HW1

Anish mohan

September 29, 2015

1. Q1

- 1a. Flexible model will generally perform better

```
 Inflexible methods can only fit to specific combination (e.g linear comb
ination) of the small number of predictors and cannot utlize the large nu
mber of samples to find a good fitting model. However, if the underlying
function is linear, then the inflexible method will do well.

  Flexible learning methods will be able to utlize the large number of sa
mples to find a reasonable fit for the true function with p predictors. H
owever, there is always the risk of over fitting the large number of inpu
t training points.
```

- 1b. Inflexible model will generally perform better

```
With large number of predictors and few number of sample points, the flex
ible model will combine the predictors but will be constrainted to the sm
all number of existing data points, thus overfitting the limited data dat
a.

An inflexible model will generally find a better fit to combine the predi
ctors to produce the results close to the few samples.
```

- 1c. Flexible model will generally perform better

```
Inflexible models cannot generalize for non linear functions. Flexible mo
dels will have better ability to fit to non-linear models, hence they wil
l generally perform better
```

- 1d. Inflexible model will generally perform better.

```
Given the high variance in noise, flexible models will tend to fit the er
ror data and give poor results. Inflexible models will do a better job of
ignoring the noise and finding a reasonable fit
```

2. Q2:

- 2a. Regression. Inference. n=500 p=3;

```
Inference problem because we are interested in finding out how the input
factors(profit, # of employees and industry) have an impact on the Salary
of CEO. The output:-CEO of salay is a quantifiable quantity hence, this i
s a regression problem
```

```
n= Sample size=500
p= number of predictors =3 (profit, # of employees, industry)
o/p: Salary of CEO
```

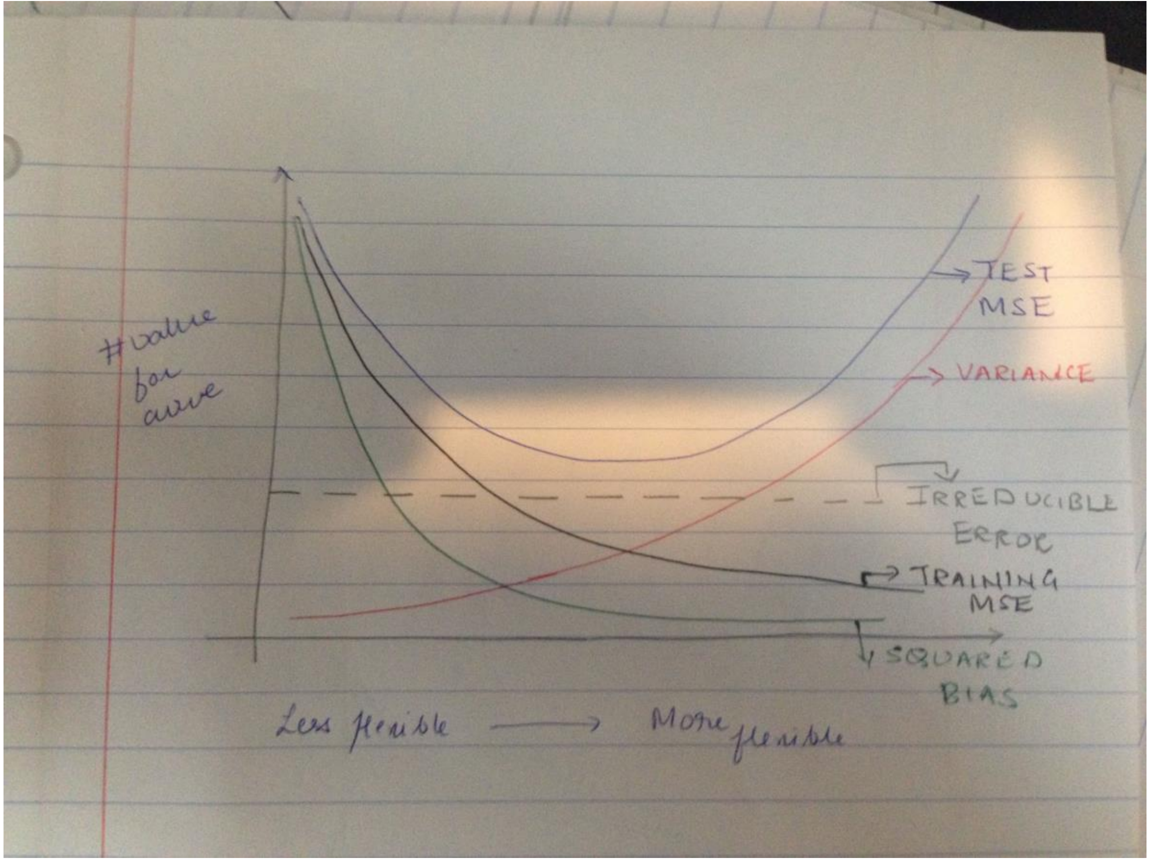
- 2b. Classification. Prediction n=20 p=13

```
Classification problem because the output variable is qualitative i.e suc
cess or
failure. Prediction problem because we just want to know if the product w
ill be success/failure and do not necessarily want to understand the inte
rplay between input variables and o/p

n=# of similar products=20
p= price charged,marketing budget, competiton price and 10 other variable
s.
o/p= if the product was a success or failure.
```

- 2c. Regression.Prediction n=52, p=3

```
 Regression problem because the output variable is a quantitative. Predic
tion problem because we are interested in predicting the % change in US d
ollar market but not necessarily interested how the input factors impact 
this output.

 n=# of weeks in a year =52
 p= % change in British, US and German Market=3
 o/p=% change in dollar
```

3. Q3



- Bias: Bias decreases as the flexibility of methods increases because in general more flexible functions can do a better approximation of the true function
- variance: Variance increases as the flexibility of method increases because results from function estimation from flexible will vary quite a bit when different sets of data are chosen.
- Training error: Training error decreases as flexibility of method increases as the model fits the training data better.
- Test error: Test error decreases upto a certain point as flexibility increases. However, after a particular point, the the test error starts increasing as the more flexible method is specialized for training data and does not generalize for the true function f.
- Irreducible error: It is assumed to remain constant and independent of the flexibility of the methods.

4. Q4

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

Obs	X1	X2	X3	Y	Distance to X1=X2=X3=0
1	0	3	0	Red	3
2	2	0	0	Red	2
3	0	1	3	Red	3.16
4	0	1	2	Green	2.23
5	-1	0	1	Green	1.414
6	1	1	1	Red	1.73

b. Green.
K=1, Then Y for(X1=0,X2=0,X3=0) is Green as Obs# is the closest.

c. What is our prediction with K = 3? Why?
Green
K=3. Closes points are Obs#4,5,6. Majority of the observations are Green, hence the Y= GREEN

d. Small. High values of K would give a boundary that tends is linear

5. Q5

- 5a. We cannot directly calculate the expected Test MSE $E(y_0-\hat{f}(x_0))$ as y_0 is not known. However we can try to estimate it.
- 5b. We cannot estimate the bias as we do not know the true function.
- 5c. We can estimate the variance of $\hat{f}$ by resampling the training data and finding the variance of the particular $
- 5d. The irreducible error at x_0 cannot be estimated as this error is random.

6. Q6

```
#Q6 Part a

in_training_target<-read.csv("training_target.csv")
print(paste0("Number of patients: ", nrow(in_training_target)))

## [1] "Number of patients: 2424"
```

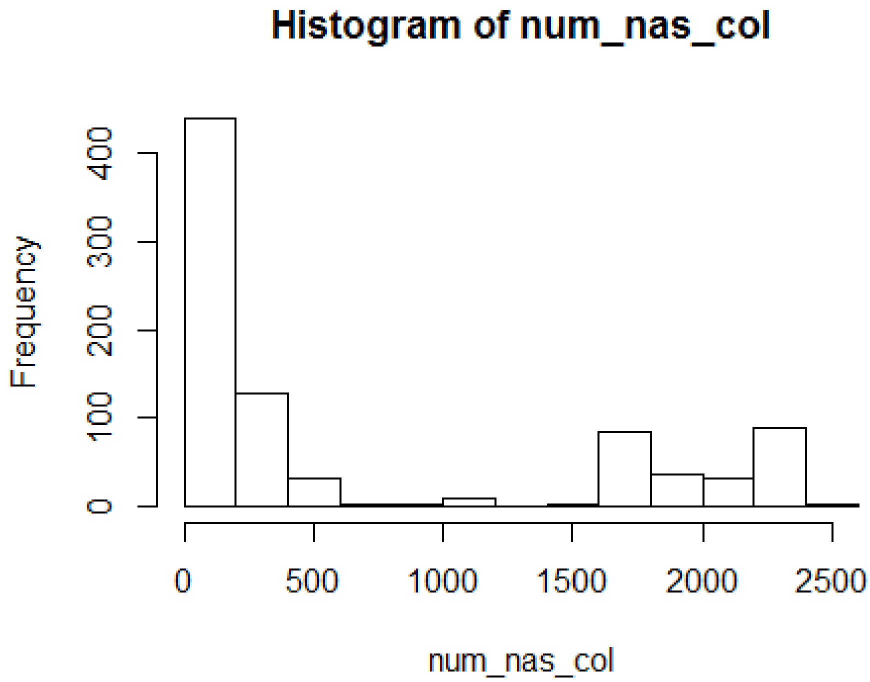
```
summary(in_training_target["ALSFRS_slope"])
```

```
##   ALSFRS_slope
##  Min.   :-4.3452
##  1st Qu.:-1.0863
##  Median :-0.6207
##  Mean   :-0.7308
##  3rd Qu.:-0.2742
##  Max.   : 1.2070
```

```
# Q6. Part b

in_training_features<-read.csv("training_features.csv")
print(paste0("Number of features: ", ncol(in_training_features)))
```

```
## [1] "Number of features: 858"
```

```
  numnas<-function(x){
    sum_na<-0
    for (n in x){
      if(is.na(n)==T)
        sum_na=sum_na+1
    }
    return(sum_na)
  }

  num_nas_col=apply(in_training_features,2,numnas)
  hist(num_nas_col)
```



```
#Q6 Part c

  feature.names<-names(in_training_features)

  for(feature.name in feature.names[-1]){
    dummy_name<-paste0("is.na.",feature.name)
    is.na.feature <-is.na(in_training_features[,feature.name])
    in_training_features[,dummy_name]<-as.integer(is.na.feature)
    in_training_features[is.na.feature,feature.name]<-median(in_training_feat
ures[,feature.name], na.rm=TRUE)

  }


#Q6 Part d
  valid_features<-read.csv("validation_features.csv")
  valid_target<-read.csv("validation_target.csv")
  print(paste0("Number of validation patients: ", nrow(valid_target)))

## [1] "Number of validation patients: 101"

  print(paste0("Number of validation features: ", ncol(valid_features)))

## [1] "Number of validation features: 858"

  summary(in_training_target["ALSFRS_slope"])

##   ALSFRS_slope
##  Min.   :-4.3452
##  1st Qu.:-1.0863
##  Median :-0.6207
##  Mean   :-0.7308
##  3rd Qu.:-0.2742
##  Max.   : 1.2070

  summary(valid_target["ALSFRS_slope"])

##   ALSFRS_slope
##  Min.   :-3.0417
##  1st Qu.:-1.2674
##  Median :-0.6565
##  Mean   :-0.7859
##  3rd Qu.:-0.3259
##  Max.   : 0.3694
```

- The minimum and maximum values in ALFRS_Slope between the validation and input patients are markedly different but the other statistical measures like mean, median, 25th & 75th quartile are comparable

```
summary(valid_features["weight.slope"])
```

```
##   weight.slope
##  Min.   :-10.4453
##  1st Qu.: -0.7609
##  Median :  0.0000
##  Mean   : -0.1319
##  3rd Qu.:  0.6997
##  Max.   :  2.7055
##  NA's   :24
```