Yifan\_Wu\_HW1

Yifan Wu

January 27, 2018

# 2.4 4. You will now think of some real-life applications for statistical learning.

## (a) Describe three real-life applications in which classification might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

Example 1: Classify high-rish patient out of low-risk patients to offer them emergent care due to limited hospital resources. Response: High-risk patient/Low-rish patient. Predictors: 17 variables (blood pressure, age) measured in the emergency room of a hospital. The goal is inference. We want discriminate high-rish patient from low-risk patient so we can better relocate hospital resources.

Example 2: Classify different applicants for new credit cards into the group that have good credit, bad credit and fall into grey area. Response: High-credit group/Low-credit group/Grey area group. Predictors: applicants information that contains related information such as anual salary, any outstanding debt, age etc. The goal is inference. We want to know who is eligible to be issued a new credit card.

Example 3: Astronomers needs to label distant objects in the sky into star, nebula, galaxy etc from the long exposure CCD images. Response: star, nebula, galaxy etc. Predictor: long exposure CCD images. The goals is inference. We need to classify different celestial bodies we discovers into known types.

## (b) Describe three real-life applications in which regression might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

Example 1: Predict sale for a perticular weather-sensitive item in a supermarket using previous time-based sales and weather data. Responses: predicted sale under certain weather condition. Predictors: Previous sale for that item under different weather conditions. The goal of this application is prediction. We want to gain knowledge about the sales in the future when a perticular weather condition happens.

Example 2: Discover the relationship between the monthly e-commerce sales and the online advertising costs. Responses: monthly e-commerce sales. Predictors: online advertising costs. The goal of this application inference. Its purpose is to find out the relationship between monthly e-commerce sales and online advertising costs.

Example 3: Examine the relationship between the age and price for used cars sold in the last year by a car dealership company. Responses: Car price. Predictors: Car age. The goal of this application is inference. We simply want to see if car price and car age are positvely correlated or negatively correlated.

## (c) Describe three real-life applications in which cluster analysis might be useful.

Example 1: Classifying groups of houses according to their house type, value and geographical location.

Example 2: Grouping customers with similar behavior in a large customer database containing their properties and past buying records.

Example 3: A retailer company want to group their customer into 3 segments by 4 dimentions(frequent spend, higher spend, higher ticket size and diversified spend) so they can target different groups using dfferent marketing strategies.

# 2.4 7.

## (a). Euclidean distance:

Observation 1:

Observation 2:

Observation 3:

Observation 4:

Observation 5:

Observation 6:

## (b). What is our prediction with K = 1? Why?

The test point (0,0,0) is closest to observatoin 5. Hence it should be classified as “Green”.

## (c). What is our prediction with K = 3? Why?

Find out 3 closest point to (0, 0, 0). They are (-1, 0, 1), (1, 1, 1), and (2, 0, 0). They are observation 2, 5, 6 and have label: Red, Green, and Red. The label “Red” appears of the time so if we choose K = 3, the test point (0, 0, 0) should be classified as “Red”.

## (d). If the Bayes decision boundary in this problem is highly nonlinear, then would we expect the best value for K to be large or small? Why?

If the Bayes decisoin boundary is highly nonlinear. we would expect the best value for K when using KNN to be very small. When Bayes decisoin boundary is highly nonlinear, the decision boundary is very flexible. It is a classifier that has low bias but very high variance, which correspond to a small K value when using KNN Classificaiton method.

# 9. This exercise involves the Auto data set studied in the lab. Make sure that the missing values have been removed from the data.

## (a) Which of the predictors are quantitative, and which are qualitative?

library(ISLR)  
dim(Auto)

## [1] 392 9

Quantitative predictors: mpg, cylinders, displacement, horsepower, weight, acceleration,

qualitative predictors: Year, Origin, name.

## (b) What is the range of each quantitative predictor? You can answer this using the range() function.

for(i in c('mpg', 'cylinders', 'displacement', 'horsepower', 'weight', 'acceleration')){  
 print(paste("The range of ", i, " in Auto is from", range(Auto[i])[1], "to", range(Auto[i])[2]))  
 cat("\n")  
 }

## [1] "The range of mpg in Auto is from 9 to 46.6"  
##   
## [1] "The range of cylinders in Auto is from 3 to 8"  
##   
## [1] "The range of displacement in Auto is from 68 to 455"  
##   
## [1] "The range of horsepower in Auto is from 46 to 230"  
##   
## [1] "The range of weight in Auto is from 1613 to 5140"  
##   
## [1] "The range of acceleration in Auto is from 8 to 24.8"

## (c) What is the mean and standard deviation of each quantitative predictor?

for(i in c('mpg', 'cylinders', 'displacement', 'horsepower', 'weight', 'acceleration')){  
 print(paste("The mean of ", i, " in Auto is", round(mean(Auto[[i]]),2)))  
 print(paste("The standard deviation of ", i, " in Auto is", round(sd(Auto[[i]]),2)))  
 cat("\n")  
 }

## [1] "The mean of mpg in Auto is 23.45"  
## [1] "The standard deviation of mpg in Auto is 7.81"  
##   
## [1] "The mean of cylinders in Auto is 5.47"  
## [1] "The standard deviation of cylinders in Auto is 1.71"  
##   
## [1] "The mean of displacement in Auto is 194.41"  
## [1] "The standard deviation of displacement in Auto is 104.64"  
##   
## [1] "The mean of horsepower in Auto is 104.47"  
## [1] "The standard deviation of horsepower in Auto is 38.49"  
##   
## [1] "The mean of weight in Auto is 2977.58"  
## [1] "The standard deviation of weight in Auto is 849.4"  
##   
## [1] "The mean of acceleration in Auto is 15.54"  
## [1] "The standard deviation of acceleration in Auto is 2.76"

## (d) Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

df<- Auto[-(10:84),]  
head(df)

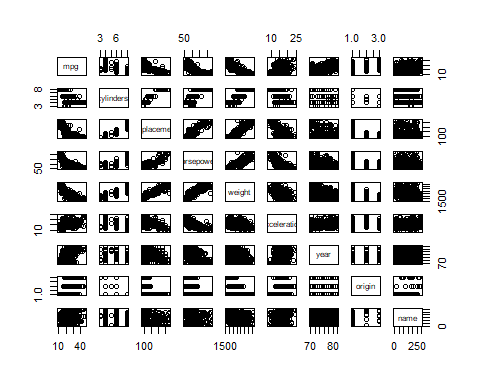
## mpg cylinders displacement horsepower weight acceleration year origin  
## 1 18 8 307 130 3504 12.0 70 1  
## 2 15 8 350 165 3693 11.5 70 1  
## 3 18 8 318 150 3436 11.0 70 1  
## 4 16 8 304 150 3433 12.0 70 1  
## 5 17 8 302 140 3449 10.5 70 1  
## 6 15 8 429 198 4341 10.0 70 1  
## name  
## 1 chevrolet chevelle malibu  
## 2 buick skylark 320  
## 3 plymouth satellite  
## 4 amc rebel sst  
## 5 ford torino  
## 6 ford galaxie 500

for(i in c('mpg', 'cylinders', 'displacement', 'horsepower', 'weight', 'acceleration')){  
 print(paste("The range of ", i, " in the subset of Auto that remains is from",  
 range(df[i])[1], "to", range(df[i])[2]))  
 print(paste("The mean of ", i, " in the subset of Auto that remains is", round(mean(df[[i]]),2)))  
 print(paste("The standard deviation of ", i, " in the subset of Auto that remains is",  
 round(sd(df[[i]]),2)))  
 cat("\n")  
}

## [1] "The range of mpg in the subset of Auto that remains is from 11 to 46.6"  
## [1] "The mean of mpg in the subset of Auto that remains is 24.37"  
## [1] "The standard deviation of mpg in the subset of Auto that remains is 7.88"  
##   
## [1] "The range of cylinders in the subset of Auto that remains is from 3 to 8"  
## [1] "The mean of cylinders in the subset of Auto that remains is 5.38"  
## [1] "The standard deviation of cylinders in the subset of Auto that remains is 1.66"  
##   
## [1] "The range of displacement in the subset of Auto that remains is from 68 to 455"  
## [1] "The mean of displacement in the subset of Auto that remains is 187.75"  
## [1] "The standard deviation of displacement in the subset of Auto that remains is 99.94"  
##   
## [1] "The range of horsepower in the subset of Auto that remains is from 46 to 230"  
## [1] "The mean of horsepower in the subset of Auto that remains is 100.96"  
## [1] "The standard deviation of horsepower in the subset of Auto that remains is 35.9"  
##   
## [1] "The range of weight in the subset of Auto that remains is from 1649 to 4997"  
## [1] "The mean of weight in the subset of Auto that remains is 2939.64"  
## [1] "The standard deviation of weight in the subset of Auto that remains is 812.65"  
##   
## [1] "The range of acceleration in the subset of Auto that remains is from 8.5 to 24.8"  
## [1] "The mean of acceleration in the subset of Auto that remains is 15.72"  
## [1] "The standard deviation of acceleration in the subset of Auto that remains is 2.69"

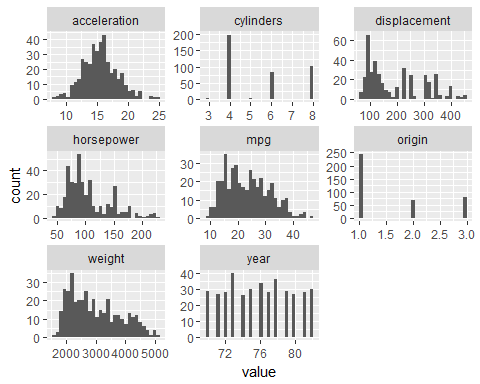
## (e) Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.

library(purrr)  
library(tidyr)  
library(ggplot2)  
pairs(Auto)



Auto %>%  
 keep(is.numeric) %>%   
 gather() %>%   
 ggplot(aes(value)) +  
 facet\_wrap(~ key, scales = "free") +  
 geom\_histogram()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



library(ggplot2)  
library(reshape)

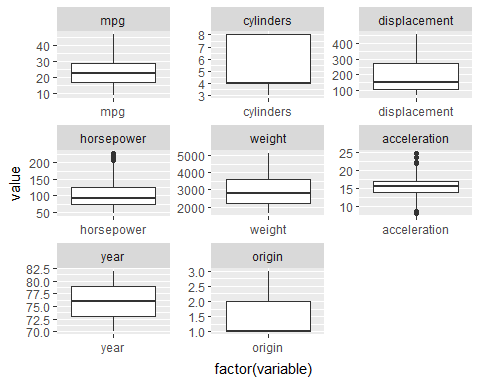
##   
## Attaching package: 'reshape'

## The following objects are masked from 'package:tidyr':  
##   
## expand, smiths

meltData <- melt(Auto)

## Using name as id variables

p <- ggplot(meltData, aes(factor(variable), value))   
p + geom\_boxplot() + facet\_wrap(~variable, scale="free")



From the scatterplot matrix we can see that “horsepower”, “weight”, and “displacement” are all positively correlated in a linear pattern.

There is a weak negative linear relationship between this 3 categories and “accleration”.

The category “mpg” are negatively correlated with the above three categories.

In “horsepower” and “acceleration”, there exists outliers.

“Acceleration” is roughly normally distributed around 16.

## (f) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer.

The category “mpg” are negatively correlated with “horsepower”, “weight”, and “displacement”s. We can expect “mpg” to be small if “horsepower”, “weight”, and “displacement” appears to be large. There an obvious negative linear relationship between “mpg” and the other 3 variables so we could use any of the three variables to perform linear regression analysis to predict “mpg”.

# The last problem

## (a). Make 10 different simulations with model complexity = 1. Compute the average Residual SSE. Also find the approximate range of the highest order coefficient for these 10 simulations. This is a measure for the baseline variance for a low complexity model.

SSE pproxiate range: 70 to 110.

## b) Make 10 different simulations with model complexity = 3. Compute the average Residual SSE. Which coefficient has the largest range in this case? What is that range? This is a measure for the variance for a medium complexity model.

SSE pproxiate range: 15 to 80.

## c) Repeat this for model complexity = 15. Which coefficient has the largest range for these 10 simulations? What is that range?

SSE pproxiate range: 0 to 10.

When the model complexity is 1, the range of the SSE is the largest. The approximate range is from 70 to 110.

## d) How do your results illustrate the bias - variance trade-off? The answer should be a short paragraph.

Variance refers to the degree of non-linearity of the fitted model. The more zigzaged the boundary line is, the higher the degree of variance this model has. Bias refers to the deviation of our fitted model to the real data. Here the bias could be measured as in SSE. We can see from the result of our simulations is that the more shapely the fited line is, the smaller the SSE is, and vice versa. So the higer the degree of the polynomial that is being fitted(variance), the more shapely the fitted line is, the smaller the SEE(bias) will be. So bias and variance are negatively correlated with each other. The trade-off between variance and bias will always exist.

## e) For which model complexity do you typically obtain a curve which is most similar and overall close to the unknown curve that is to be estimated? Try multiple simulation for several different model complexities, summarize what you see, and explain your answer. Pictures or numerical results are not required.

When we fit a model with 2 degree of complexity, the fitted line resembles the line from which we generated the data. Since the black line has one curve, it is a part of a two degree polynomial. So it makes sense that when we set the model complexity to 2, the fitted line is the closest to the unknown curve overall.