MATH 4322 Homework 1

Saim Ali

1/28/23

Instructions

1. Due date: January 31, 2023, 11:59 PM
2. Answer the questions fully for full credit.
3. Scan or Type your answers and submit only one file. (If you submit several files only the recent one uploaded will be graded).
4. Preferably save your file as PDF before uploading.
5. Submit in Canvas under Homework 1.
6. These questions are from *An Introduction to Statistical Learning*, second edition by James, et. al., chapter 2.
7. The information in the gray boxes are R code that you can use to answer the questions.

Problem 1

Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

a) We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.
Regression, Prediction; n=52 (52 weeks), p=3 (3 Predictors)

b) An online store is determining whether or not a customer will purchase additional items. This online store collected data from 1500 customers and looked at cost of initial purchase, if there was a special offer, type of item purchased, number of times the customer logged into their account, and if they purchased additional items.
Classification, Prediction; n=1500 (1500 customers), p=5 (5 Predictors)

Problem 2

This is an exercises about bias, variance and MSE.

Suppose we have n independent Bernoulli trails with true success probability p. Consider two estimators of p: $\hat{p}_1 = \hat{p}$ where $\hat{p}$ is the sample proportion of successes and $\hat{p}_2 = \frac{1}{2}$, a fixed constant.

a) Find the expected value and bias of each estimator.

E($\hat{p}_1$) = 1/n(x_1 +…+x_n) = 1/n(p+…+p)= p

E($\hat{p}_2$) = E(1/2) = 1/2s

Bias($\hat{p}_1$) = E($\hat{p}_1$)-p = p-p = 0

Bias($\hat{p}_1$) = E($\hat{p}_1$)-p = p-p = 0

Bias($\hat{p}_2$) = E($\hat{p}_2$)-p = 1/2-p; so it will be biased unless p = 1/2.

b) Find the variance of each estimator.

Var($\hat{p}_1$) = Var(1/n(x_1+…+x_n)) -> (1/(n^2))(pq+…+pq) == (1/n)(pq)

Var($\hat{p}_2$) = Var(1/n(x_1+…+x_n)) -> Var(1/2) == 0

Therefore, we have determined that $\hat{p}_2$ has a smaller variance

c) Find the MSE of each estimator and compare them by plotting against the true p. Use $n = 4$. Comment on the comparison.

MSE($\hat{p}_1$) = Var($\hat{p}_1$)+(Bias($\hat{p}_1$))^2 = (1/n)(pq)+0 = (1/n)(pq) == (1/n)((p)(1-p))

***MSE($\hat{p}_2$) = Var($\hat{p}_2$)+((Bias($\hat{p}_2$))^2 = 0+((1/2)-p)^2 == ((1/2)-p)^2**

This means that $\hat{p}_1$ will typically have a MSE for most values of p but not if p is near 1/2; if p is near 1/2 then $\hat{p}_2$ has a smaller MSE

Problem 3

Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a non-parametric approach)? What are its disadvantages?

In a parametric statistical learning approach models that are used on the data are characterized by a certain number of parameters and distribution of the data is assumed to belong to a specific known family of distributions.

Advantages include: simpler computations, easier interpretations, and gives betters performance when data is limited.

Disadvantages include: Models are less flexible as they are limited by assumptions we make, performance can be poor if the distribution of data doesn't match the assumed family of distributions

On the other hand non-parametric the model used to fit the data is characterized by a flexible.The parameters are estimated using methods such as kernel regression, decision trees, and nearest neighbor methods.

Advantages include: More flexibility, better performance with complex data distributions

Disadvantages include: Harder/more complex computations, Harder to interpret, can lead to over-fitting when data is limited

Problem 4

This exercise involves the `Auto` data set in `ISLR` package. Make sure that the missing values have been removed from the data.

(a) Which of the predictors are quantitative, and which are qualitative? **Quantitative: mpg, displacement, horsepower, weight, acceleration Qualitative: cylinders, year, origin, name**

(b) What is the range of each quantitative predictor? You can answer this using the `summary()` function.
mpg: (9,46.60), displacement: (68,455), horsepower: (46,230), weight: (1613,5140) acceleration: (8,24.80)

(c) What is the mean and standard deviation of each quantitative predictor?
mpg: 23.45, 7.805007; displacement: 194.4, 104.644; horsepower: 104.5, 38.49116; weight: 2978, 849.4026; acceleration: 15.54, 2.758864

(d) Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

mpg: (11,46.6), 24.4, 7.867283; cylinders: (3,8), 5.373, 1.654179; displacement: (68,455), 187.2, 99.67837; horsepower: (46,230), 100.7, 35.70885; weight: (1649,4997), 2936, 811.3002; acceleration: (8.5,24.8), 15.73, 2.693721; year: (70, 82), 77.15, 3.106217; origin: (1,3), 1.6001, 0.81991

(e) Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.
I found that many of these different attributes in this data set have strong linear relationships with each other both positive and negative. I saw strong evidence of these attributes being strongly influential in the change in values for each other.

(f) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer.
Yes when I evaluated the scatterplot of mpg~weight I found a linearly negative relationship occuring. This suggests that the heavier your car is -> the lower your miles per gallon will be which makes sense in a real-life scenario.

Problem 5

This exercise relates to the `College` data set, which can be found in the file `College.csv` attached to this homework set in Blackboard. It contains a number of variables for 777 different universities and colleges in the US. The variables are

- Private : Public/private indicator
- Apps : Number of applications received
- Accept : Number of applicants accepted
- Enroll : Number of new students enrolled
- Top10perc : New students from top 10% of high school class
- Top25perc : New students from top 25% of high school class
- F.Undergrad : Number of full-time undergraduates
- P.Undergrad : Number of part-time undergraduates
- Outstate : Out-of-state tuition
- Room.Board : Room and board costs
- Books : Estimated book costs
- Personal : Estimated personal spending
- PhD : Percent of faculty with Ph.D.'s
- Terminal : Percent of faculty with terminal degree
- S.F.Ratio : Student/faculty ratio
- perc.alumni : Percent of alumni who donate
- Expend : Instructional expenditure per student
- Grad.Rate : Graduation rate

Before reading the data into `R`, it can be viewed in Excel or a text editor.

a) Use the `read.csv()` function to read the data into `R`. Call the loaded data `college`. Make sure that you have the directory set to the correct location for the data. You can also import this data set into `RStudio` by using the **Import Dataset → From Text** drop down list in the Environment window.

b) Look at the data using the `View()` function. You should notice that the first column is just the name of each university. We will not use this column as a variable but it may be handy to have these names for later. Try the following commands in `R`:

```
View(college)
rownames(college) <- college[,1]
college <- college[,-1]
View(college)
summary(college)
```

If you are getting an error make sure your data frame is named with a lowercase "c".
Give a brief description of what you see in the data frame.
In the college dataset I see 777 entries, 18 total columns which have various different variables some of which are words instead of integers.

c) Use the `summary()` function to produce a numerical summary of the variables in the data set. Is there any variables that do not show a numerical summary?
The 'Private' variable does not show a numerical summary, I believe this is due to that variable only containing a 'Yes' or 'No' response.

Type in the following in `R`:

```
college$Private <- as.factor(college$Private)
View(college)
attach(college)
#pairs(college)
p<- pairs(college[,1:5])

#plot(Accept~Apps)
#plot(Outstate~Private)
```

d) Use the `pairs()` function to produce a scatterplot matrix of the first five columns or variable of the dataset. Describe any relationships you see in these plots.
I see a few positively linear relationships such a Apps~Accept which means that the higher the application count for the college is -> the higher the acceptance rate is. I also observed that some of the plots given did now indicate any possible relationship between the two given variables being plotted, this was shown in the form of a plot which shows no sign of a pattern present.

e) Use the `plot()` function to produce a plot of `Outstate` versus `Private`. What type of plot was produced? Give a description of the relationship. *Hint: `Outstate is in the y-axis.*
"plot(Outstate~Private)" created a boxplot with two variables 'Yes' and 'No'. The Y-axis consists of the 'Outstate' variable while 'Private' is represented by the 'Yes' and 'No' shown in the x-axis.The relationship shows that the spread of 'Yes' (middle 95%) is higher when 'Outstate' is higher. There are a few outliers also shown in the boxplots by dots outside of the whiskers. Therefore we can assume that the higher the 'Outstate' value the higher chance it will be a private college.

f) Create a new qualitative variable, called `Elite`, by *binning* the `Top10perc` variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50%. Type in the following in `R`:

```
Elite <- rep("No", nrow(college)) #this gives a column of No's for the same
number of rows college.
Elite[college$Top10perc > 50] <- "Yes" #changes to Yes if top 10% is greater
than 50
Elite <- as.factor(Elite)
college <- data.frame(college,Elite) #adds Elite as a column
summary(college)
```

Use the `summary()` function to see how many elite universities there are.
Using the requirements above we derive that there is 78 elite colleges and 699 colleges that are not elite in our given dataset.

Problem 6

This exercise involves the Boston housing data set.

(a) To begin, load in the Boston data set. The Boston data set is part of the ISLR2 library. You may have to install the ISLR2 library then call for this library.

```
library(ISLR2)
```

Now the data set is contained in the object Boston.

Read about the data set:

```
##?Boston
attach(Boston)
plot(Boston)

plot(crim~dis)
```

How many rows are in this data set? How many columns? What do the rows and columns represent?
Boston is "A data frame with 506 rows and 13 variables."

(b) Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.
Initially I found that there was alot of predictors present in this dataset so I was unable to read the pairwise scatterplots, to fix this issue I ran the plot function for only a few variables at a time to get a better understanding of the given plots. After this, I found many of these variables don't actually have relationships and only a few plots can be seen with any sort of relationship being seen.

(c) Are any of the predictors associated with per capita crime rate? If so, explain the relationship.
I found it hard to find any sort of relationship for the 'crim' predictor, the closest thing to a relationship I found for this variable was: 'crim' plotted with 'dis' which showed me that a very low dis value results in a higher crim value. It did seem like the slope flattens out after '4=dis' so it seems exponentially negative.

(d) Do any of the census tracts of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.
For "per capita crime rate by town" there are many outliers in the upper extreme of the boxplot. Most of the towns are seen to have low crim values are between 0-5. Outliers range from 10-above 80. For tax rates the boxplot we create shows no outliers and the median value is slightly above 300 which means that the data for tax is skewed as the data ranges from 200 to 700 as shown from the boxplot whiskers.
For Pupil-teacher ratios the boxplot shows us that the variable has a few

outliers in the lower section of the boxplot. The boxplot data is shown to range from 12 to 22 and median is around 19.

```
par(mfrow=c(1,3))
boxplot(Boston$crim, xlab = "crim")
boxplot(Boston$tax, xlab = "tax")
boxplot(Boston$ptratio, xlab = "ptratio")
table(Boston$chas)
library(tidyverse)

median(Boston$ptratio)
test <- Boston[order(Boston$medv),]
test[1,]
summary(Boston)

seven <- subset(Boston, rm>7)
nrow(seven)
eight <- subset(Boston, rm>8)
nrow(eight)
```

(e) How many of the census tracts in this data set bound the Charles river?
There is 35 tracts bound the Charles river.

(f) What is the median pupil-teacher ratio among the towns in this data set?
The median pupil-teacher among the towns in this data set is: 19.05

(g) Which census tract of Boston has lowest median value of owner occupied homes? What are the values of the other predictors for that census tract, and how do those values compare to the overall ranges for those predictors? Comment on your findings.
Census 399 had the lowest median value of owner occupied homes with med=5 ($5,000)
After viewing the summary of the other predictors lowest median values I found: crim to be higher, min to be at the median, indus to be in the 3rd quartile range, chas to be the median, nox to be higher than the 3rd quartile, rm to be less than the 1st quartile, age to be the max, dis ot be less than the 1st quartile, rad to be the max, tax to be the 3rd quartile, pratio to be at the 3rd quartile, and lstat to be above the 3rd quartile.
I found that these values compare to the overall ranges of the predictors by being within the 1st to 3rd quartile typically and sometimes being the values of the min/max.

(h) In this data set, how many of the census tracts average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the census tracts that average more than eight rooms per dwelling.
64 census tracts average more than seven rooms per dwelling.
8 census tracts average more than seven rooms per dwelling.

I looked into the census tracts and found that there is a lower lstat and lower crim value present in these tracts.