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

Candidate Number:14318

03 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.
   I would recommend an inflexible approach because: i) The large number of predictors creates a high potential for a flexible approach to overfit the model, even with the relatively large number of observations. This is because the flexible approaches are more likely to overfit a model on high dimensional data. Overfitting refers to a situation where the model attributes predictive power to parameters based on noise in the data rather than true properties of the unknown function. ii) Flexible approaches are more computationally expensive than inflexible approaches. Given the large number of observations and dimensions of the data, it is likely that a flexible approach would exceed the computational resource limits of the client.

   b) Large sample size of $N = 5$ billion, and small number of predictors $p = 6$.
   I would recommend a flexible learning method because: i) The low number of dimensions means flexible approaches are likely to be able to quickly produce the most accurate model without overfitting. ii) The client has not indicated a preference for an interpretable model - a major drawback of inflexible methods is their lack of interpratibility however this does not appear to be relevant in this case.

   c) Large number of predictors, $p = 125,000$, sample size $N = 2000$ is relatively small.
   I would recommend an inflexible approach because the low ratio of observations to dimensions creates a lot of potential for a flexible model to overfit the data. Therefore, an inflexible model with pre-chosen parameters of interest would likely produce a better model.

   d) Based on exploratory analysis of the data, it appears that the predictors and the response have a non-linear relationship. I would recommend a flexible learning method because flexible models

   are better able to estimate parameters for non-linear data. This is because flexible models can fit many different possible functional forms for f and so they are more likely to find the function which accurately models the non-linear data.

   e) The error term has very large variance.
   I would recommend an inflexible model. When the the error term has very large variance, flexible learning methods are likely to capture variance-related noise in the functional form. An inflexible model is less likely to suffer from this issue so it is preferable in this case.

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)

Parametric approaches make assumptions about the shape of F and use a training dataset to fit the model.

Advantages
1. Assuming a parametric form f simplifies the problem of estimating f because it is generally much easier to estimate a set of parameters in the linear model than it is to fit an entirely arbitrary function f.
2. More interpretable
3. When the functional form is well known or we would like to see how well a pre-defined set of parameters predict an outcome, a parametric model is preferable

Disadvantages
1. The potential disadvantage of a parametric approach is that the model we choose will usually not match the true unknown form of f. If the chosen model is too far from the true f, then our estimate will be poor.
2. Does not exploit the full predictive power in the data

Non-parametric methods do not make explicit assumptions about the functional form of f
Advantages
1. By avoiding the assumption of a particular functional form for f, they have the potential to accurately fit a wider range of possible shapes for f.
2. Do not require expert knowledge to assume functional form of f

Disadvantages
1. Large number of observations is required
2. Larger potential for overfitting
3. Low interpretability

3. *ISL 2.4 Exercise 2*

2a) Regression, inference
b) Classification, prediction
c) Regression, prediction

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)

a) large n and large p

b) Large n - more predictive power but more computationally expensive, large P - able to investigate more theories through observational evidence, able to make better predictions by including more parameters in the model but still not able to make causal claims since not all important parameters can be measured.
c) big data is unlikely to solve the fundamental problem of causal inference since large n still requires judgment of which paremteres are of interest and large P cannot (yet) capture all important parameters

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.

```
Fibonacci <- numeric(50)
Fibonacci[1] <- 0
Fibonacci[2] <- 1
for (i in 3:50)
  Fibonacci[i] <- Fibonacci[i-2] + Fibonacci[i-1]
  sum(Fibonacci)
```

```
## [1] 20365011073
```

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.)

```
library(MASS)
head(Boston)
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21
##   medv
## 1 24.0
## 2 21.6
## 3 34.7
## 4 33.4
## 5 36.2
## 6 28.7
```

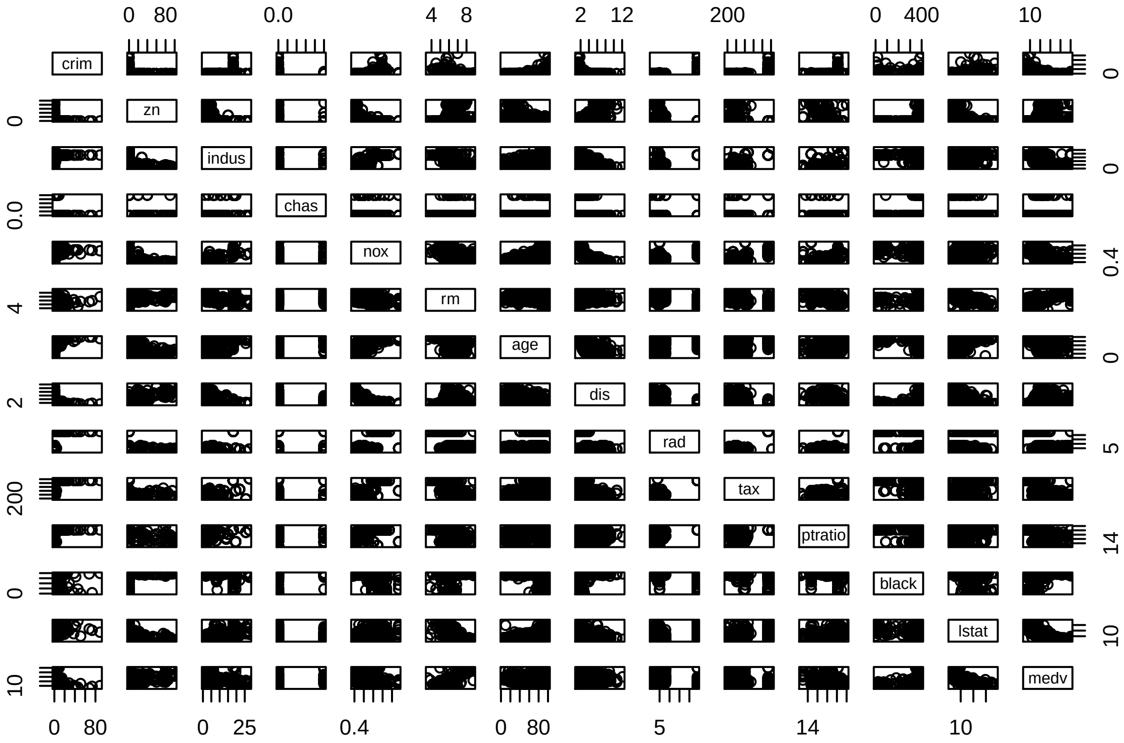
```
nrow(Boston)
```

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

```
ncol(Boston)
```

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

```
# Code for 10 b) goes here
Boston$chas <- as.numeric(Boston$chas)
Boston$rad <- as.numeric(Boston$rad)
pairs(Boston)
```



These plots are not very useful as they are so small - we can perhaps make out some correlations between variables.

```
# Code for 10 c) goes here
cor(Boston, method = c("pearson"))
```

```
##                crim          zn       indus         chas         nox
## crim     1.00000000 -0.20046922  0.40658341 -0.055891582  0.42097171
## zn      -0.20046922  1.00000000 -0.53382819 -0.042696719 -0.51660371
## indus    0.40658341 -0.53382819  1.00000000  0.062938027  0.76365145
## chas    -0.05589158 -0.04269672  0.06293803  1.000000000  0.09120281
## nox      0.42097171 -0.51660371  0.76365145  0.091202807  1.00000000
## rm      -0.21924670  0.31199059 -0.39167585  0.091251225 -0.30218819
## age      0.35273425 -0.56953734  0.64477851  0.086517774  0.73147010
## dis     -0.37967009  0.66440822 -0.70802699 -0.099175780 -0.76923011
## rad      0.62550515 -0.31194783  0.59512927 -0.007368241  0.61144056
## tax      0.58276431 -0.31456332  0.72076018 -0.035586518  0.66802320
## ptratio  0.28994558 -0.39167855  0.38324756 -0.121515174  0.18893268
## black   -0.38506394  0.17552032 -0.35697654  0.048788485 -0.38005064
## lstat    0.45562148 -0.41299457  0.60379972 -0.053929298  0.59087892
## medv    -0.38830461  0.36044534 -0.48372516  0.175260177 -0.42732077
##                  rm         age         dis          rad         tax    ptratio
## crim    -0.21924670  0.35273425 -0.37967009  0.625505145  0.58276431  0.2899456
## zn       0.31199059 -0.56953734  0.66440822 -0.311947826 -0.31456332 -0.3916785
## indus   -0.39167585  0.64477851 -0.70802699  0.595129275  0.72076018  0.3832476
## chas     0.09125123  0.08651777 -0.09917578 -0.007368241 -0.03558652 -0.1215152
## nox     -0.30218819  0.73147010 -0.76923011  0.611440563  0.66802320  0.1889327
## rm       1.00000000 -0.24026493  0.20524621 -0.209846668 -0.29204783 -0.3555015
## age     -0.24026493  1.00000000 -0.74788054  0.456022452  0.50645559  0.2615150
## dis      0.20524621 -0.74788054  1.00000000 -0.494587930 -0.53443158 -0.2324705
## rad     -0.20984667  0.45602245 -0.49458793  1.000000000  0.91022819  0.4647412
## tax     -0.29204783  0.50645559 -0.53443158  0.910228189  1.00000000  0.4608530
## ptratio -0.35550149  0.26151501 -0.23247054  0.464741179  0.46085304  1.0000000
## black    0.12806864 -0.27353398  0.29151167 -0.444412816 -0.44180801 -0.1773833
## lstat   -0.61380827  0.60233853 -0.49699583  0.488676335  0.54399341  0.3740443
## medv     0.69535995 -0.37695457  0.24992873 -0.381626231 -0.46853593 -0.5077867
##               black      lstat       medv
## crim    -0.38506394  0.4556215 -0.3883046
## zn       0.17552032 -0.4129946  0.3604453
## indus   -0.35697654  0.6037997 -0.4837252
## chas     0.04878848 -0.0539293  0.1752602
## nox     -0.38005064  0.5908789 -0.4273208
## rm       0.12806864 -0.6138083  0.6953599
## age     -0.27353398  0.6023385 -0.3769546
## dis      0.29151167 -0.4969958  0.2499287
## rad     -0.44441282  0.4886763 -0.3816262
## tax     -0.44180801  0.5439934 -0.4685359
## ptratio -0.17738330  0.3740443 -0.5077867
## black    1.00000000 -0.3660869  0.3334608
## lstat   -0.36608690  1.0000000 -0.7376627
## medv     0.33346082 -0.7376627  1.0000000
```

There is an association between the per capita crime rate (crim) and the other predictors.

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

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
##  0.00632  0.08204  0.25651  3.61352  3.67708 88.97620
```

```
summary(Boston$tax)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   187.0   279.0   330.0   408.2   666.0   711.0
```

```
summary(Boston$ptratio)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   12.60   17.40   19.05   18.46   20.20   22.00
```

```
# Code for 10 e) goes here
nrow(subset(Boston, chas ==1))
```

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

35 suburbs fit this criteria

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

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   12.60   17.40   19.05   18.46   20.20   22.00 
```

Median pupil-teacher ratio = 19

```r
# Code for 10 g) goes here
df <- Boston[order(Boston$medv),]
lowest <- df[1,]
summary(df$crim)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
##  0.00632  0.08204  0.25651  3.61352  3.67708 88.97620 
```

```r
summary(lowest$crim)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   38.35   38.35   38.35   38.35   38.35   38.35 
```

Suburb 399 has the lowest median value of owner occupied homes. The crime rate is much greater than the mean/median in all neighbourhoods.

```r
# Code for 10 h) goes here
rm_over_7 <- subset(Boston, rm>7)
nrow(rm_over_7)
```

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

```r
## [1] 64
rm("rm_over_7")

rm_over_8 <- subset(Boston, rm>8)
nrow(rm_over_8)
```

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

```r
## [1] 13


summary(rm_over_8)
```

```
##       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
```

There are 64 suburbs with more than 7 rooms per dwelling.

There are 13 suburbs with more than 8 rooms per dwelling

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 approaches

Key features
* There is an observed response in y for each observation of the predictor measurement * Model relates the response to the predictors

Unsupervised approaches * For every observation we observe a vector of measurements but no associated y * The lack of a response variable means predictive/inferential modelling is not possible * Instead we can seek to understand relationships between the variables or between observations e.g cluster analysis * sometime we have incomplete data (i.e a proportion of responses are missing) this is referred to as a semi-supervised learning problem