Friday, February 6, 2015

DS502 Statistical Methods for Data Science

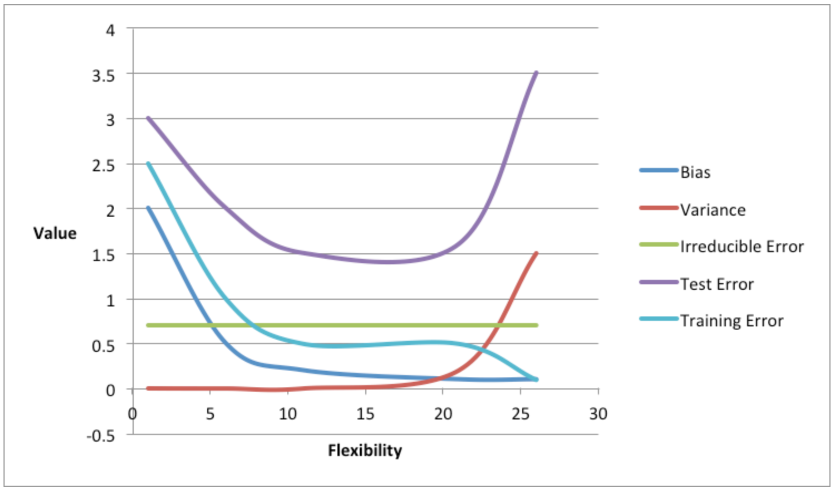
HW1 Congyuan Tang

1. Section 2.4, page52, Question 1

- (a) The sample size n is extremely large, and the number of predictors p is small. This suggests that our training data has too few degrees-of-freedom, and this is more likely because we used a more inflexible statistical learning method.
- (b) The number of predictors p is extremely large, and the sample size n is small. This suggests that our training data has too many degrees-of-freedom, and this is more likely because we used a more flexible statistical learning method.
- (c) The relationship between the predictors and the response is highly non-linear. This suggests our linear regression model has too many degrees-of-freedom, and this is more likely because we used a more flexible statistical learning method.
- (d) The variance of the error terms is extremely high. This suggests that our model might be overfitting the training data, and the model is often result from a more flexible statistical learning method.

2. Section 2.4, page52 - 53, Question 3

- (a) Graph:



- (b) Explanation:

- Test Error = Bias + Variance + Bayes Error;
- Training Error monotonously decrease as flexibility increases;
- Bayes Error is irreducible and constant due to our measuring method;
- Bias and Variance have trade-off effect that as the flexibility increases, Bias decreases and Variance increases.

3. Section 2.4, page53, Question 6

- Describe the difference 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 non-parametric approach)? What are its disadvantages?
  - parametric approach is a model-based approach whereas non-parametric approach do not make explicit assumptions about the target function.
  - Advantages: By assuming a particular function form of target function f, parametric approach reduces the problem of estimating f down to estimating a set of parameters.
  - Disadvantages: It need to make explicit assumption about the target function, and it only fit a narrow range of target function form.

4. Section 2.4, page54 - 55, Question 8

```
# Set working directory.
setwd('/Users/Eric/GoogleDrive/2015@WPI/DS502/HW1/')
# Read in college dataset.
college <- read.csv('College.csv')
# Set DataFrame's rowname to college names, and its not a data column.
rownames(college) = college[,1]
# Examine the dataset.
fix(college)
# Remove college names from original dataset.
college = college[,-1]
# Examine the dataset.
fix(college)
# Numerical summary of the variables in the dataset.
summary(college)
# Scatter plot of the first ten columns of the dataset.
pairs(college[,1:10])
```

```
# Boxplots of Outstate and Private
plot(as.factor(college$Private), college$Outstate)
# Binning the Top10perc variable.
Elite = rep("No", nrow(college))
Elite[college$Top10perc > 50] = "yes"
Elite = as.factor(Elite)
college = data.frame(college, Elite)
# Examine Elite
summary(Elite)
# Boxplots of Outstate verse Elite
plot(as.factor(college$Elite), college$Outstate)
# Plot histogram
par(mfrow=c(2,2))
hist(college$Apps)
hist(college$Accept)
hist(college$Enroll)
hist(college$Top10perc)

par(mfrow=c(1,1))
hist(college$Top10perc)
# Interesting findings:
# After ploting the histogram of college$Top10perc, we find the most common frequency is still around 10%.
# It suggests that most colleges are just as selective as high schools are. However, there are some colleges which is very selective with Top10perc higher than 80%.
# This also suggests that in general, colleges are not extremely selective.
par(mfrow=c(1,1))
plot(college$Apps, college$Accept)
highRatio = (college$Apps/college$Accept) > 5
college[highRatio,]
# Interesting findings 2:
# Most school recieves more applications will also accept more students, but there are also some colleges with high application number and low acceptance number.
# The most selective university (with application number five times more than the actual acceptance number) are Harvard and Princeton.
```

5. Section 2.4, page56, Question 9

```
# Set working directory.
setwd('/Users/Eric/GoogleDrive/2015@WPI/DS502/HW1/')
# Read in the dataset.
auto <- read.csv('Auto.csv')
# Summary of the dataset.
summary(auto)
```

```
# From the summary we can see, that 'mpg', 'cylinders', 'displacement', 'weight', 'acceleration', 'year', 'origin', are quantitative.
# 'horsepower', 'name' are qualitative.

# Finding range of each quantitative predictor.
range(auto$mpg)
range(auto$cylinders)
range(auto$displacement)
range(auto$weight)
range(auto$acceleration)
range(auto$year)
range(auto$origin)

# Mean and standard deviation of each quantitative predictor.
mean(auto$mpg)
sd(auto$mpg)

mean(auto$cylinders)
sd(auto$cylinders)

mean(auto$displacement)
sd(auto$displacement)

mean(auto$weight)
sd(auto$weight)

mean(auto$acceleration)
sd(auto$acceleration)

mean(auto$year)
sd(auto$year)

mean(auto$origin)
sd(auto$origin)

# Remove from 10th to 85th records and culculate Range, Mean, Standard Deviation again.
t1 = rep(TRUE,9)
t2 = rep(FALSE, 76)
t3 = rep(TRUE, nrow(auto)-85)
t = c(t1, t2, t3)
auto = auto[t,]
# Finding range of each quantitative predictor.
# Mean and standard deviation of each quantitative predictor.
# Read in Auto.csv dataset again to get full dataset.
```

```
auto <- read.csv('Auto.csv')
# Try to find interesting stuff.
plot(auto$cylinders, auto$mpg)
plot(as.factor(auto$cylinders), auto$mpg)
# It is obvious that even number of cylinders are more common in cars, and in average,
with more cylinders, the less mpg would be.

plot(auto$cylinders, auto$acceleration)
plot(as.factor(auto$cylinders), auto$acceleration)
# It is also obvious that in average, 4-cylinder cars have the same acceleration of 6-
cylinder cars, but have more acceleration than 8-cylinder cars (due to lack of records,
we ignore comparing to odd-cylinder cars).
# Predict mpg and justify it.
# These four predictors are helpful.
par(mfrow = c(2,2))
plot(auto$cylinders, auto$acceleration)
plot(auto$displacement, auto$mpg)
plot(as.numeric(auto$horsepower), auto$mpg)
plot(auto$weight, auto$mpg)
# These three predictors are not very helpful.
par(mfrow = c(2,2))
plot(auto$acceleration, auto$mpg)
plot(auto$year, auto$mpg)
plot(auto$origin, auto$mpg)
```

6. Section 3.7, page120, Question 1

- Null hypothesis for each of the three predictors in table3-4 is, the advertising in TV has no relation to the sales; the advertising in radio has no relation to the sales; the advertising in newspaper has no relation to the sales.
- The extremely small p-value (great less than 0.05) for TV and radio suggesting that, holding other predictors constant, each of the predictors will, with high likelihood, change the amount of sales. That is, fixing the money spent on radio and newspaper, and increasing the amount of TV advertising, will very likely lead to an increase in sales because the p-value in table 3.4 is very small for TV and TV’s coefficient is positive. On the other hand, since the p-value for newspaper is quite large (great larger than 0.05), spending more money in newspaper advertising may not have any effect on the sales when advertising in TV and radio are fixed.

7. Section 3.7, page121, Question 5

- Fitted values from linear regression without the intercept:

$$\hat{y}_i = x_i\beta \qquad (1)$$

$$\beta = \frac{\sum_{i'=1}^{n} x_{i'}y_{i'}}{\sum_{i'=1}^{n} x_{i'}^2} \qquad (2)$$

- We plug formula (2) to (1), we can get:

$$\hat{y}_i = x_i \frac{\sum_{i'=1}^{n} x_{i'}y_{i'}}{\sum_{i'=1}^{n} x_{i'}^2}$$

- From the definition given that:

$$\hat{y}_i = \sum_{i'=1}^{n} a_{i'}y_{i'}$$

- Then we can have:

$$a_{i'} = \frac{\sum_{i'=1}^{n} x_{i'}x_i}{\sum_{i'=1}^{n} x_{i'}^2}$$

8. Section 3.7, page121, Question 6

- Argue that in the case of simple linear regression, the least square line always passes through the point ($\underline{x}$, $\underline{y}$).
- The formula for simple linear regression:

$$\hat{y} = \beta_0 + \beta_1 x$$

- From formula 3.4 :

$$\beta_0 = \bar{y} - \beta_1\bar{x}$$

- By plugging formula 3.4 to simple linear regression:

$$\hat{y} = \bar{y} - \beta_1 \bar{x} + \beta_1 x$$

- From the equation above, we can see that, whatever beta1 equals to, x.bar and y.bar always satisfy this equation.

---

9. Section 3.7, page121 - 122, Question 8