STA 545 Statistical Data Mining I, Fall 2020

Homework 8, due: Wednesday 11/4/2020 (1PM)

Please submit the pdf file generated by R markdown in UBlearns. Please use tables, figures, or a few sentences to answer data analysis questions.

1. (50 points) In this exercise, we will predict the number of applications received using the other variables in the College data set in the ISLR R package.
   (a) Split the data set into a training set (70%) and a test set (30%).
   (b) Fit a ridge regression model on the training set, with λ chosen by 5-fold cross-validation. Report the final model and the test error obtained.
   (c) Fit a ridge regression model on the training set, with λ chosen by the generalized cross-validation. Please create your own R function for this generalized cross-validation procedure. Report the final model and the test error obtained.
   (d) Fit a PCR model on the training set, with M chosen by 5-fold cross-validation. Please create your own R function for this cross-validation procedure. Report the test error obtained, along with the value of M selected by cross-validation.
2. (50 points) In this exercise, we will compare the performance of 1-NN, LDA, and QDA on the **iris** data. Please use the 10-fold cross-validation method to estimate the test error of each method.