Solution to Homework 1

Thanks to Yuwen Zhang

1 Problem 1

(a) It is a regression problem. And we are more interested in inference. n = 500, p = 3

(b) It is a classification problem. And we are more interested in prediction. n = 20, p = 13

(c) It is a regression problem. And we are more interested in prediction. n = 366 / 7 = 52, n = 3

2 Problem 2

Advantages for a very flexible approach:

(1) model is richer and adapt to change. We can keep improving the model as we add more data to fit while the less flexible one has a limit of fit quality.

(2) The model can catch some local variations compared to a less flexible one.

(3) less bias compared to a less flexible one

Disadvantage for a very flexible approach:

(1) hard to interpret compared to less flexible one

(2) may have overfit problems that will influence accuracy

(3) more variance compared to a less flexible one

If we are only interested in prediction accuracy and the interpretability of the predictive model is not a concern, a more flexible models can be a good choice, though not always in terms of prediction accuracy. But if we are mainly interested in inference, the restrictive/less flexible models are more interpretable and preferred.

3 Problem 3

(a) (1) **E-mail spam**: we can classify an e-mail to be a spam or not so we can deliver it to the junk folder. The predictor (input variable) is the body text of the e-mail. The response is whether the e-mail is spam or not (0/1). This is a prediction problem, because our concern is to classify an e-mail and we don't really care the relationship between the predictor and the response.

(2) **Image classification**: we can classify a picture to be a cat (1) or not (0). The predictor is the RGB image (length × width × channel) and the response is whether the picture is a cat or not. This is a prediction problem, since we don't really need to know the relation between the input and the output.

(3) **product classification**: we can distinguish products between several main product categories by giving some features (price, sales volume...). The predictors are the features and the response is the categories (0,1,2,...etc.). It's a prediction problem.

(b) (1) **House price prediction**: we can predict housing sales prices based on predictors like location, area of housing, housing utility, year built and several other features. The response will be the housing sale price. Since the price is continuous, it is a regression problem. It's also a prediction problem since we care about the prediction result more than the relationships between the result and the features.

(2) **Taxi fare prediction**: we can predict the fare amount for a taxi ride in New York City given the pickup and dropoff locations. The predictors can be pickup time, pickup location, dropoff time, dropoff location and passenger count. The response is the fare amount. It is a prediction problem.

(3) **Fire damage and distance to fire station**: for a fire insurance company, it wants to relate the amount of fire damage in major residential fires to the distance between the burning house and the nearest fire station. The predictor is the distance from fire station and the response is the extent of damage (represented by continuous numbers). It is an inference problem (a simple linear regression problem) because we want to know the relationship between the distance and the damage for further analysis.

(c) (1) **Market segmentation**: We can cluster customers based on the database so we can use this information to start some target marketing programs.

(2) **Friend recommendation**: We can build friends connection network graph by clustering users and recommend new friends to them based on the connection of their nearest friends.

(3) **Biological taxonomy**: We can cluster creatures into plants and animals, and in animal cluster, they can be further clustered into with

spine and no spine ... So if we have a new creature, we can figure out which cluster it is in by computing and comparing the distance of it from each cluster center.

4 Problem 4

(a) Euclidean distance:

$$d(\mathbf{p},\mathbf{q}) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

Observation 1:

$$d1 = \sqrt{(0-0)^2 + (3-0)^2 + (0-0)^2} = 3$$

Observation 2:

$$d2 = \sqrt{(2-0)^2 + (0-0)^2 + (0-0)^2} = 2$$

Observation 3:

$$d3 = \sqrt{(0-0)^2 + (1-0)^2 + (3-0)^2} = 3.16$$

Observation 4:

$$d4 = \sqrt{(0-0)^2 + (1-0)^2 + (2-0)^2} = 2.236$$

Observation 5:

$$d5 = \sqrt{(-1-0)^2 + (0-0)^2 + (1-0)^2} = 1.414$$

Observation 6:

$$d6 = \sqrt{(1-0)^2 + (1-0)^2 + (1-0)^2} = 1.732$$

(b) For KNN classifier, we know that:

$$Pr(Y = j|X = x_0) = \frac{1}{K}\sum_{i \in N_0} I(y_i = j)$$

When $K = 1$, we pick up observation 5:

$$Pr(Y = Green|X = x_0) = 1$$

So we can predict that Y should be green.

(c) When $K = 3$, we pick up observation 5, 6, 2

$$Pr(Y = Green|X = x_0) = \frac{1}{3}$$

$$Pr(Y = Red|X = x_0) = \frac{2}{3}$$

So we can predict that Y should be red.

(d) From textbook P39 - P40 we know that, as $\frac{1}{K}$ increases, the method becomes more flexible (non-linear), so to get a highly non-linear Bayes decision boundary, we need smaller K to get larger $\frac{1}{K}$.

5 Problem 5

(a) There are 506 rows and 14 columns in this data set. The row represents each house sample and the column represents each feature (i.e. crim, zn, indus, chas ...).

(b) We run `pairs(Boston)` and please see attached page for the scatterplot. From this figure we can see the relationship between different columns (features). For example, there seem to be some observable relationships between rm and medv, lstat and medv, nox and dis.
When rm goes up, medv goes up accordingly.
When lstat goes up, medv goes down accordingly.
When nox goes up, dis goes down accordingly.

(c) We can see that there are several predictors associated with per capita crime rate. i.e. nox, rm, age, dis.

(1) For age (portion of owner-occupied units built prior to 1940), when age increases, the crime rate increases as well.

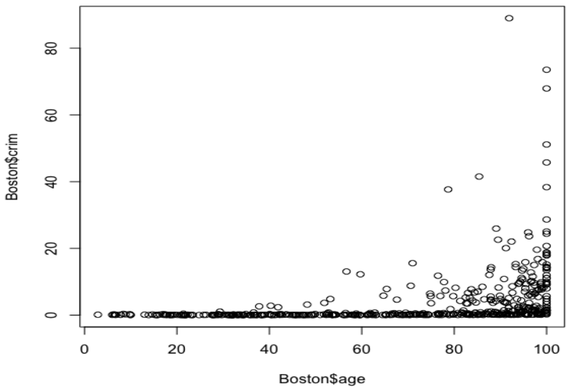

Figure 1: age vs. crim

(2) For dis (weighted mean of distances to five Boston employment centres), when distance increases, the crime rate decreases.

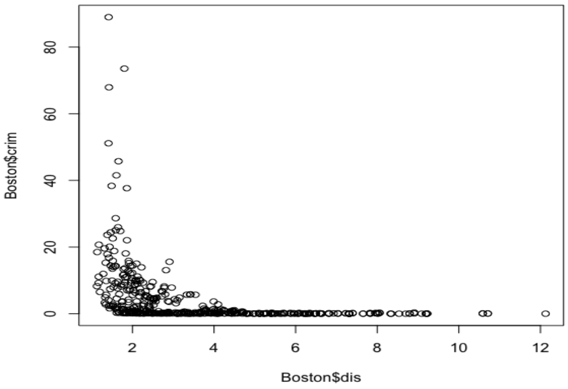

Figure 2: dis vs. crim

(d) (1) For crime rate, we can see from histogram figure 5 that most suburbs have relatively low crime rate while the frequency for higher crime rates is relatively low and not significant. I divide the range into 20 breaks and the distribution can be seen from figure 5.

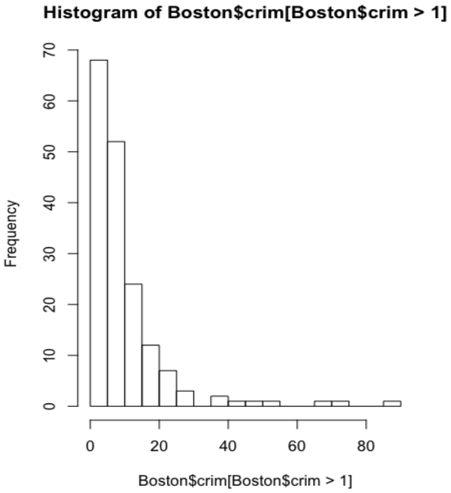

Figure 3: crime rate

(2) For tax rates, we can see from histogram figure 6 that there are suburbs with particular high tax rates. I divide the range into 20 breaks and the distribution can be seen from figure 6.

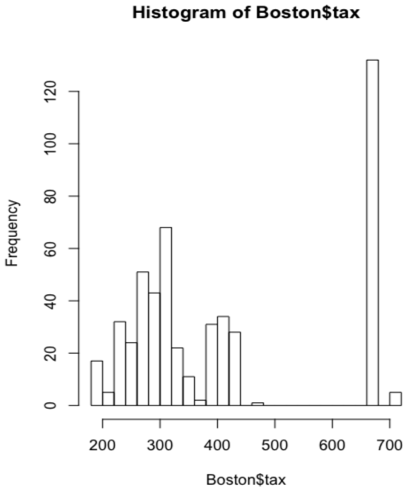

Figure 4: tax

(3) For pupil-teacher ratios, we can see from histogram figure 7 that there are quite a few moderately high pupil ratios around 20, but no particularly high ones. I divide the range into 20 breaks and the distribution can be seen from figure 7.

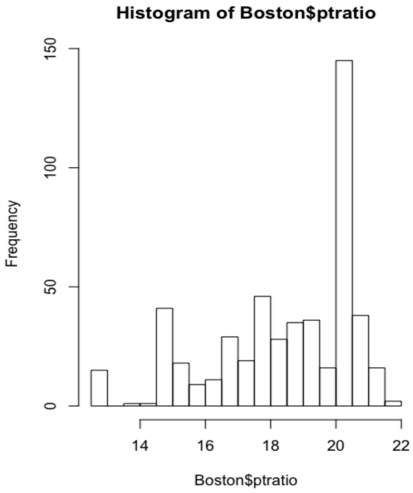

Figure 5: pupil teach ratio

(e) From the data set explanation we know that, when chas == 1, the suburb is bound to the Charles river. We run subset(Boston, chas == 1) and figure out the dimension by dim. The result is that there are 35 suburbs in this data set bound the Charles river.

(f) We run median(Boston$ptratio) to figure out the median value is 19.05

(g) First we figure out that the lowest medv is 5, and then we run subset(Boston, medv = 5) to figure out the suburbs with the lowest median value of owner-occupied homes. And here are two results:

	399	406
crim	38.3518	67.9208
zn	0.0000	0.0000
indus	18.1000	18.1000
chas	0.0000	0.0000
nox	0.6930	0.6930
rm	5.4530	5.6830
age	100.0000	100.0000
dis	1.4896	1.4254
rad	24.0000	24.0000
tax	666.0000	666.0000
ptratio	20.2000	20.2000
black	396.9000	384.9700
lstat	30.5900	22.9800
medv	5.0000	5.0000

We run summary(Boston) to get the overall range of predictors:

```
     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
```

For the crime rate, both 399 and 406 are relatively high (38%, 48% respectively), and they fall in the range of low frequency for this predictor.
For zn, both of them are zero and falls into the highest frequency range of this predictor.
For indus, both of them falls into the highest frequency range of this predictor.
For chas, both of them are zero and falls into the highest frequency range of this predictor.
For nox, both of them are relatively high and larger than the 3rd Qu.
For rm, both of them are relatively low and smaller than 1st Qu.
For age, both of them are maximum value.
Similar analysis can be done on other features like dis, rad, tax, ptratio, black, lstat and medv. So we find that the suburbs with lowest median value of owner-occupied homes have higher crime rate, and the proportion of residential land zoned for lots over 25,000 sq.ft are both zero, and they are both not bound to Charles river and etc..

(h) There are 64 suburbs average more than seven rooms per dwelling. And there are 13 suburbs average more than eight rooms per dwelling. We did a summary of suburbs that average more than eight rooms per dwelling and the results are as follows:

```
      crim               zn             indus             chas
 Min.   :0.02009   Min.   : 0.00   Min.   : 2.680   Min.   :0.0000
 1st Qu.:0.33147   1st Qu.: 0.00   1st Qu.: 3.970   1st Qu.:0.0000
 Median :0.52014   Median : 0.00   Median : 6.200   Median :0.0000
 Mean   :0.71879   Mean   :13.62   Mean   : 7.078   Mean   :0.1538
 3rd Qu.:0.57834   3rd Qu.:20.00   3rd Qu.: 6.200   3rd Qu.:0.0000
 Max.   :3.47428   Max.   :95.00   Max.   :19.580   Max.   :1.0000
      nox               rm             age             dis
 Min.   :0.4161   Min.   :8.034   Min.   : 8.40   Min.   :1.801
 1st Qu.:0.5040   1st Qu.:8.247   1st Qu.:70.40   1st Qu.:2.288
 Median :0.5070   Median :8.297   Median :78.30   Median :2.894
 Mean   :0.5392   Mean   :8.349   Mean   :71.54   Mean   :3.430
 3rd Qu.:0.6050   3rd Qu.:8.398   3rd Qu.:86.50   3rd Qu.:3.652
 Max.   :0.7180   Max.   :8.780   Max.   :93.90   Max.   :8.907
```

```
      rad              tax           ptratio          black      
 Min.   : 2.000   Min.   :224.0   Min.   :13.00   Min.   :354.6  
 1st Qu.: 5.000   1st Qu.:264.0   1st Qu.:14.70   1st Qu.:384.5  
 Median : 7.000   Median :307.0   Median :17.40   Median :386.9  
 Mean   : 7.462   Mean   :325.1   Mean   :16.36   Mean   :385.2  
 3rd Qu.: 8.000   3rd Qu.:307.0   3rd Qu.:17.40   3rd Qu.:389.7  
 Max.   :24.000   Max.   :666.0   Max.   :20.20   Max.   :396.9  
     lstat           medv     
 Min.   :2.47   Min.   :21.9  
 1st Qu.:3.32   1st Qu.:41.7  
 Median :4.14   Median :48.3  
 Mean   :4.31   Mean   :44.2  
 3rd Qu.:5.12   3rd Qu.:50.0  
 Max.   :7.44   Max.   :50.0  
```

Compared with the summary of the whole data set we can see that, the mean of crime rate goes down, the mean of zn goes up, the mean of indus goes down, the mean of chas goes up, the mean of nox goes up, the mean of rm goes up, the mean of age goes down, the mean of dis goes up, the mean of rad goes down....
We find that suburbs with more than eight rooms have relatively low crime rate, more average number of rooms per dwelling, larger distance from five Boston employment center and etc..