Homework 1

Due Date 09/20/2018

By: Marc Mailloux

M3127024

Q1. For each part below indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

(i) The sample size n is extremely large, and the number of predictors p is small.

a. (Better) I would choose a **flexible** learning model as the better, because all the larger the n we most likely wont overfit.

(ii) The number of predictors p is extremely large, and the number of observations n is small.

a. (worse) I would choose a **inflexible** method as the better since the other would over fit the data, and create a bias

(iii) The relationship between the predictors and response is highly non-linear.

a. (Better)I would choose a **flexible** model as better because a flexible model will fit the non-linearity better

(iv) The variance of the error terms, i.e. σ^2 = Var(ϵ), is extremely high.

a. (worse) I would choose the **inflexible** model since the flexible model will catch too much noise.

Q2. We have data from the questionnaires survey (to ask people opinion) and objective testing with two attributes (acid durability and strength) to classify whether a special paper tissue good or not. The following table gives the training sample

X_1= Acid Durability (seconds)	X_2= Strength (kg/square meter)	Y
7	7	Bad
7	4	Bad
3	4	Good
1	4	Good
5	3	Good

7	5	Good

Now the factory produces a new paper tissue that pass laboratory test with X_1= 3 and X_2= 7. Suppose we wish to use this data set to make a prediction for Y when X_1= 3 and X_2= 7 using K-nearest neighbors.

(i) Compute the Euclidean distance between each observation and the test point X_1= 3 and X_2= 7.

a. MATH HERE: Using the distance formula for each training set

i. Dn = sqrt((x1-y1)^2+(x2-y2)^2)

1. Where x is the test test and y is training

ii. D1 = 4 , Prediction Bad

iii. D2 = 5 , Prediction Bad

iv. D3 = 3, Prediction Good

v. D4 = 3.6, Prediction Good

vi. D5 = 4.123, Prediction Good

(ii) What is our prediction with K = 1? Why?

a. GOOD Looking at the calculations when k=1 observation 3 is the lowest distance when sorted

(iii)What is our prediction with K = 3? Why?

a. GOOD Looking at the distance calculations when k=3, there are two distances corresponding to good and one to bad so we choose good.

(iv) If the Bayes decision boundary in this problem is highly non-linear, then would we expect the best value for K to be large or small? Why?

a. Since KNN is flexible as it, if we were to increase K to be large then this would lead the algorithm to be inflexible so a small K should be used instead.

Q3. Question 10 from Chapter 2 exercise

Q4. Suppose you have regression data generated by a polynomial of degree 3. Characterize the bias-variance of the estimates of the following models on the data with respect to the true model by circling the appropriate entry.

	Bias	Variance
Linear Regression (HIGH/LOW) Polynomial Regression with degree 3(LOW/ LOW) Polynomial Regression with degree 10(LOW/ HIGH)	low/**high** **low**/high **low**/high	**low**/high **low**/high low/**high**

Q5. Carefully explain the differences between the KNN classifier and KNN regression methods.

The main difference is primarily at the difference types of data they are used to analyze. Specifically the KNN Classifier is used for more qualitative methods like groups or categories, where the KNN regression is used for more quantitative method or numerical methods. More closely, someone would use the KNN classifier to classify groups, and the KNN regression method to make a prediction from quantitative methods.

Q6. Question 3 from Chapter 3 exercise
Q7. Question 6 from Chapter 3 exercise
Q8. Question 7 from Chapter 3 exercise
Q9. Question 11 from Chapter 3 exercise
Q10. Question 5 from Chapter 4 exercise
Q11. Question 2 from Chapter 4 exercise
Q12. Question 11 from Chapter 4 exercise

RCODE :

Q3

```
library(MASS)
library(ggplot2)
library(plyr)

?Boston
#Q1: How many rows are in this data set? How Many Columns? What do the Rows and Columns represent?
#506 rows and 14 Columns
#The Columns represent the predictors and the rows are the values for each observation of the suburb
dim(Boston)
str(Boston)
print(Boston)
#partb
par(mfrow = c(1, 2))
plot(Boston$rm, Boston$crim)
plot(Boston$age, Boston$crim)
pairs(Boston)
#From looking at the scatter plots it seems that there are some correlations between various variables
#Specfically Age has correlations between other variables like nox,Istat, and crim
#it also shows which variables are categorical and which are quantitative, where the qualitative variables are either 1 to 5
#part c
hist(Boston$crim, breaks = 50)
#the data is for the most part under 18
#Using less than 20 to have a better look at the data
pairs(Boston[Boston$crim < 18, ])
#Yes the variables crim seems to be associated with AGE,NOX,rm,dis,medv, and Istat. Other variable correlations are difficult to tell
#looking only at the age and crim it clear that there is some sort of relationship

#partd
hist(Boston$crim, breaks = 50)
nrow(Boston[Boston$crim > 20, ])
#only 18 areas have higher crime rate of 20% or higher than the rest of Boston
nrow(Boston[Boston$crim > 60, ])
#only 3 urban areas have crime rates of 60% and higher
hist(Boston$tax, breaks = 50)
nrow(Boston[Boston$tax > 600, ])
#137 urban areas have a tax rate of 600 or higher
#from the data set we see alot are at 666 tax rate so lets find out how many areas that is exaclty
nrow(Boston[Boston$tax == 666, ])
#There are 132 urban areas with the tax rate of 666.
hist(Boston$ptratio, breaks = 50)
#from this historgram it can be seen that there are alot of areas near 21 so lets determine how many there
nrow(Boston[Boston$ptratio > 20 ,])
#There are 201 homes with a high ptratio
#Yes some of the suburbs appear to have high crime rates, Tax Rates appear to have some urban areas with high tax rates, although majority are below 50%
#The pupil teacher ratio seems to be uniform over the population set

#parte
y = count(Boston$chas)
print (y)
nrow(Boston[Boston$chas == 1 , ])
#there are 35 cities touching the Charles River

#partf

pupil = Boston$ptratio
s = summary(pupil)
print (s)
# the Median Student to Pupil ratio is 19.05

#partg
newBoston <- Boston[order(Boston$medv),]
print (newBoston)

row.names(Boston[min(Boston$medv), ])
```

```
#so suburb 5 has the smalled medv
range(Boston$ptratio)
#the range of ptratio for Boston is 12.6 to 22.0
Boston[min(Boston$medv), ]$ptratio
#finding out what suburb 5's ptratio: 18.7 so there are more students in this area than children

#part h

nrow(Boston[Boston$rm > 7.0,])
#There are 64 urban areas with homes with more than 7 rooms on average
nrow(Boston[Boston$rm > 8.0,])
#There are 13 urban areas with homes more than 8 rooms
#so 64-13 = 51 and so there are 51 areas that have 7 rooms on average
```

Chapter 3 Questions

```
#chapter 3 Question 11

set.seed(1)
x = rnorm(100)
y = 2*x+rnorm(100)
p = lm(y~x+0)
summary(p)
#The value of B is 1.9939
#The SE is 0.1065
#the t-score is 18.73
#p-value is 2.2e-16
#since our p is small we reject the Ho

#partb

q = lm(x~y+0)
summary(q)
#The value of B is .39111
#The SE is 0.02098
#the t-score is 18.73
#p-value is 2.2e-16
#since our p is small we reject the Ho

#partC

#the values of B for both results do differ as for the error.
#the t-scores and p-values are equal to each other
#This shows that we are analyzing the same line that was initially created.

#part d

w = length(x)

z = sqrt(w-1)*(x %*% y)
e = sqrt(sum(x**2)*sum((y**2))-(x%*%y)**2)
t.stat = z/e
as.numeric(t.stat)
#18.72593

#parte argue why the t stats woudl be the same if we replace x and y with each other.

#this can be seen from the formulas above. If i were to replace them mathematically nothing will change
#so of course they will be the same.

#part f

p = lm(y~x)
summary(p)

q= lm(x~y)
summary(q)
#the t stats for x and y are 18.56 with a p value of 2.2e-16
#again showing that the regression for y onto x is the same as x onto y
```

Chapter 4 Questions

```
library(ISLR)
attach(Auto)
library(class)
mpg01 <- rep(0, length(mpg))
mpg01[mpg > median(mpg)] = 1
Auto <- data.frame(Auto, mpg01)
Auto

pairs(Auto)
plot(Auto$mpg01,Auto$horsepower)
plot(Auto$mpg01,Auto$weight)
plot(Auto$mpg01,Auto$mpg)
plot(Auto$mpg01,Auto$acceleration)
plot(Auto$mpg01,Auto$cylinders)

#looking at these scatter plots above acceleration and cylinders arent ideal predictors
#This is to due to there being some overlap and an ideal predictor would like the comparison to mpg
#the predictors I would choose are horsepower, weight, and mpg

#part c

train = (year %% 2 == 0 )
data.train = Auto[train,]
data.test = Auto[!train,]
mpg01.test <- mpg01[!train]

#part d
model = mpg01 ~ horsepower + weight + mpg
d.lda = lda(model)
d.lda

p.lda = predict(d.lda,data.test)
table(p.lda$class,mpg01.test)
#produced the confusion matrix

mean(p.lda$class != mpg01.test)
#so the test Error Rate is 3.29%

#part e


e.qda = qda(model)
p.qda = predict(e.qda,data.test)
table(p.qda$class,mpg01.test)

mean(e.qda$class !=mpg01.test)
#test error rate is 4.95%

#part f

f.glm = glm(model)
re= predict(f.glm,data.test, type = 'response')
p.glm = rep(0, length(re))
p.glm[re > 0.5] = 1
table(p.glm, mpg01.test)
mean(p.glm !=mpg01.test)
#The error rate is 3.2%

#part g

train.auto <- cbind(mpg, weight, horsepower)[train, ]
test.auto <- cbind(mpg, weight, horsepower)[!train, ]
train.mpg01 <- mpg01[train]
set.seed(1)
p.knn <- knn(train.auto, test.auto, train.mpg01, k = 20)
table(p.knn, mpg01.test)
mean(p.knn != mpg01.test)

#The error for:
```

```
#k=1 is 14.83516
#k=3 is 15.38
#k=10 is 15.93
#k=50 is 14.29
#k=100 is 14.835
#k=200 is 54.94%
# so it seems right between say 25 and 75 is a spot where the optimal k lies
```