Data Algorithms II

Chelsea Nowlin

Assignment 2

**Exercise 2, 9, 10, 12**

**2.** **Carefully explain the differences between the KNN classifier and KNN regression methods.**

Key differences between KNN regression and KNN classifier are the KNN regression predicts the value of the output using the value of the average while the KNN classifier predicts the class the output variable belongs to using probability.

**9. This question involves the use of multiple linear regression on the Auto data set.**

**(a) Produce a scatterplot matrix which includes all of the variables in the data set.**

**(b) Compute the matrix of correlations between the variables using the function cor(). You will need to exclude the name variable, cor() which is qualitative.**

**(c) Use the lm() function to perform a multiple linear regression with mpg as the response and all other variables except name as the predictors. Use the summary() function to print the results. Comment on the output. For instance: i. Is there a relationship between the predictors and the response? ii. Which predictors appear to have a statistically significant relationship to the response? iii. What does the coefficient for the year variable suggest?**

The small p-value indicates a relationship between mpg and the other variables.

All predictors except the horsepower, cylinders, and acceleration are significant.

With each increase in year, an increase of 0.750773 in mpg is expected, assuming all other predictors remain constant.

**(d) Use the plot() function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?**

The residuals vs fitted model is slightly nonlinear.

The residuals plot shows a few upper and lower outliers with an extreme point at 14 leverage.

**(e) Use the \* and : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?**

Based on the p-values, the interaction between weight and displacement is statistically significant but the interaction between displacement and cylinders in not significant based on the high p-value.

**(f) Try a few different transformations of the variables, such as log(X), √ X, X2. Comment on your findings.**

The log transformation provides the most linear model.

**10. This question should be answered using the Carseats data set.**

**(a) Fit a multiple regression model to predict Sales using Price, Urban, and US.**

**(b) Provide an interpretation of each coefficient in the model. Be careful—some of the variables in the model are qualitative!**

The model provides a coefficient estimate for the variables ShelveLocGood and ShelveLocMedium. Both coefficients are positive so the car seat shelving location has a positive effect on the price of the car seat and performance of sales.

**(c) Write out the model in equation form, being careful to handle the qualitative variables properly.**

Y = 5.539 + 4.849x1 + 1.956 x2 + E

**(d) For which of the predictors can you reject the null hypothesis H0 : βj = 0?**

Both qualitative predictors can be used to reject the null hypothesis based on the p-value of the predictors.

**(e) On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.**

**(f) How well do the models in (a) and (e) fit the data?**

Both models fit the data based on the p-values and the r-square

**(g) Using the model from (e), obtain 95 % confidence intervals for the coefficient(s).**

**(h) Is there evidence of outliers or high leverage observations in the model from (e)?**

It is possible to have outliers, especially if other variables and combinations of predictive coefficients are explored to rule out the outliers that may skew the data.

**12. This problem involves simple linear regression without an intercept.**

**(a) Recall that the coefficient estimate βˆ for the linear regression of Y onto X without an intercept is given by (3.38). Under what circumstance is the coefficient estimate for the regression of X onto Y the same as the coefficient estimate for the regression of Y onto X?**

B0 = 0

When the model takes on a quadratic shape

**(b) Generate an example in R with n = 100 observations in which the coefficient estimate for the regression of X onto Y is different from the coefficient estimate for the regression of Y onto X.**

**(c) Generate an example in R with n = 100 observations in which the coefficient estimate for the regression of X onto Y is the same as the coefficient estimate for the regression of Y onto X.**