Stats 101C Homework 1

Charles Liu (304804942)

10/16/2020

Loading Necessary Packages:

```
library(MASS)
```

Problem 1 (Exercise 5) 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?

ANSWERS BELOW:

Advantages of a very flexible model:

1. Possibly gives a better fit for non-linear models
2. Higher flexibility means lower bias
3. Gives a more complex model, if needed

Disadvantages of a very flexible model:

1. Can cause the training data to be overfit (follows noise too closely)
2. Higher flexibility means higher variance
3. It causes the model to estimate a greater number of predictors (not always effectient)

"As for less flexible model, it is just the opposite of a very flexible model"

Under the circumstances:

A more flexible model approach would be preferred to a less flexible approach when we are more interested in predicting the results rather than interpreting the results. It's also useful when the model's relationship is non-linear, many data points to locate a pattern from, and when it has a low irreducible error.

A less flexible model approach would be preferred to a more flexible approach when we are more interested in interpreting the results rather than predicting the results. It's also useful when the model's relationship is linear, less data points to locate the linear pattern, and when it has a high irreducible error.

Problem 2 (Exercise 6) 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?

ANSWERS BELOW:

Parametric Approach:

1. Assumes a form or shape for f
2. Uses a model-based approach for finding f
3. Estimate f down to a single set of parameters needed for the model
4. Advantages for regression or classification are simplifying the model f to a few parameters and not as many observations are needed, compared to non-parametric approach
5. Disadvantages for regression or classification are inaccuracy if the form/shape of f is assume incorrectly and it'll overfit the observations if you use a very flexible model

Non-Parametric Approach:

1. Does **NOT** make any (or little) assumptions about the form of f
2. Requires more samples to better estimate f
3. Great for fitting non-linear patterns from the model
4. Advantage for regression or classification is it is better for fitting non-linear models
5. Disadvantage for regression or classification is it does not offer easy interpretability for the data because little to no assumptions have been made on the model

Problem 3 (Exercise 10) 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 MASS library in R.Now the data set is contained in the object Boston. Read about the data set. How many rows are in this data set? How many columns? What do the rows and columns represent?

```r
attach(Boston)
dim(Boston)
```

```
## [1] 506  14
```

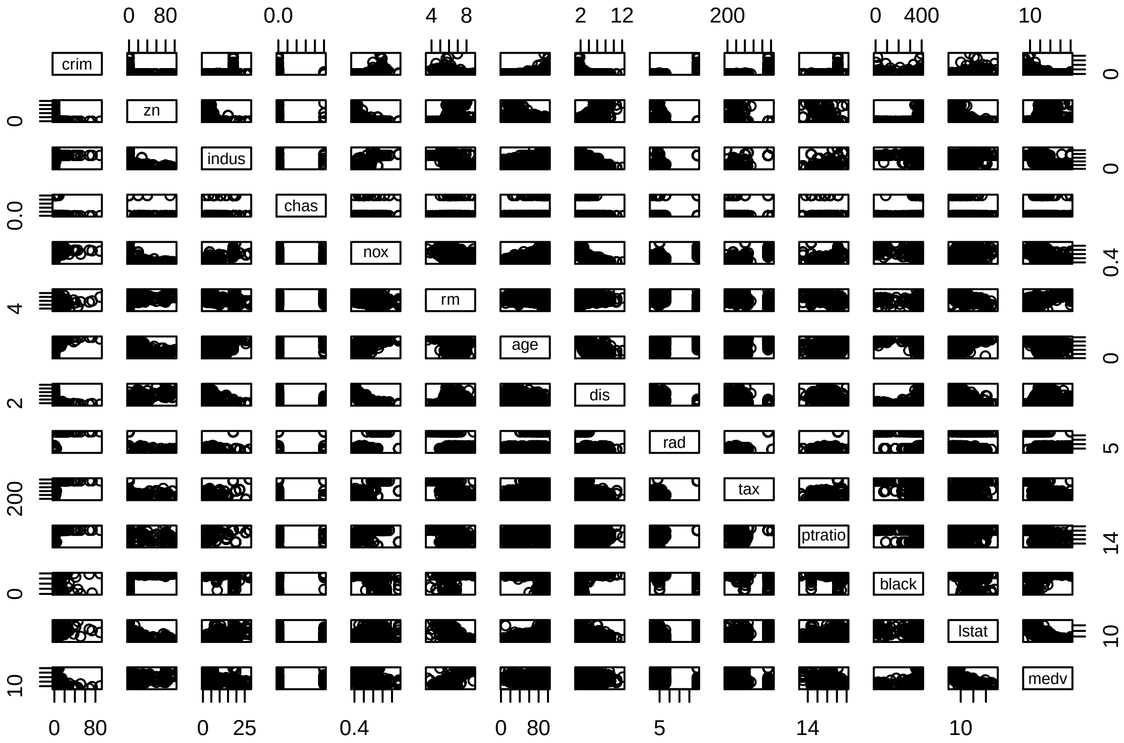
```r
names(Boston)
```

```
##  [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"
##  [8] "dis"     "rad"     "tax"     "ptratio" "black"   "lstat"   "medv"
```

ANSWER: There are 506 rows with 14 columns. The rows are the observations, and the columns are the variables (or the predictors).

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

```r
pairs(Boston)
```



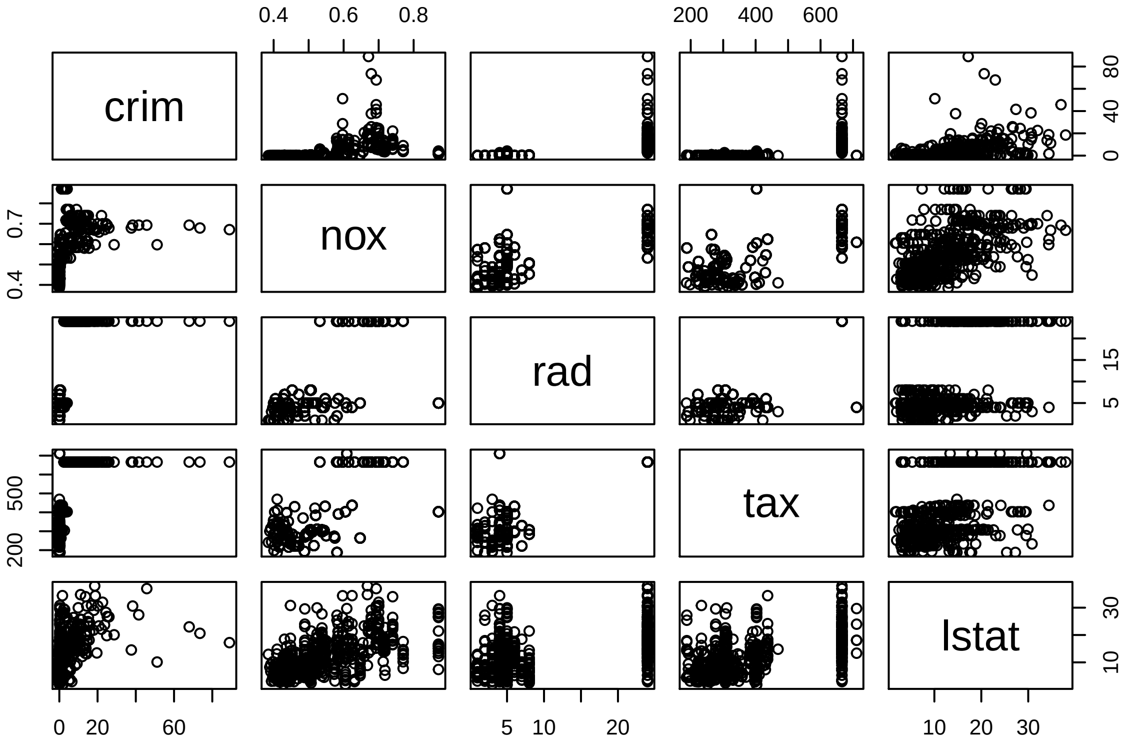
COMMENTS: From our findings, we can see a high level of **rad** contained in our **cri** parameter.The **medv** paramater seems to have an inverse relationship between **indus**, **lstat**, and **nox**. We would need to observe further into the data to see why this may be the case. We also see a relationship proportional for **rm** and **medv**.

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

```r
corr_Boston <- cor(Boston)
# We want to see the predictos associated with "crim"
corr_Boston_crim <- corr_Boston[-1, 1]
corr_Boston_crim
```

```
##          zn       indus        chas         nox          rm         age
## -0.20046922  0.40658341 -0.05589158  0.42097171 -0.21924670  0.35273425
##         dis         rad         tax     ptratio       black       lstat
## -0.37967009  0.62550515  0.58276431  0.28994558 -0.38506394  0.45562148
##        medv
## -0.38830461
```

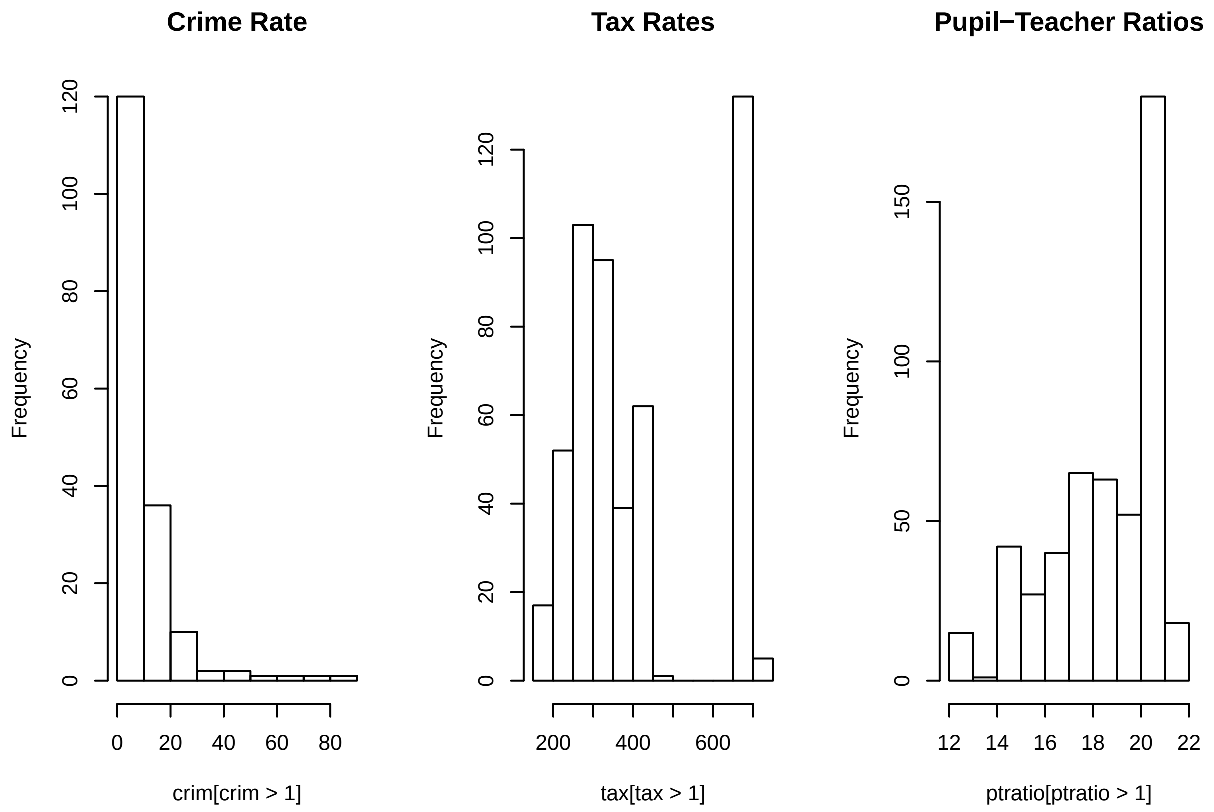
```r
# We see the ones with the highest correlation with "crim" are: "rad", "tax", lstat", and "nox"

pairs(Boston[, c(1, 5, 9, 10, 13)])
```



COMMENTS: We can see here (both numerically and graphically) that “nox”, “rad”, “tax” and “lstat” have the highest correlation with “crim” parameter. To also note, “rad” and “tax” has the strongest correlation with “crim”, meanwhile “lstat” and “nox” have a medium correlation with “crim”. They are all proportionally related with each other.

(d) Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.

```
par(mfrow=c(1,3))
hist(crim[crim > 1], breaks = 10, main = "Crime Rate")
hist(tax[tax > 1], breaks = 10, main = "Tax Rates")
hist(ptratio[ptratio > 1], breaks = 10, main = "Pupil-Teacher Ratios")
```



COMMENTS: We see that for the most part, the cities have lower crime rates. However, there is a long tail that shows 18 suburbs that appears to have a crime rate greater than 20 (frequency). We can see that there is a certain "divide" between the "Tax Rates" for people in suburbs. The divide is between people who have tax rates 600+ and less than 500 rate-wise. We can see the "Pupil-Teacher Ratios" is somewhat skewed to the left, but there is not any particularly high ratios present.

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

```
sum(chas == 1)
```

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

ANSWER: There is a total of 35 suburbs in this data set bound the Charles river.

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

```
median(ptratio)
```

```
## [1] 19.05
```

ANSWER: The median pupil-teacher ratio among the towns in this data set is 19.05.

(g) Which suburb of Boston has lowest median value of owneroccupied homes? What are the values of the other predictors for that suburb, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

```
# We see observation "399" has the minimum, but let's observe further.
which.min(medv)
```

```
## [1] 399
```

```
summary(Boston)
```

```
##       crim                zn             indus            chas        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
##  1st Qu.: 0.08204   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  
```

```
# Find the minimum "medv"
min_medv <- Boston[medv == min(medv), ]
# Find the range for those predictors
range_Boston <- sapply(Boston, range)

# Combine it all together and rename the items
exercise10_g <- rbind(min_medv, range_Boston)
rownames(exercise10_g) <- c("1st Lowest medv", "2nd Lowest medv", "Minimum", "Maximum")
exercise10_g
```

```
##                     crim zn indus chas   nox    rm   age    dis rad tax
## 1st Lowest medv 38.35180  0 18.10    0 0.693 5.453 100.0 1.4896  24 666
## 2nd Lowest medv 67.92080  0 18.10    0 0.693 5.683 100.0 1.4254  24 666
## Minimum          0.00632  0  0.46    0 0.385 3.561   2.9 1.1296   1 187
## Maximum         88.97620 100 27.74    1 0.871 8.780 100.0 12.1265  24 711
##                 ptratio  black lstat medv
## 1st Lowest medv    20.2 396.90 30.59    5
## 2nd Lowest medv    20.2 384.97 22.98    5
## Minimum            12.6   0.32  1.73    5
## Maximum            22.0 396.90 37.97   50
```

COMMENTS: We see that "crim", "indus", "tax", "ptratio", "black", and "lstat" is rather high for theese two suburbs with the lowest "medv". We also noticed that "age" and "rad" are equal to the Maximum amount. These two suburbs are very similar across all the predictors.

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

```
# Find the amounts
sum(rm > 7)
```

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

```
sum(rm > 8)
```

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

```
# Find the range for the "rm" greater than 8 rooms
range_Boston_eight <- sapply(Boston[rm > 8, ], range)
# Find the range for those predictors
range_Boston <- sapply(Boston, range)

# Combine it all together and rename the items
exercise10_h <- rbind(range_Boston_eight, range_Boston)
rownames(exercise10_h) <- c("Minimum (greater than 8 Rooms)", "Maximum (greater than 8 Rooms)", "Minimu
exercise10_h
```

```
##                                     crim  zn indus chas    nox    rm   age
## Minimum (greater than 8 Rooms)  0.02009   0  2.68    0 0.4161 8.034   8.4
## Maximum (greater than 8 Rooms)  3.47428  95 19.58    1 0.7180 8.780  93.9
## Minimum (Boston Dataset)        0.00632   0  0.46    0 0.3850 3.561   2.9
## Maximum (Boston Dataset        88.97620 100 27.74    1 0.8710 8.780 100.0
##                                     dis rad tax ptratio  black lstat medv
## Minimum (greater than 8 Rooms)  1.8010   2 224    13.0 354.55  2.47 21.9
## Maximum (greater than 8 Rooms)  8.9067  24 666    20.2 396.90  7.44 50.0
## Minimum (Boston Dataset)        1.1296   1 187    12.6   0.32  1.73  5.0
## Maximum (Boston Dataset        12.1265  24 711    22.0 396.90 37.97 50.0
```

ANSWER: We see there are 64 of the suburbs average more than 7 rooms per dwelling. We also see there are 13 of the suburbs average more than 8 rooms per dwelling.

COMMENTS: From this, we can see that there is a lower "crim" and "lstat" when we compare the ranges. Judging from how small the sample set is, the ranges are relatively similar to the Boston dataset in its entirety. The only major difference we see here is the "black" variable has a far greater Minimum for "greater than 8 Rooms" than the "Boston Dataset".