STATS 415 Homework 3

Due Thursday Feb 1, 2018

Please write your name, uniqname, and section (number, time, or GSI name) on the front page of your homework. Turn in a printout of your homework in the lecture or in your GSI's mailbox across room 305A West Hall, no later than 5pm on the due date.

This exercise relates to the `Carseats` data set used in Homework 2. Before you proceed, divide the data into training and test sets, using the first 90% of the observations for training, and the remaining 10% for testing.

1. Fit a multiple regression model to predict `Sales` using all other variables (model 1), and a reduced model with everything except for `Population, Education, Urban, and US` (model 2), using only the training data to estimate the coefficients. For both models, report training and test error. Comment on how they differ.

2. Suppose we fit KNN regression to predict `Sales` from the variables used in model 2, except for `ShelveLoc`. Would you expect a better training error with $K = 1$ or $K = 10$? How about test error? Explain your answer (without actually computing the errors).

3. Fitting a KNN regression requires computing distances between data points. Would you standardize the variables in this dataset first? Why or why not?

4. Fit the KNN regression to predict `Sales` from the variables used in model 2, except for `ShelveLoc`. Plot the training and test errors as a function of K. Report the value of K that achieves the lowest training error and the lowest test error. Comment on the shape of the plots and the optimal K in each case.

5. Make a plot of residuals against fitted values for both model 2 and for KNN regression with K of your choice, for the *test data*. Make sure the scale of the axes is the same in both plots. Comment on any similarities or differences.

Please limit your solution to at most 6 pages.