STAT 542, Homework 1

January 26, 2018
Due date: Feb 9, 5:00 pm

Requirements: This homework consists of two problems. You should submit your **report and R code** as separate files to Compass2g. The report should be in PDF/MS Word format, with no more than 6 pages, including plots. Your report should focus on describing the analysis strategy, demonstrating and interpreting the main result. No excessive R code should be included in the report. Font size should be 12pt and plots need to be clearly labeled. If you use R markdown and need a template, there is an example source file provided at our course website.

Question 1 [50 points]: Load the `BostonHousing` data from the `mlbench` R package. The goal is to model the median house value `medv` using other variables. In this question, we will only use linear models. You can use any additional R packages as you like.

a. [10 points] Perform a descriptive analysis on all variables. Comment on any potential issues and address them if needed.

b. [15 points] Perform the best subset selection using BIC criterion. Report the best model (the selected variables and their parameters.

c. [15 points] Perform i) forward stepwise selection using AIC criterion; and ii) backward stepwise selection using Marrow's C_p criterion. Compare these two models with the model in part b).

d. [5 points] Comment on the advantages and disadvantages of the selection algorithms (best subset, forward, backward and stepwise). If you get different results using these three algorithms (assume that you use the same selection criterion), would you prefer some over others? Why?

d. [5 points] Comment on the advantages and disadvantages of the three selection criteria (AIC, BIC, C_p). If you get different results using these three criteria (assuming the same algorithm), would you prefer some over others? Why?

Question 2 [50 points]: Write your own R code that can fit a k nearest neighbor classification model. For this question, you cannot use any R package that directly fits kNN. Then apply your code to the Fashion MNIST dataset on Kaggle (this is a very similar problem as the hand written digit example we used):

https://www.kaggle.com/zalando-research/fashionmnist

This dataset is already separated into a training set and a testing set. Perform the following:

a. [5 points] Provide a short summary of the dataset and the research goal.

b. [20 points] Write your R code for kNN. This is a fairly large dataset, so you should consider writing an efficient algorithm. If it is very slow, you can consider doing part d) first. In addition, how do you deal with ties when k is even.

c. [20 points] Fit your kNN model to the training data, and predict the labels using the testing data. Tune the parameter k to obtain the best testing error and comment on the effect of k. Summarize the performance of the final model. What is the degrees of freedom?

d. [5 points] Can you suggest some approaches that can speed up the computation (even at some minor cost of prediction accuracy)?