hw1math4323

anthonycastillo UH ID: 1670011

2022-09-03

Question 1a:

classification -> prediction n=20 p=14

Question 1b:

regression->inference n=200 p=15

Question 1c:

regression->inference n = 1500 p=10 Question 1d:

classification->prediction n = 300 p =8

Question 2a:

supervised learning

Question 2b:

1. for example (a) and (d) f\_hat can be treated as a black box. The other examples (b) and (c) they can not be treated as a black box.

ii.for inference examples we need to know the exact form of f\_hat. The predication example we can use the treat as black box because the form of f\_hat is not very important as long the prediction is accurate.

Question 2c:

reducible and irreducible errors. The reducible error can be improved using a better statistical learning technique. The irreducible error can not be improved because we can not avoid randoms errors in the model.

Question 3:

example 1: streaming services trying to group similar tv show/movies that the consumer has watched

example 2: selling a product that targets a specific audience like video games

example 3: in sports you can group together many differnet statistics to help the team win/lose the game

Question 4a:

Use RStudio’s drop-down menu (Environment → Import Dataset → From Text (base) …) to read the data into R. Make sure the Heading is set to Y es. Call the loaded data Credit.

Question 4b-1:

summary(Credit)

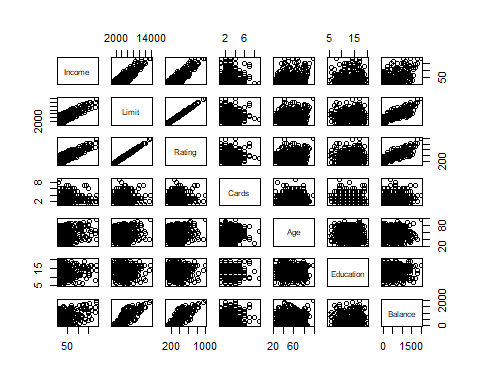
## Income Limit Rating Cards   
## Min. : 10.35 Min. : 855 Min. : 93.0 Min. :1.000   
## 1st Qu.: 21.01 1st Qu.: 3088 1st Qu.:247.2 1st Qu.:2.000   
## Median : 33.12 Median : 4622 Median :344.0 Median :3.000   
## Mean : 45.22 Mean : 4736 Mean :354.9 Mean :2.958   
## 3rd Qu.: 57.47 3rd Qu.: 5873 3rd Qu.:437.2 3rd Qu.:4.000   
## Max. :186.63 Max. :13913 Max. :982.0 Max. :9.000   
## Age Education Own Student Married Region   
## Min. :23.00 Min. : 5.00 No :193 No :360 No :155 East : 99   
## 1st Qu.:41.75 1st Qu.:11.00 Yes:207 Yes: 40 Yes:245 South:199   
## Median :56.00 Median :14.00 West :102   
## Mean :55.67 Mean :13.45   
## 3rd Qu.:70.00 3rd Qu.:16.00   
## Max. :98.00 Max. :20.00   
## Balance   
## Min. : 0.00   
## 1st Qu.: 68.75   
## Median : 459.50   
## Mean : 520.01   
## 3rd Qu.: 863.00   
## Max. :1999.00

Question 4b-2: numerical values - Income, Limit, Rating, Cards, Age, Education, Balance

categorical values - Own, Student, Married, Region

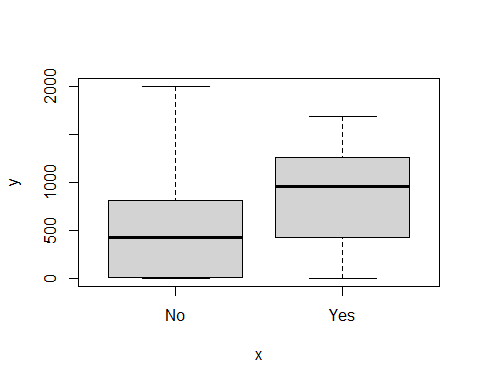
Question 4b-3:

pairs(~Income+Limit+Rating+Cards+Age+Education+Balance, Credit)



Question 4b-4:

plot(Credit$Student,Credit$Balance)



Question 4b-5:

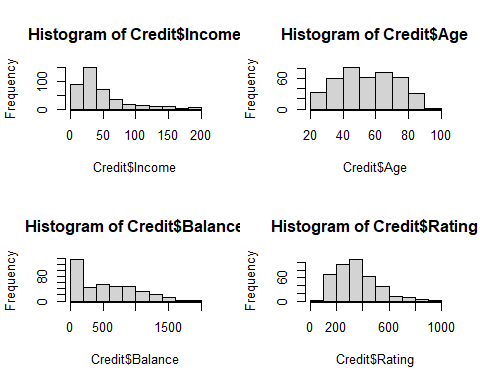
high=(Credit$Rating>680)  
table(high)

## high  
## FALSE TRUE   
## 382 18

#plot(Credit$Balance,Credit$high)

Question 4b-6:

par(mfrow = c(2,2))  
hist(Credit$Income)  
hist(Credit$Age)  
hist(Credit$Balance)  
hist(Credit$Rating)



Question 5a:

library(MASS)  
#Boston  
?Boston

## starting httpd help server ... done

506 rows and 14 columns. lstat - lower status of the population ptratio - pupil-teacher ratio by town chas - if bounded by charles river or not medv - median value of owner-occupied homes in $1000s

Question 5b:

most of the predictors are quantitative variables

Question 5c:

Boston\_range = sapply(Boston,range)  
rownames(Boston\_range) = c("min:","max:")  
Boston\_range

## crim zn indus chas nox rm age dis rad tax ptratio black  
## min: 0.00632 0 0.46 0 0.385 3.561 2.9 1.1296 1 187 12.6 0.32  
## max: 88.97620 100 27.74 1 0.871 8.780 100.0 12.1265 24 711 22.0 396.90  
## lstat medv  
## min: 1.73 5  
## max: 37.97 50

Question 5d:

Boston\_mean\_sd = sapply(Boston, function(x) c(mean(x),sd(x)))  
rownames(Boston\_mean\_sd) = c("mean:","sd:")  
Boston\_mean\_sd

## crim zn indus chas nox rm age  
## mean: 3.613524 11.36364 11.136779 0.06916996 0.5546951 6.2846344 68.57490  
## sd: 8.601545 23.32245 6.860353 0.25399404 0.1158777 0.7026171 28.14886  
## dis rad tax ptratio black lstat medv  
## mean: 3.795043 9.549407 408.2372 18.455534 356.67403 12.653063 22.532806  
## sd: 2.105710 8.707259 168.5371 2.164946 91.29486 7.141062 9.197104

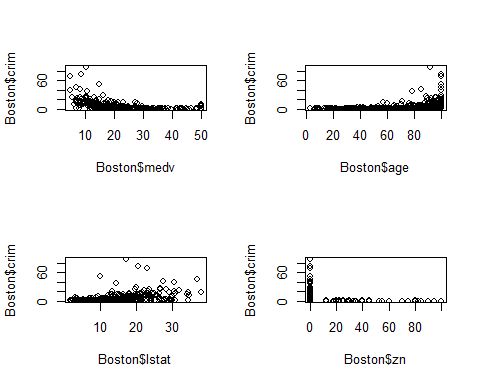
Question 5e:

newBoston = Boston[-c(50:100),]  
NB\_range\_mean\_sd = sapply(newBoston, function(x) round(c(range(x),mean(x),sd(x)),digits=1))  
#format(round(NB\_range\_mean\_sd,digits = 4),nsmall = 4)  
rownames(NB\_range\_mean\_sd) = c("min:","max:","mean:", "sd:")  
NB\_range\_mean\_sd

## crim zn indus chas nox rm age dis rad tax ptratio black lstat  
## min: 0 0.0 0.5 0.0 0.4 3.6 2.9 1.1 1.0 187.0 12.6 0.3 1.7  
## max: 89 95.0 27.7 1.0 0.9 8.8 100.0 12.1 24.0 711.0 22.0 396.9 38.0  
## mean: 4 10.5 11.7 0.1 0.6 6.3 71.1 3.6 10.2 420.8 18.5 352.7 13.1  
## sd: 9 22.8 6.9 0.3 0.1 0.7 27.7 2.1 9.0 172.3 2.3 95.4 7.3  
## medv  
## min: 5.0  
## max: 50.0  
## mean: 22.4  
## sd: 9.5

Question 5f:

#pairs(Boston)  
par(mfrow=c(2,2))  
plot(Boston$medv,Boston$crim)  
plot(Boston$age,Boston$crim)  
plot(Boston$lstat,Boston$crim)  
plot(Boston$zn,Boston$crim)

 crim and medv - as medv increases the crim rate decreases.

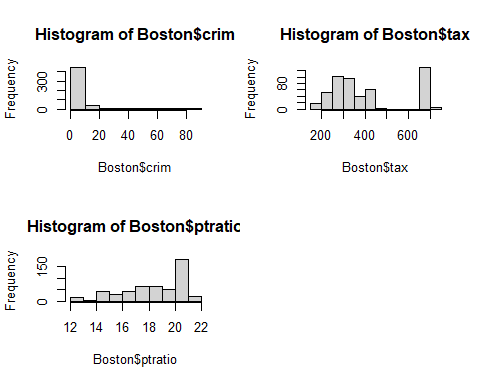
crim and lstat - as lstat increases there is somewhat of an increase in crime rates.

crim and age - as age increase the crim rate also increases.

crim and zn - 0 zn have higher crim rate than the rest of the plot.

Question 5g:

par(mfrow=c(2,2))  
hist(Boston$crim)  
hist(Boston$tax)  
hist(Boston$ptratio)



Question 5h:

nrow(subset(Boston,chas ==1))

## [1] 35

Question 5i:

#pairs(Boston)  
par(mfrow=c(2,2))  
plot(Boston$crim, Boston$medv)  
plot(Boston$lstat, Boston$medv)  
plot(Boston$black, Boston$medv)  
plot(Boston$rm, Boston$medv)

