Problem Set 1: R, R Markdown, Conceptual Foundations of ML

Candidate Number: 12889

10 February 2021

Part 1: Short Answer Questions

1. Imagine you have been hired as a data consultant. Your client has given you the task of building a classifier for a new dataset they have constructed. In each of the following 5 scenarios, would you recommend a flexible statistical learning method or an inflexible approach? Why? (2-3 sentences per scenario)

   (a) There is a large sample size of $N = 5$ billion, a large number of predictors $p = 100,000$, and the client is limited in their computing resources. *We would want to use a non flexible method here. The non flexible model requires less computing power, and can performs well with both lots of predictors and large samples*

   b) Large sample size of $N = 5$ billion, and (small number of predictors $p = 6$. *We would want to use a less flexible model since we have small amount of predictors*

   (c) Large number of predictors, $p = 125,000$, sample size $N = 2000$ is relatively small. *A more flexible model should be applied here, to take advantage of more information we have on predictors.*

   (d) Based on exploratory analysis of the data, it appears that the predictors and the response have a non-linear relationship.*A flexible model since they are better at predicting non linear relationships*

   e) The error term has very large variance. An inflexible model.*The data set appears to have a lot of noise so if it is flexible it will probably be predicting noise and not signal*

2. How is a **parametric** approach different from a **non-parametric** approach to statistical learning? How does each approach go about estimating f? Name three advantages and three disadvantages of each approach. (2-3 sentences per approach) A parametric approach assumes linearity in the functional form of the model we are trying to measure while a non-parametric form does not make such assumptions. Advantages of parametric model It can more easily be used of inference. A linear assumption allows for clear understanding of how a predictors effect outcome. Parametric models require less data Parametric models are usually less resource intense. Disadvantages.
   Parametric models tend to be worse at predicting outcomes due to linear assumption which rarely holds in real life application Parametric models usually have more difficulty with categorical variables.

Non Parametric models Advantageous Are likely to be better at predicting because of less restrictive assumptions about the underlying function. Is usually better with extremely large amount of predictors. Is sometimes better for categorical variables. Such as knn use for Wikipedia and curse words

3. *ISL 2.4 Exercise 2*

(a) We would use a regression based approach, because the outcome we are interested in is CEO pay which is a continuous variable. We care more about inference because we want to see how certain variables effect CEO pay and there are a relatively small amount of predictors. The sample size is 500, our predictors are profits, number of employees, and industry.

(b) We would use a classifier based approach in this case since we care about a discrete value of success or failure. We are more interested in prediction because we are just trying to estimate if the new product will be a success or failure based on previous products, rather then understand the causes that will make it so. Our sample size in 20, and we have 13 predictors.

(C) We would us a regression based approach, percentage change is a continuous variable. We care more about prediction because the relationship between stock market and exchange rate is quite volatile and proving causality with our variables would be hard, and many of our predictors are collinear. Our sample size is 52 weeks of exchange rate data, and our predictors are the stock markets in the US, Germany, and the UK.

4. *ISL 2.4 Exercise 3*

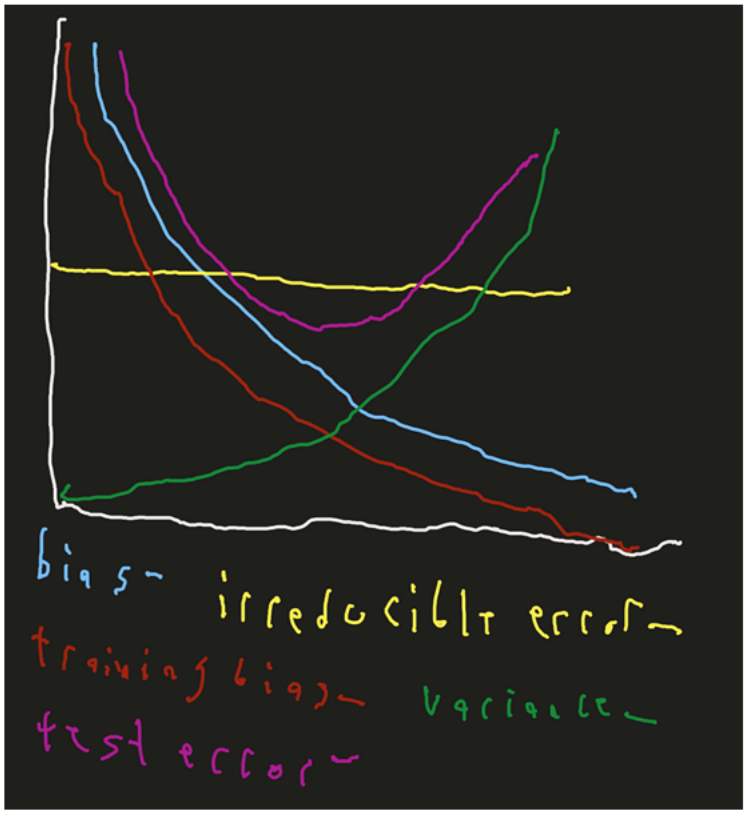


Figure 1: The irreducable error does not change with felxibility so it is a straight line. The bias is a Monotonic decreasing function of flexibility. The variance Monotonic increasing function as the flexibility increases. So does training error. Training error can go to zero if we massively overfit, but bias cannot. While test error encapsulates the bias variance trade off, so it is convex at the point where variance is increasesing faster then the bias is decreasing.

5. What are the two kinds of "big data" Rocio Titiunik wrote about in her paper on big data? What are some benefits and drawbacks of each kind of big data analysis for social scientific inquiry? Can either kind of big data solve the fundamental problem of causal inference? (5-10 sentences)

There is big data in n and big data in p. Big data in n is where the sample size is extremely large. While large p means a large amount of predictors. The benefits of big n is that it can increase the precision of our

models, as well as increase the significance of our hypothesis tests. However large data sets still does not always mean our attempt to prove casual inference will be correct. If our estimator is inconsistent, to start with, adding more data would not fix the issue. The hope with a large number of predictors is that we can capture all the variables of cause and effect, and eliminate items such as omitted variable bias. However the catch with this is that we cannot determine if our model actually captures all casual variables as, well as ignores variables that are effected by the treatment. Large p and large n allow us to have more powerful descriptive analysis and a resource to draw theory from. However research design and theory to back up our models is still needed to prove causal inference. ## Part 2: Coding Questions

6. In the next problem set, we will use `for` loops and `if/else` statements to implement k-fold cross-validation. To prepare you for this, we'll practice them using the fibbonacci sequence. The fibbonacci sequence is a sequence where each number is the sum of the two preceding ones: $(0,)1, 1, 2, 3, 5, \ldots$. Using `for` loops and `if/else` statements, write code that will output the sum of the first 50 terms of the fibbonacci sequence. Include zero as the first term.

```
vec<-as.numeric(0:50)
for(i in 0:51){ifelse(vec[i]<2, vec[i]<-vec[i], vec[i]<-vec[i-1]+vec[i-2])}
sum(vec)
```

```
## [1] 32951280098
```

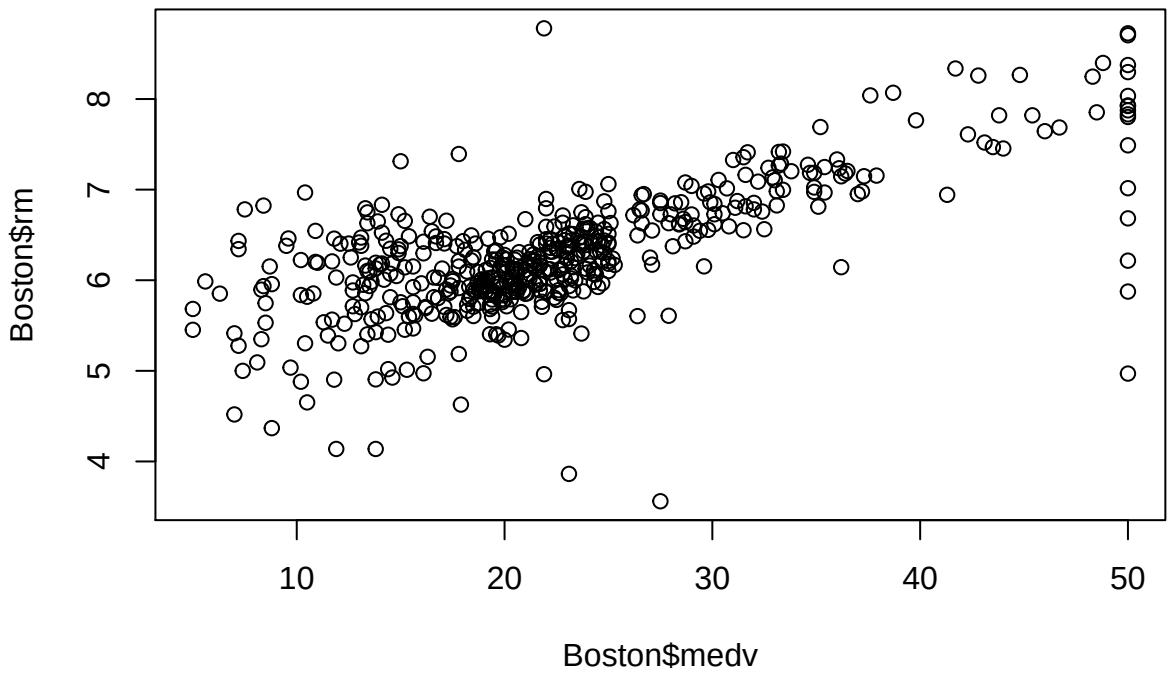
7. *ISL 2.4 Exercise 10* (Note: 1. You will need to install the `MASS` library from CRAN. 2. Please break text out of code blocks when explaining or reporting your answers.)

```
# Code for 10 a) goes here
library(MASS)
```

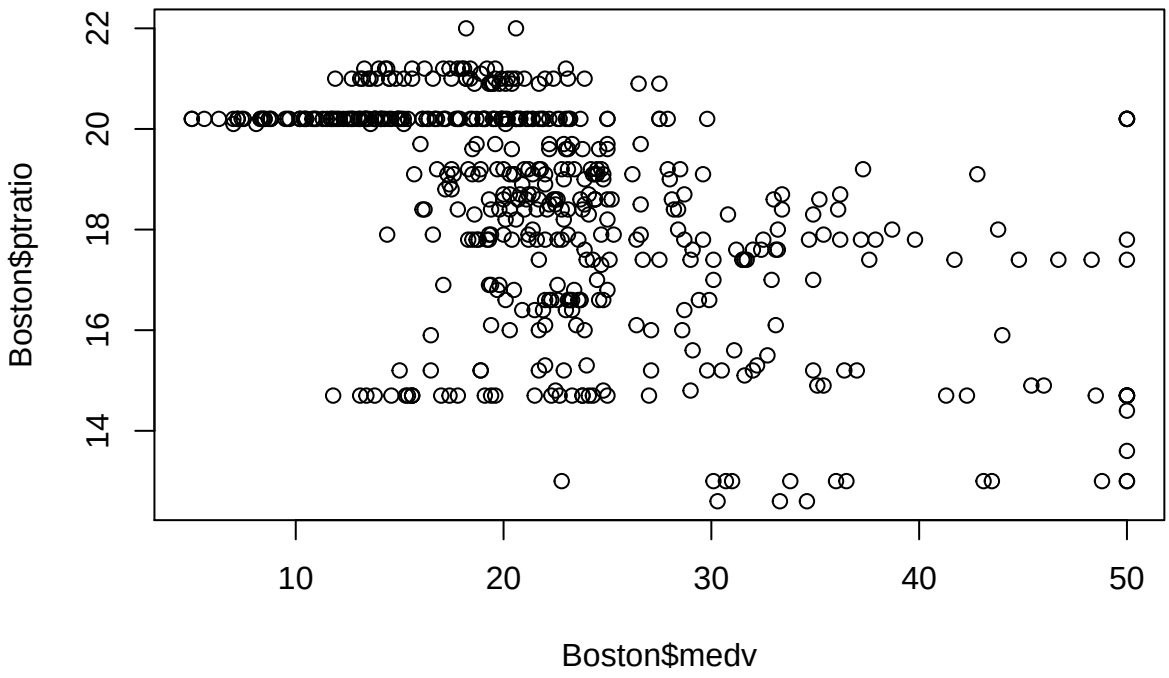
```
## Warning: package 'MASS' was built under R version 4.0.3
```

There are 506 rows and 14 columns. the rows are tracts for different parts of Boston The columns are variables such as average room per #dwelling and property value.

```
# Code for 10 b) goes here
 plot(Boston$medv, Boston$rm)
```

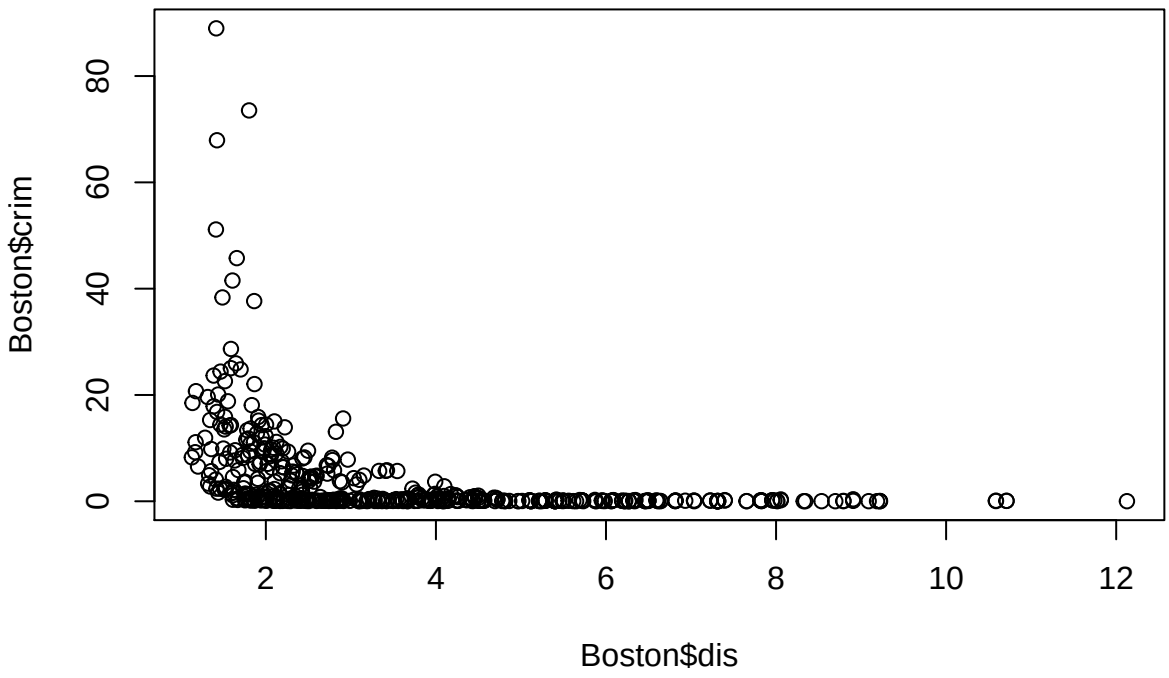



```
plot(Boston$medv, Boston$ptratio)
```

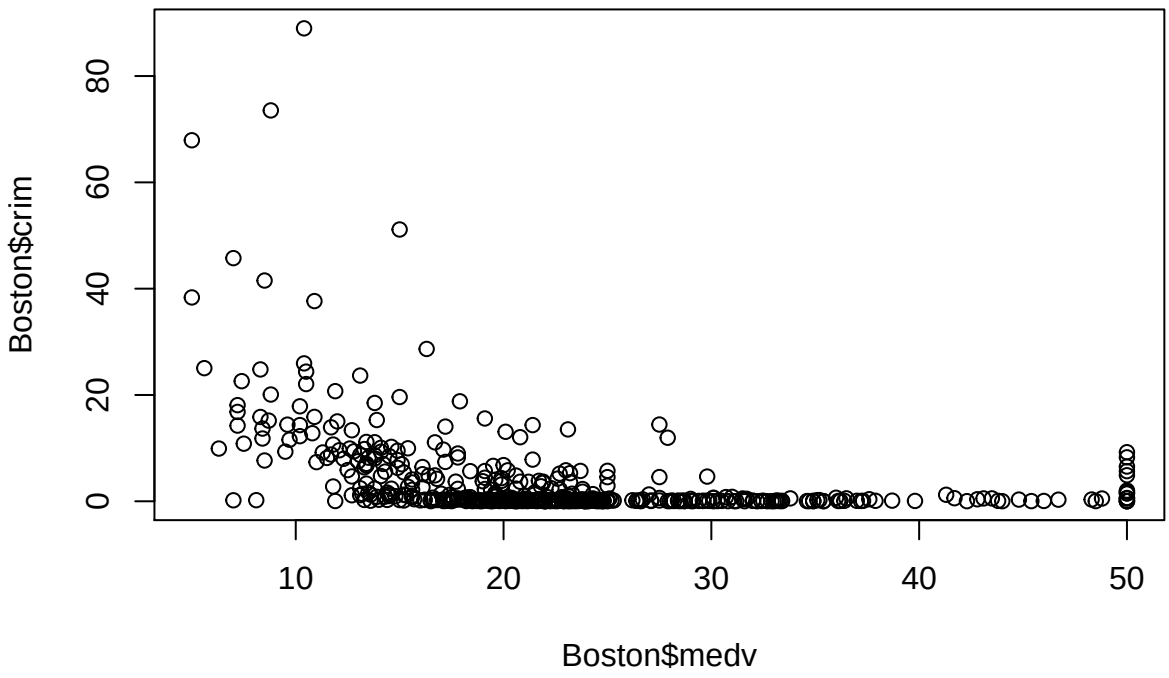


There is a strong positive correlation between housing prices and number of dwelling rooms. There is also a negative correlation between median housing price and student teacher ratio

```
# Code for 10 c) goes here
 plot(Boston$dis, Boston$crim)
```




```
plot(Boston$medv, Boston$crim)
```



median housing values does have a negative correlation with crime rates. As neighborhoods level of income increases the level of crime will probably decrease because individuals are not as impoverished

```
# Code for 10 d) goes here
summary(Boston)
```

```
##       crim                zn              indus            chas        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
##  1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000  
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917  
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000  
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000  
##       nox               rm              age              dis        
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795  
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
##       rad              tax           ptratio          black       
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32  
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38  
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44  
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67  
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23
```

```
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90
##      lstat            medv
##  Min.   : 1.73   Min.   : 5.00
##  1st Qu.: 6.95   1st Qu.:17.02
##  Median :11.36   Median :21.20
##  Mean   :12.65   Mean   :22.53
##  3rd Qu.:16.95   3rd Qu.:25.00
##  Max.   :37.97   Max.   :50.00
```

Shows the range on predictors. Crime is particularly high in some suburbs. The mean crime rate is 3.6 but there is one suburb which the value of 88 this is not true for tax, where the mean appears to be in the middle The same is true with teacher student ratio

```
# Code for 10 e) goes here
bordriv<-Boston$chas
sum(bordriv)
```

```
## [1] 35
```

the answer is 35

```
# Code for 10 f) goes here
summary(Boston)
```

```
##       crim                zn              indus            chas
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000
##  1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000
##       nox               rm              age              dis
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127
##       rad              tax           ptratio          black
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90
##      lstat            medv
##  Min.   : 1.73   Min.   : 5.00
##  1st Qu.: 6.95   1st Qu.:17.02
##  Median :11.36   Median :21.20
##  Mean   :12.65   Mean   :22.53
##  3rd Qu.:16.95   3rd Qu.:25.00
##  Max.   :37.97   Max.   :50.00
```

19.05

```
# Code for 10 g) goes here
Boston[which(Boston$medv==min(Boston$medv)),]
```

```
##        crim zn indus chas   nox    rm age    dis rad tax ptratio  black lstat
## 399 38.3518  0  18.1    0 0.693 5.453 100 1.4896  24 666    20.2 396.90 30.59
## 406 67.9208  0  18.1    0 0.693 5.683 100 1.4254  24 666    20.2 384.97 22.98
##     medv
## 399    5
## 406    5
```

there are two lots that have the lowest median value of owner-occupied homes. These two lots are 399, and 406. These areas have much higher then average crime rates, as well as higher student to teacher ratios. These are indicators that are strongly correlated to low income neighborhoods.

```
# Code for 10 h) goes here
big7=Boston$rm[which(Boston$rm>7)]
length(big7)
```

```
## [1] 64
```

```
big8=Boston$rm[which(Boston$rm>8)]
length(big8)
```

```
## [1] 13
```

8. Using R Markdown, write some notes on the differences between supervised and unsupervised approaches to statistical learning. Use headers of different sizes, italic and bold text, numbered lists, bullet lists, and hyperlinks. If you would like, use inline LaTeX (math notation).

Supervised and Unsupervised Learning

Unsurprised Learning

1. We are trying to observe characteristics of our data set.
    - We want to observe clustering if our data set is categorical possible method (kmeans)
    - We would want to reduce dimensions if our data set is numerical, data set is numerical Possible method (PCA)

Supervised Learning

1. We have a response variable for our data. We are trying to estimate f(x) that maps D to Y
    - regression if the the response variable is continuous example method (OLS)
    - classier if the response variable is categorical, example method (Knn model)