Summer Poissonnier
March 16, 2021
CAP 4773
Homework 3

Homework 3: Due Thursday, March 25 at 11:59pm

There are two parts to this homework assignment, each with multiple questions. Please insert answers under corresponding questions, then save the document as a pdf and upload it to CANVAS. *Providing your R code is not required, but it may be helpful when assigning partial credit.*

We will again be using the **`College`** dataset in the **`ISLR`** package.

1) First, we will fit a multiple linear regression model to the data to make predictions about a college's out-of-state tuition from all other variables in **`College`**.

   a. (5 points) Use mathematical notation to state the null hypothesis for your multiple linear regression model, ensuring to indicate the number of features in your model.

There are 17 features in my model. The null hypothesis is:

$$H_0: \quad \boldsymbol{\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} = \beta_{15} = \beta_{16} = \beta_{17} = 0}$$

   b. (5 points) Provide the test statistic and associated p value for your hypothesis test from 1a.

The F statistic is: 148.1 and the p value is: 2.2e-16

   c. (5 points) Draw an appropriate conclusion based on the results of your hypothesis test.

Since the p value is less than .05 or .01, we can reject the null hypothesis, H_0. We can support the alternative hypothesis, H_A: at least one B_i is not equal to 0.

   d. (5 points) Briefly explain what this conclusion means about the linear relationship between a college's out-of-state tuition and the features in your model.

This conclusion means that the relationship between a college's out-of-state tuition and at least one or more of the features are directly related.

   e. (5 points) According to your model, which features are linearly related to a college's out-of-state tuition?

The features that are linearly related to the model are: Private, Apps, Accept, Top10Perc, Room.Board, Terminal, perc.alumni, Expend, and Grad.Rate.

f. (5 points) Which of the features listed in 1e is most strongly related to a college's out-of-state tuition? How do you know?

The feature in 1e that is most strongly related to a college's out-of-state tuition is Private. I know this because the test statistic is 9.128 which means that Private is related to out-of-state tuition and the estimate is 2264 which is the largest estimate of the variables.

g. (5 points) What does the feature in 1f represent? *Hint: Think about the type of variable in the* `College` *dataset.*

The feature in 1f represents if the college is public or private. Yes for private and no for public. This is a qualitative variable.

h. (5 points) Taking into consideration your answer to 1g, briefly explain the relationship between the feature in 1f and a college's out-of-state tuition. Be specific about *how much* the average tuition will change with changes to the value of this feature.

The cost of a private out of state college increases the cost of the out of state tuition because since it is private the tuition will cost more than for a public out of state college. The variables are directly related. The average tuition will change drastically with changes to the values of Private.

i. (5 points) Provide the values of the two goodness-of-fit statistics computed for this model.

The Residual standard error is: 1958 and the R^2 is: 0.7684.

j. (5 points) Compare the statistics from 1g to those obtained for the simple linear regression model from Homework 2. Based on your comparisons, which model has a better fit?

The model with the best fit is the multiple linear regression model which has an RSE of 1958 and a R^2 of 0.7684 versus the simple linear model with an RSE of 3329 and R^2 of 0.3162.

2) Next, we will apply K-nearest neighbors to make predictions about whether a college is private using all other variables in `College` as features.

a. (5 points) Designate rows 1-500 as the training data, and rows 501-777 as the test data. Which college is in the top row of your test data?

The college in the top row of the test data is: Saint Mary of the Woods College.

b. (10 points) How many colleges in the test data are classified by KNN as private when $K = 1, 10,$ and 100? ***For reproducibility, set the seed to a value of "1" before running each KNN.***

There are 179 colleges classified as private when K=1, 175 when K=10, and 197 when K =100.

c. (10 points) Examine tables of predicted classes of your KNN classifiers vs. the observed classes from your test data. When your predictions are incorrect, does it tend to be because you classified a public college as private (false positive) or a private college as public (false negative)?

When a prediction is incorrect it tends to be because I classified a public college as private (false positive).

d. (10 points) Provide test accuracy rates for your KNN classifier when $K = 1, 10,$ and $100.$

The test accuracy rate for K=1 is: 0.888, for K=10 is: 0.924, and for K=100 is: 0.859.

e. (10 points) Provide test error rates for your KNN classifier when $K = 1, 10,$ and $100.$

The test error rate for K=1 is: 100-88.8 = 11.2%, K=10 is: 100-92.4 = 7.6%, and K = 100 is: 100-85.9 = 14.1%.

f. (5 points) Which of these values of K likely produces a KNN decision boundary with a shape that is closest to that of the Bayes decision boundary?

The value of K that likely produces a KNN decision boundary with a shape closet to that of Bayes decision boundary is when K = 10.

Code:

```
# Question 1

lm.fit <- lm(Outstate ~ ., data = College)
lm.fit

summary(lm.fit)

#Question 2 a
install.packages("class")
library(class)

train <- seq(1,500)
train.x <- cbind(Apps, Accept, Enroll,Top10perc, Top25perc,
              F.Undergrad, P.Undergrad, Outstate, Room.Board,
              Books, Personal, PhD, Terminal, S.F.Ratio,
              perc.alumni, Expend, Grad.Rate)[train,]
test <- seq(501,777)
test.x <- cbind(Apps, Accept, Enroll,Top10perc, Top25perc,
             F.Undergrad, P.Undergrad, Outstate, Room.Board,
             Books, Personal, PhD, Terminal, S.F.Ratio,
             perc.alumni, Expend, Grad.Rate)[test,]
train.Y <- Private[train]

# 2b
set.seed(1)
knn.pred <- knn(train.x, test.x,train.Y, k=1)
summary(knn.pred)

set.seed(1)
knn.pred1 <- knn(train.x, test.x,train.Y, k=10)
summary(knn.pred1)

set.seed(1)
knn.pred2 <- knn(train.x, test.x,train.Y, k=100)
summary(knn.pred2)

#2c
test.Y <- Private[test]
table(knn.pred, test.Y)
table(knn.pred1, test.Y)
table(knn.pred2, test.Y)

#2d
```

```
mean(knn.pred == test.Y)
mean(knn.pred1 == test.Y)
mean(knn.pred2 == test.Y)
```