

# 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 combination) of the small number of predictors and cannot utilize the large number 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 utilize the large number of samples to find a reasonable fit for the true function with  $p$  predictors. However, there is always the risk of over fitting the large number of input training points.

- 1b. Inflexible model will generally perform better

With large number of predictors and few number of sample points, the flexible model will combine the predictors but will be constrained to the small number of existing data points, thus overfitting the limited data data.

An inflexible model will generally find a better fit to combine the predictors 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 models will have better ability to fit to non-linear models, hence they will generally perform better

- 1d. Inflexible model will generally perform better.

Given the high variance in noise, flexible models will tend to fit the error 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 salary is a quantifiable quantity hence, this is 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 success or failure. Prediction problem because we just want to know if the product will be success/failure and do not necessarily want to understand the interplay between input variables and o/p

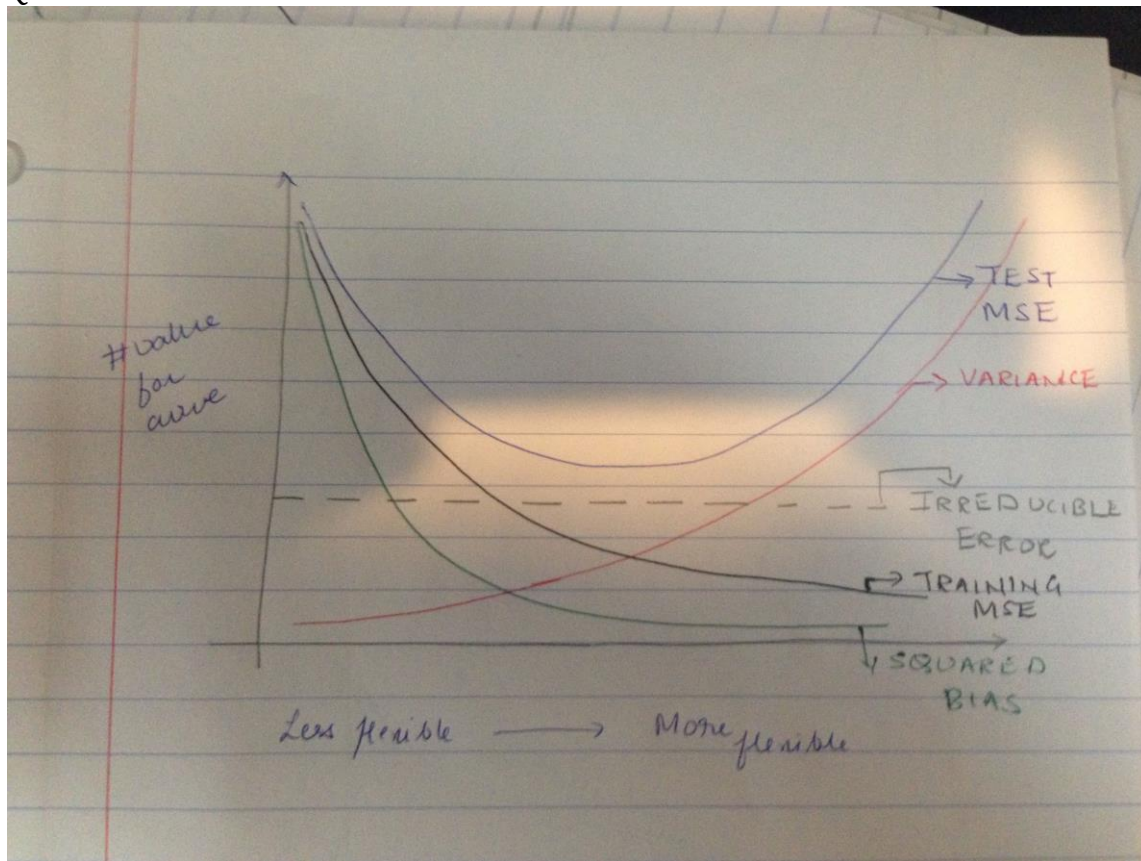
n=# of similar products=20  
p= price charged, marketing budget, competition price and 10 other variables.  
o/p= if the product was a success or failure.

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

Regression problem because the output variable is a quantitative. Prediction problem because we are interested in predicting the % change in US dollar 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

- 4a. Classification Examples
  - Is a newly discovered star in milky way or in an outside galaxy ?
    - Response: Inside or outside Milky way.

- Predictors: True Brightness, Motion vector.
  - Goal: Prediction, as we are trying to predict the position of the star.
- Is the a fish caught in the sea to be seabass or tuna?
  - Response: Is the fish Seabass or Tuna ?
  - Predictors: Length, width and weight.
  - Goal: Prediction, as are trying to predict the class of a newly caught fish.
- Is this new pixel in an image as background or foreground?
  - Response: Is the pixel in foreground or background
  - Predictors: Pixel brightness, Pixel brightness of 9 neighbors, Edge passing through pixel ?
  - Goal: Prediction as are trying to predict the class of the pixel based on input data.
- 4b. Regression Examples:
  - Value of the house and factors that impact it.
    - Response: How much would sale value increase if I add an additional bedroom.
    - Predictors: # of bedrooms, bathrooms, size, location etc.
    - Goal: Inference, as we trying to find if addition of addition of a bedroom has more significant impact than other factors.
  - What is the increas in sale volume due to additional investment in advertising on internet.
    - Response: How much would sale volume go up, if we increased advertising on internet.
    - Predictors: Advertising in various media channels: radio, newspaper, tv and internet.
    - Goal: Inference, as are trying to find if advertising on interet has more significant impact than other mediums.
  - Temperature of a particular place.
    - Response: How much do the critical factors that impact temperature of a particular place ?
    - Predictors: Lat/Long of the place, time of day, day of the year etc.
    - Goal: Inference, as we are trying to find out the impact of various predictors on the temperature.

## 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*

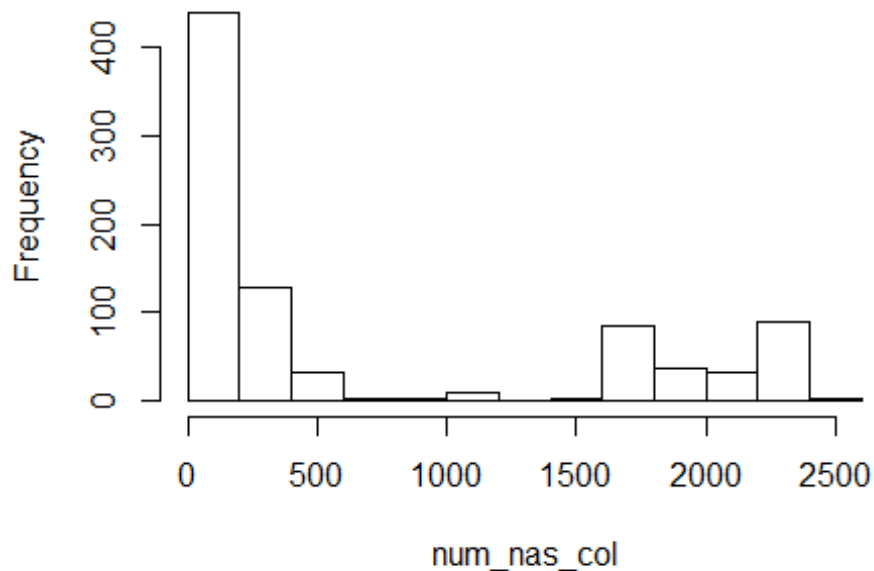
```
in_training_features<-read.csv("training_features.csv")
print(paste0("Number of features: ", ncol(in_training_features)))
```

```
## [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)
```

Histogram of 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_features[,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
```