Homework 4

Big Data Analytics

Due time: January 3, 2019, 11:59 PM

Instructions:
1. Please turn in your homework in a pdf format generated by R Markdown.
2. Discussion are encouraged, but you need to work on the problems by yourself.

1 (20 points)
Problem 1 (Problem 6, Chapter 7)

In this exercise, you will further analyze the *Wage* data set consider throughout this chapter.

(a) Perform polynomial regression to predict *wage* using *age*. Use cross-validation to select the optimal degree d for the polynomial. What degree was chosen, and how does this compare to the results of hypothesis testing using ANOVA? Make a plot of the results polynomial fit to the data.

(b) Fit a step function to predict *wage* using *age*, and perform cross-validation to choose the optimal number of cuts. Make a plot of the fit obtained.

2 (40 points)
Problem 2 (Problem 9, Chapter 6)

This question uses the variables *dis* (the weighted mean of distances to five Boston employment centers) and *nox* (nitrogen oxides concentration in parts per 10 million) from the *Boston* data. You will treat *dis* as the predictor and *nox* as the response.

(a) Use the *poly()* function to fit a cubic polynomial regression to predict *nox* using *dis*. Report the regression output, and plot the resulting data and polynomial fits.

(b) Plot the polynomial fits for a range of different polynomial degrees (say, from 1 to 10), and report the associated residual sum of squares.

(c) Perform cross-validation or another approach to select the optimal degree for the polynomial, and explain your results.

(d) Use the *bs()* function to fit a regression spline to predict *nox* using *dis*. Report the output for the fit using four degrees of freedom. How did you choose the knots? Plot the resulting fit.

(e) Now fit a regression spline for a range of degrees of freedom, and plot the resulting fits and report the resulting RSS. Describe the results obtained.

(f) Perform cross-validation or another approach in order to select the best degrees of freedom for a regression spline on this data. Describe your results.

3 (40 points)
Problem 3 (Problem 8, Chapter 7)

In the Lab of the textbook, a classification tree was applied to the *Carseats* data set after converting *Sales* into a qualitative response variable. Now you will seek to predict *Sales* using regression trees and related approaches, treating the response as a quantitative variable.

(a) Split the data set into a training set and a test set.

(b) Fit a regression tree to the training set. Plot the tree, and interpret the results. What test MSE do you obtain?

(c) Use cross-validation in order to determine the optimal level of tree complexity. Does pruning the tree improve the test MSE?

(d) Use the bagging approach in order to analyze this data. What test MSE do you obtain? Use the *importance()* function to determine which variables are most important.

(e) Use random forests to analyze this data. What test MSE do you obtain? Use the *importance()* function to determine which variables are most important. Describe the effect of m, the number of variables considered at each split, on the error rate obtained.