show your work in detail. Include comments for every step.

Upload the final report as a **pdf** file to this form. You may submit each task as a separate file.

Each one of you is randomly assigned a dataset for this exam. **Download the dataset that has your name** from the OneDrive folder called “Midterm exam_datasets”. The variable names are the same across datasets, but the data and content are different.

Use ***ggplot*** for all plots.

Good Luck!

Task I:

The response variable in this task is *resp1*. Consider *resp2* as other variables in this case.

1- Load your dataset into R. Then, use suitable functions to explore and summarize the variables. (2pts)

2- Check whether the dataset includes missing values. (1pt)

3- Check the pairwise correlation between *resp1* and all quantitative variables. Are there obvious trends that you can see? If yes, what are they? (2pts)

4- Plot *Var11* against *resp1*, group by *Var11*. Comment on the output. (2pts)

5- Randomly split your data into training (70%) and testing (30%) subsets. Generate a simple linear regression model with *resp1* as response and *Var11* as predictor. Comment on the model summary. Calculate the test error based on predictions of the response for instances in the testing subset. (4pts)

6- Use a leave-one-out cross-validation approach to generate a multiple linear regression model with *resp1* as response and all other variables as predictors. Comment on the model summary. Calculate the test error based on predictions of the response for instances in the testing subset. (4pts)

7- Use a 10-fold cross-validation approach to generate a multiple linear regression model with *resp1* as response and *Var5*, *Var6*, and *Var7* as predictors, while considering possible interaction among them. Calculate the test error based on predictions of the response for instances in the testing subset. (4pts)

8- For a polynomial regression model, use a validation set approach (training 80%, testing 20%) with *resp1* as response and *Var2* as predictor. Generate a quadratic, a cubic, and a quartic model. Which one is better? Show all steps of your answer. – define a function to generate and compare all three models at once, for a full grade on this question. (5pts)

Task II:

The response variable in this task is *resp2*. Consider *resp1* as other variables in this case.

1- While handling *resp2* as categorical, use a 5-fold cross-validation to generate a logistic regression model with *Var2* and *Var10* as predictors. Calculate and comment on the model prediction error. (4pts)

2- While handling *resp2* as categorical, use a 5-fold cross-validation to generate a linear discriminant analysis model with *resp2* as response and *Var2* and *Var10* as predictors. What are the model's sensitivity, specificity, and accuracy? (4pts)

3- Repeat question 2 but with a quadratic discriminant analysis model. What are the model's sensitivity, specificity, and accuracy? (4pts)

4- What is a suitable way to compare the models generated in the previous two questions. Use proper functions and visualizations. Which model is better? (4pts)