MATH 4322 Homework 1

Dr. Cathy Poliak

Spring 2023

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.

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.

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.

b) Find the variance of each estimator.

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

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?

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?

(b) What is the range of each quantitative predictor? You can answer this using the `summary()` function.

(c) What is the mean and standard deviation of each quantitative predictor?

(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?

(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.

(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.

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`:

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

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?

Type in the following in `R`:

```
college$Private <- as.factor(college$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.

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.*

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
```

Use the `summary()` function to see how many elite universities there are.

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.

```
Boston
```

Read about the data set:

```
?Boston
```

How many rows are in this data set? How many columns? What do the rows and columns represent?

(b) Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.

(c) Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

(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.

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

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

(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.

(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.