Math 4005 Final Project

Project Due: Midnight Dec 13, 2017 .
(*All works must be done independently to receive any credit. No late submission will be accepted.)*

The final project worths 200 points. You again have the wine quality dataset, "winequality_red.csv". Recall that the response in this dataset is the variable *quality* which is an integer value score between 0 and 10 on the taste of the wines. There are eleven physicochemical variables as potential predictors in the dataset. The study was to investigate how these variables are related to the quality of the wines. In this project, please perform the following tasks and report your work using R Notebook (in HTML format) containing all the discussions, the R codes and the outputs of the codes.

1. In this problem, you are asked to use the logistic regression to analyze the data.
   (a) Create a new data frame in which the original response variable *quality* is replaced with a new binary response variable which takes value 1 (the wine quality is high) whenever *quality* has a value equal or greater than 6 and value 0 (the wine quality is low) otherwise.
   (b) Find a best logistic regression model to fit the data with the new data obtained in (a). Give a discussion on the outputs of the computation.
   (c) Give an estimated formula for the probability that the wine quality is high, as a function of the values of the predictors, and demonstrate how you can use this formula to predict wine quality based on new values of the predictors.
   (d) Use bootstrap method to find 95% confidence intervals for the regression coefficients in your model and compare these intervals to the ones you calculate directly from the outputs of the glm function.
2. Apply the proportional odds regression to the original wine quality data and do the following.
   (a) Fit a proportional odds model to the data. Report your findings from the fitted model. Demonstrate how the fitted probabilities for each row in the data can be derived from the fitted log(odds).
   (b) Write a short R program that implements the bagging method for prediction based on the proportional odds regression. The program will take values of the predictors as input and outputs a quality rating of the wine.
   (c) Use a 5-fold cross-validation to compare the predictive accuracies (mean square prediction errors) of the methods used in (a) and (b) above. Does the bagging improve the prediction? Give an explanation to your answer.
3. Use regression tree to analyze the original wine quality data:
   (a) Determine the best value for the complexity parameter and then use it to fit a regression tree to the data. Report your findings from the fitted model.
   (b) Demonstrate how you can use the bootstrap method to obtain a 95% confidence interval for a fitted response value from the regression tree (use one fitted value for demonstration).
   (c) Write a short program in R that implements the random forests method for regression using trees generated from the *rpart* function, using the value of the complexity parameter obtained in (a). Identify and report the important predictors in the data.
   (d) Compare the mean square prediction errors of the tree method and the random forests method using a 5-fold cross-validation.