Write up for Homework1

Hyungkyu Lim

Question 1

(10 points) Consider the Student Performance Data Set on the UCI machine learning repository (https://archive.ics.uci.edu/ml/datasets/student+performance). Suppose that you are getting this data in order to build a predictive model for First Period Grades. Using the full dataset, investigate the data using exploratory data analysis such as scatterplots, and other tools we have discussed in class. Preprocess this data and justify your choices (elimination of outliers, elimination of variables, variable transformations, etc.) in your write up. Submit the cleaned dataset as an *.RData file.

As far as I concerned, there might be three first period grades; Math, Portuguese and combined grades with math and Portuguese. I know that the question1 says that I have to build a predictive model for first period grade but there are predictors that only explains for each grade such as failures, absences and paid for math and Portuguese grades. I think these variables might be important to explain the relationship with each grade so I tried to do EDA for math, Portugueses and combined grades for them. For data cleaning, one student who has grades for math and Portuguese answered twice for each predictors there are a lot of redundant predictors such as guardian,dalc and famsup so I dropped redundant predictors in merged data. Also I dropped second and three periods for math and Portugueses. Because these grades happened after first period grades, I guess these grade might not affect first period grades.

Question2

(10 points) Perform a multiple regression on the dataset you pre-processed in question one. The response are the first period grades. Use the lm() function in R.

a) Which predictors appear to have a significant relationship to the response.

Predictors such as schooolsup, studytime, famsize, famsup, higher,goout and health have significant relationship with combined grade for math and Portuguese .Predictors such as sex, Fjoboher, schooolsup, studytime, famsup, higher, goout and failures of math have significant relationship with first period grade for math.Predictors such as schoolMS, sex, famsize, studytime, schoolsup, higher, failures of Portuguese, health and absences of Purtuguese have significant relationship with first period grade for Portuguese.

b) What suggestions would you make to a first-year student trying to achieve good grades.

Based on the a), I suggested that most students who have low grade for math failed one or twice on math class previously, so students should focus on math studying to pass at once otherwise, they have to retake the math class and it might affact thier math grade. Futhermore studytime might be important to get both grades so student should spend more time to study math and Portuguese.

c) Use the * and : symbols to fit models with interactions. Are there any interactions that are significant?

Interaction term between schoolsup and studytime have significant relationship with combined grade for math and Portuguese. Also interaction term between each fauilures for math and portuguese and higher have significant realtationship with each grades.

Question3

(10 points) ISL textbook exercise 2.10 modified: This exercise concerns the boston housing data in the MASS library (>library(MASS) >data(Boston)).

a) Make pairwise scatterplots of the predictors, and describe your findings.

It seems that some pairs high correlated such as the relationship between tax and rad, the relationship between medv and lstat.

b) Are any of the predictors associated with per capita crime rate?

In Boston data set, many areas have low crime rate regardless of tax rate, accessbility of the highway or lower status of population. However there are some predictors such as rad (accessibility of the highway), tax (full-value property-tax rate per 10,000 dollars), lstat(lower status of the population), dis (weighted mean of distances to five Boston employment centres.) and medv (median value of owner-occupied homes in 1000 dollars) have is quite highly correlated with per capita crime rate.

Question4

(10 points) ESL textbook exercise 2.8 modified: Compare the classification performance of linear regression and k-nearest neighbor classification on the zipcode data. In particular, consider only the 2's and 3's for this problem, and k = 1,3,5,7,9,11, 13,15. Show both the training and the test error for each choice of k. The zipcode data is available in the ElemStatLearn package.

I'm not sure how linear regression works with classification problem, but first tried to subset 2's and 3's data and then using predictor() to generate predict value to caclulate mse. For KNN, I used KNN function to classify the value and calculate mse. I expected that when k is increasing, mse might decrease. It seems that mse is decreasing when k is getting graeter and greater based on the two graph for mse of train and test.