**Md Reza**

**HW2:** ISLR 2.4 Exercises 1 - 8

**Due Date:** 09-22-2019

**Q1.** For each of parts (a) through (d), indicate whether would generally expect the performance of a flexible statistical learning method to perform better or worse than an inflexible method when:

1. The sample size *n* is extremely large, and the number of predictors *p* is small?

A flexible method would fit the data closer and extract information from the large *n*.

1. The number of predictors’ *p* is extremely large, and the number of observations *n* is small?

A flexible method would cause noises, and it would also overfit if the number of observation *n* is small.

1. The relationship between the predictors and response is highly non-linear?

A flexible method with more degrees of freedom would fit better with the highly non-linear model.

1. The variance of the error terms, i.e. *σ2 = Var(ε)*, is extremely high ?

The high variance of the error term would cause lots of noise. Therefore, a flexible method would increase the variance to fit the noise.

**Q2.** 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*.

1. We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

This is a regression problem with inference (*response understanding which factors would affect CEO salary*); given *n* = 500 (i.e. data point) and *p* = 3 (i.e. profit, number of employee, and industry).

1. We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

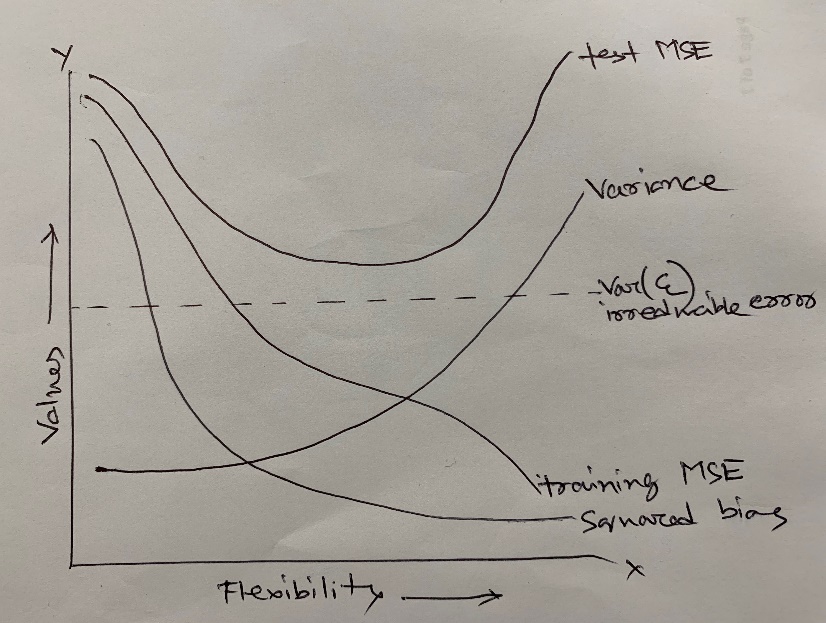
This is a classification problem with prediction (*response whether launching a new product will a success or a failure*); given *n* = 20 (i.e. similar products previously launched) and *p* = 13 (i.e. product, marketing budget, competition price, and ten other variables).

1. We are interesting 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.

This is a regression problem with prediction (*response % change is USD/Euro in relation to weekly changes in the world market*); given *n* = 52 (i.e. number of weeks) and *p* = 3 (i.e. US, British & German Markets).

**Q3.** We now revisit the bias-variance decomposition.

1. Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The *X*-axis should represent the amount of flexibility in the method, and the *Y*-axis should represent the values for each curve. There should be five curves. Make sure to label each one.

****

1. Explain why each of the five curves has the shape displayed in part (a).

Training MSE - As the level of flexibility increased the training MSE is decreased, this is because the bias is large when the model is inflexible, but as the complexity increased it try to fit the data hence the bias is reduced resulting reduction in the training error.

Test MSE - Per bias and variance trade-off, the variance is smaller and bias is larger when the model is inflexible, but inversely as the model becomes more complex the bias reduces faster than the variance resulting large test error, and that why test MSE has U shape carve.

The squared bias decreases monotonically as the flexibility increased this is because building complicated problem on top of a simple linear regression model will have a high bias.

The variance increased monotonically as the flexibility increased this is because the general rule of thumb the model with high flexibility would produce high variance.

The irreducible error is a constant parallel line where the increased flexibility wouldn’t impact this line. It also provides the upper bound to validate the accuracy of the model.

**Q4.** You will now think of some real-life applications for statistical learning.

1. Describe three real-life applications in which classification might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

Classification:

1. Identify whether a given email can be assigned to spam or non-spam
   1. Response: Spam/Non-Spam
   2. Predictors: Display name, Subject line, Spelling mistakes, Email body, Signature
   3. Goals: Prediction
2. Determined whether credit card applicant’s has good credit or bad credit
   1. Response: Good credit/Bad credit
   2. Predictors: Annual salary, Outstanding debts, Age
   3. Goals: Prediction
3. Should an applicant be hired by or not
   1. Response: Hire/Don’t hire
   2. Predictors: Cognitive ability, Experience, Creativity, Past performance
   3. Goals: Prediction
4. Describe three real-life applications in which regression might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

Regression:

1. Forecast products sale
   1. Response: What is the sale of the products predicted to be by September 2020?
   2. Predictors: Demographic information, Past buying behavior
   3. Goals: Inference
2. Economic growth in the state of Michigan
   1. Response: What is the economic growth of the state of Michigan to be by the year 2030?
   2. Predictors: Natural resources, Physical capital, Human Capital, Labor, Technology
   3. Goals: Inference
3. House price prediction
   1. Response: Michigan housing market forecast for the year ending 2020
   2. Predictors: Per capita income, Sale price, Unemployment rate, Economic growth
   3. Goals: Inference

**(c)** Describe three real-life applications in which cluster analysis might be useful.

Cluster:

1. Clustering Michigan population into different age group for health care purposes
   1. Response: By 2030 the percentage of the population of Michigan will fall under Baby Boomer, Generation X, Millennials and so forth
   2. Predictors: Demographic, Age
   3. Goals: Prediction
2. Clustering five-star hotel service ratings to best/good/average/bad
   1. Response: This hotel has best/good/average/bad service rating
   2. Predictors: Reliability, Assurance, Tangibles, Responsiveness
   3. Goals: Prediction
3. Marketing and sales prediction
   1. Response: Clustering different targeted audiences to implement audience specific marketing strategy
   2. Predictors: Group people with similar traits and likelihood of purchase
   3. Goals: Prediction

**Q5.** What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

* The advantages and disadvantages of a very flexible approaches are as follows:

Advantages:

* it may give a better fit when the relationship is non-linear
* it decreases the bias
* it has low irreducible error

Disadvantages:

* it could overfit the training data resulting in large test error
* it requires a large number of sample data to avoid the overfitting issue
* it follows the noise closely, might also increase the variance
* A more flexible approach is preferred when we are interested more in prediction than inference.
* A less flexible approach is preferred when we are interested in the inference that is the interpretability of the results.

**Q6.** 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 nonparametric approach)? What are its disadvantages?

With the parametric approach, we try to fit the data to the assumed functional form (i.e. linear/non-linear, flexible/less flexible).

While with the nonparametric approach instead of assuming the functional form we try to estimate the function that would closely fit the data.

The advantages of a parametric approach to regression or classification are:

* the parametric approach could fit a wide variety of data into a functional form and also work well with a simple model
* compare to the nonparametric approach it doesn’t require a large number of parameters

The disadvantages of a parametric approach to regression or classification are:

* large numbers of parameters would overfit the training data resulting test error
* to train the model it required large training data that would result in large bias

**Q7.** The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Obs. | X1 | X2 | X3 | Y |
| 1 | 0 | 3 | 0 | Red |
| 2 | 2 | 0 | 0 | Red |
| 3 | 0 | 1 | 3 | Red |
| 4 | 0 | 1 | 2 | Green |
| 5 | 1 | 0 | 1 | Green |
| 6 | 1 | 1 | 1 | Red |

Suppose we wish to use this data set to make a prediction for *Y* when *X1 = X2 = X3* = 0 using *K*-nearest neighbors.

1. Compute the Euclidean distance between each observation and the test point, *X1 = X2 = X3* = 0

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Obs. | X1 | X2 | X3 | Y | Calculation | Distance |
| 1 | 0 | 3 | 0 | Red | sqrt[(0-0)^2+(3-0)^2+(0-0)^2] | 3 |
| 2 | 2 | 0 | 0 | Red | sqrt[(2-0)^2+(0-0)^2+(0-0)^2] | 2 |
| 3 | 0 | 1 | 3 | Red | sqrt[(0-0)^2+(1-0)^2+(3-0)^2] | 3.16 |
| 4 | 0 | 1 | 2 | Green | sqrt[(0-0)^2+(1-0)^2+(2-0)^2] | 2.24 |
| 5 | 1 | 0 | 1 | Green | sqrt[(-1-0)^2+(0-0)^2+(1-0)^2] | 1.41 |
| 6 | 1 | 1 | 1 | Red | sqrt[(1-0)^2+(1-0)^2+(1-0)^2] | 1.73 |

1. What is our prediction with *K* = 1? Why?

Our prediction is Green because the closest single neighbor is observation 5, which is Green.

1. What is our prediction with *K* = 3? Why?

Our prediction is Red because the closest neighbors are observations 2(Red), 5(Green), 6(Red), where two out three are Red.

1. If the Bayes decision boundary in this problem is highly nonlinear, then would we expect the best value for *K* to be large or small? Why?

Small, because the boundary becomes inflexible and linear as the *K* becomes larger.

**Q8.** This exercise relates to the College data set

**(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.

> library(ISLR)

> data("College")

> college <- read.csv("College.csv")

**(b)** Look at the data using the fix() function. You should notice that the first column is just the name of each university. We don’t really want R to treat this as data. However, it may be handy to have these names for later.

> head(college[, 1:5])

X Private Apps Accept Enroll

1 Abilene Christian University Yes 1660 1232 721

2 Adelphi University Yes 2186 1924 512

3 Adrian College Yes 1428 1097 336

4 Agnes Scott College Yes 417 349 137

5 Alaska Pacific University Yes 193 146 55

6 Albertson College Yes 587 479 158

> rownames <- college[, 1]

> college <- college[, -1]

> head(college[, 1:5])

Private Apps Accept Enroll Top10perc

1 Yes 1660 1232 721 23

2 Yes 2186 1924 512 16

3 Yes 1428 1097 336 22

4 Yes 417 349 137 60

5 Yes 193 146 55 16

6 Yes 587 479 158 38

**(c)**

**i.** Use the summary() function to produce a numerical summary of the variables in the data set.

> summary(college)

Private Apps Accept Enroll Top10perc Top25perc

No :212 Min. : 81 Min. : 72 Min. : 35 Min. : 1.00 Min. : 9.0

Yes:565 1st Qu.: 776 1st Qu.: 604 1st Qu.: 242 1st Qu.:15.00 1st Qu.: 41.0

Median : 1558 Median : 1110 Median : 434 Median :23.00 Median : 54.0

Mean : 3002 Mean : 2019 Mean : 780 Mean :27.56 Mean : 55.8

3rd Qu.: 3624 3rd Qu.: 2424 3rd Qu.: 902 3rd Qu.:35.00 3rd Qu.: 69.0

Max. :48094 Max. :26330 Max. :6392 Max. :96.00 Max. :100.0

F.Undergrad P.Undergrad Outstate Room.Board Books Personal

Min. : 139 Min. : 1.0 Min. : 2340 Min. :1780 Min. : 96.0 Min. : 250

1st Qu.: 992 1st Qu.: 95.0 1st Qu.: 7320 1st Qu.:3597 1st Qu.: 470.0 1st Qu.: 850

Median : 1707 Median : 353.0 Median : 9990 Median :4200 Median : 500.0 Median :1200

Mean : 3700 Mean : 855.3 Mean :10441 Mean :4358 Mean : 549.4 Mean :1341

3rd Qu.: 4005 3rd Qu.: 967.0 3rd Qu.:12925 3rd Qu.:5050 3rd Qu.: 600.0 3rd Qu.:1700

Max. :31643 Max. :21836.0 Max. :21700 Max. :8124 Max. :2340.0 Max. :6800

PhD Terminal S.F.Ratio perc.alumni Expend Grad.Rate

Min. : 8.00 Min. : 24.0 Min. : 2.50 Min. : 0.00 Min. : 3186 Min. : 10.00

1st Qu.: 62.00 1st Qu.: 71.0 1st Qu.:11.50 1st Qu.:13.00 1st Qu.: 6751 1st Qu.: 53.00

Median : 75.00 Median : 82.0 Median :13.60 Median :21.00 Median : 8377 Median : 65.00

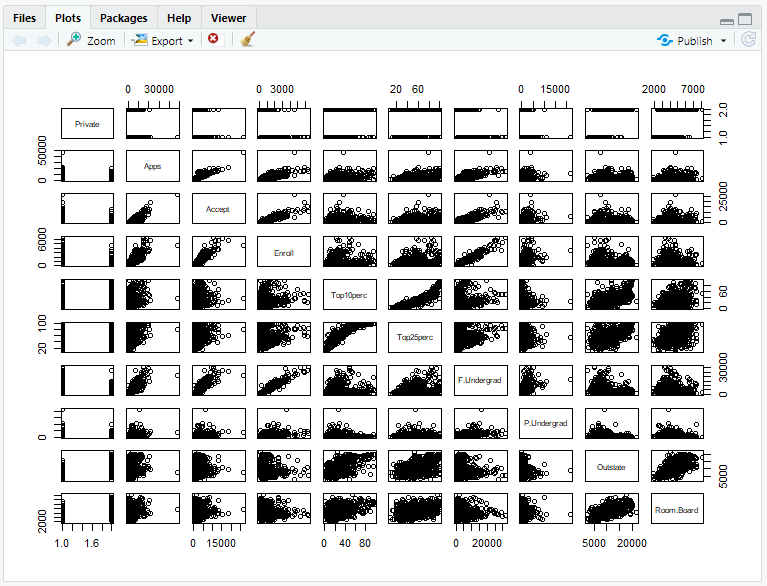
Mean : 72.66 Mean : 79.7 Mean :14.09 Mean :22.74 Mean : 9660 Mean : 65.46

3rd Qu.: 85.00 3rd Qu.: 92.0 3rd Qu.:16.50 3rd Qu.:31.00 3rd Qu.:10830 3rd Qu.: 78.00

Max. :103.00 Max. :100.0 Max. :39.80 Max. :64.00 Max. :56233 Max. :118.00

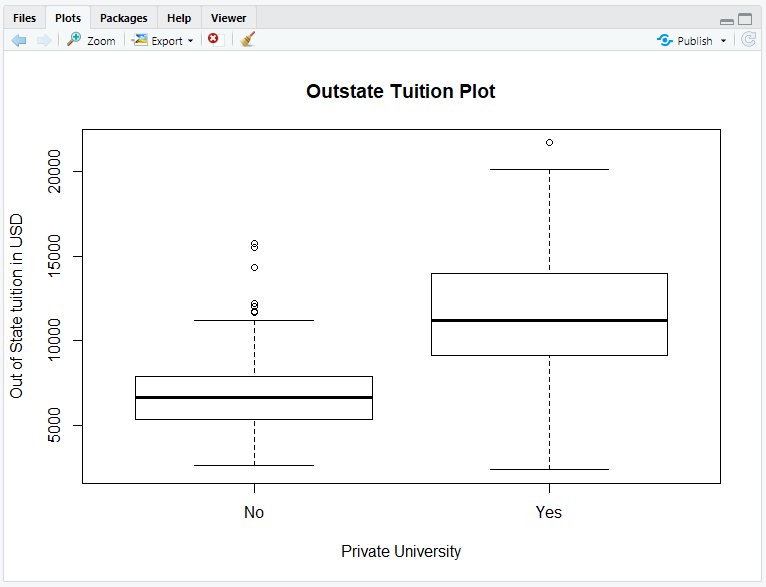
**ii.** Use the pairs() function to produce a scatterplot matrix of the first ten columns or variables of the data.

> pairs(college[, 1:10])



**iii.** Use the plot() function to produce side-by-side boxplots of “Outstate” versus “Private”.

> plot(college$Private, college$Outstate, xlab = "Private University", ylab ="Out of State tuition in USD", main = "Outstate Tuition Plot")



**iv.** Create a new qualitative variable, called “Elite”, by binning the “Top10perc” variable. Use the summary () function to see how many elite universities there are. Now use the plot() function to produce side-by-side boxplots of “Outstate” versus “Elite”.

> Elite <- rep("No", nrow(college))

> Elite[college$Top10perc > 50] <- "Yes"

> Elite <- as.factor(Elite)

> college$Elite <- Elite

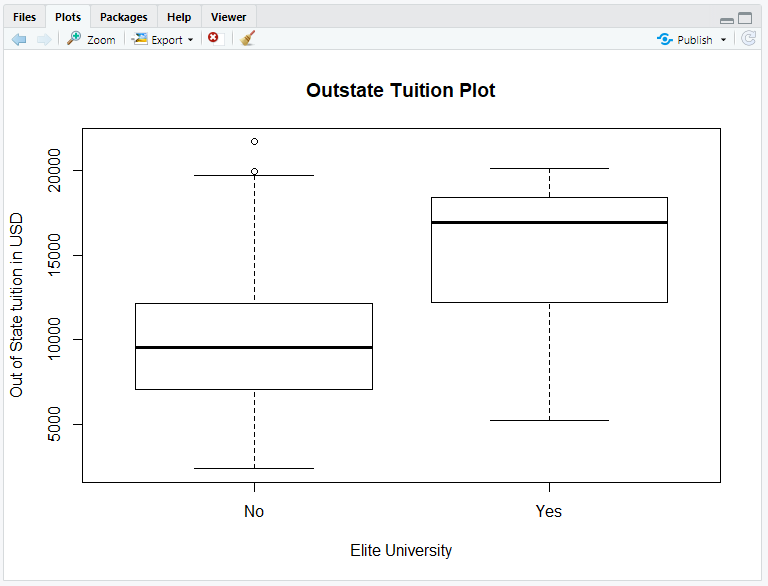
> summary(college$Elite)

There are 78 elite university

No Yes

699 78

> plot(college$Elite, college$Outstate, xlab = "Elite University", ylab ="Out of State tuition in USD", main = "Outstate Tuition Plot")



**v.** Use the hist() function to produce some histograms with numbers of bins for a few of the quantitative variables.

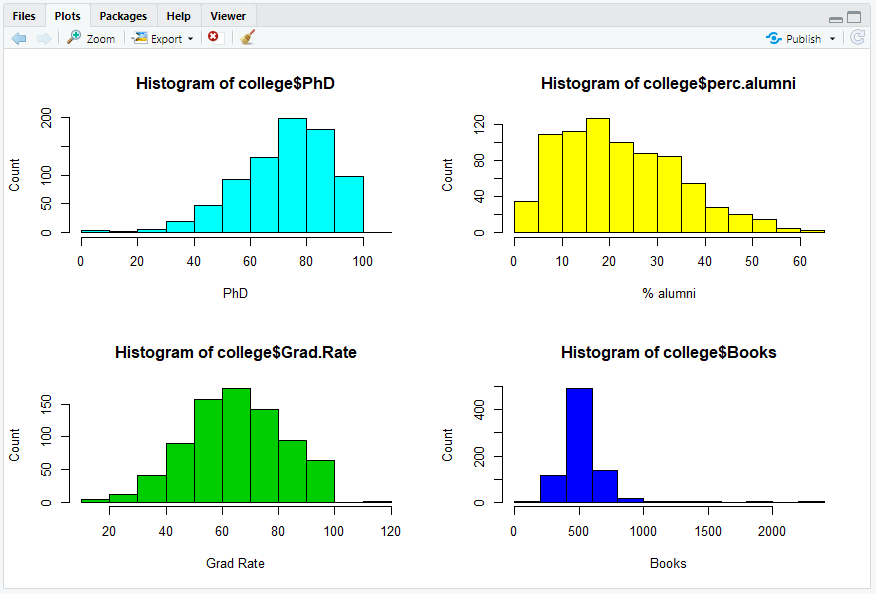
> par(mfrow = c(2,2))

> hist(college$PhD, col = 5, xlab = "PhD", ylab = "Count")

> hist(college$perc.alumni, col = 7, xlab = "% alumni", ylab = "Count")

> hist(college$Grad.Rate, col = 3, xlab = "Grad Rate", ylab = "Count")

> hist(college$Books, col = 4, xlab = "Books", ylab = "Count")



**vi.** Continue exploring the data, and provide a brief summary of what you discover.

> summary(college$Grad.Rate)

Min. 1st Qu. Median Mean 3rd Qu. Max.

10.00 53.00 65.00 65.46 78.00 118.00

Interesting to see which universities have 118% grad, also try to discover how many universities have similar percentages?

> explore.grad <- college[college$Grad.Rate == 118, ]

> nrow(explore.grad)

[1] 1

> rownames[as.numeric(rownames(explore.grad))]

[1] Cazenovia College

777 Levels: Abilene Christian University Adelphi University Adrian College ... York College of Pennsylvania