ISYE 6501 HW10

April 2021

1 Question 14.1

Read the data into R and review the summary of each feature, I found that the 7th column, which is Bare Nuclei, contain character values. Since the data should be numbers from 1-10, the character values indicate missing value in that column. Then I checked the unique values for each column to see if there are any value that is outside of the range. It seems Bare Nuclei is the only one that has missing values. Then I counted that there are 16 missing values in Bare Nuclei, which is less than 5% of the total number of data points 699. Thus we can perform imputation for the missing values.

```
 V7
 Length:699
 Class :character
 Mode  :character

Unique values:
 [2]  5  3  6  4  8  1  2  7 10  9
 [3]  1  4  8 10  2  3  7  5  6  9
 [4]  1  4  8 10  2  3  5  6  7  9
 [5]  1  5  3  8 10  4  6  2  9  7
 [6]  2  7  3  1  6  4  5  8 10  9
 [7]  1 10  2  4  3  9  7 NA  5  8  6
 [8]  3  9  1  2  4  5  7  8  6 10
 [9]  1  2  7  4  5  3 10  6  9  8
 [10] 1  5  4  2  3  7 10  8  6
 [10] 2  4
```

2 Question 14.1.1 Mean/Mode

First I will fill in the missing values with the mode of Bare Nuclei due to that all the values are discrete number from 1-10. The mode for Bare Nuclei is value 1.

2.1 My R code is:

```
set.seed(416)

df <- read.table("breast-cancer-wisconsin.data.txt", sep = ',',
                 stringsAsFactors = FALSE, header = FALSE)
head(df)
summary(df)
missing_value <- which(df$V7 == "?")
df[df == "?"] <- NA
# review missing value
for (i in c(2,3,4,5,6,7,8,9,10,11)){
  print(unique(df[,i]))
}

sum(is.na(df$V7))

# fill in with mode
library(modeest)
Mode = mlv(df[,7], method = "mfv")

df[,7] <- as.integer(replace(df[,7], is.na(df[,7]), Mode))
```

3 Question 14.1.2 Regression

Using a regression to fill in missing value, I will use independent variables except Bare Nuclei to predict value of missing Bare Nuclei. I used linear regression to fit the model and use stepwise regression to select the best model for prediction. I also rounded the predicted values to nearest integer so that it is categorical.

```
Result for predicted missing value:
5 8 1 2 1 2 3 2 2 6 1 3 5 2 1 1
```

3.1 My R code is:

```
# fill in with regression
set.seed(416)
df2 <- read.table("breast-cancer-wisconsin.data.txt", sep = ',',
                  stringsAsFactors = FALSE, header = FALSE)
head(df2)
summary(df2)
df2[df2 == "?"] <- NA
sum(is.na(df2$V7))
fit <- lm(V7~V2+V3+V4+V5+V6+V8+V9+V10,df2)
summary(fit)
AIC_fit <- stepAIC(fit, direction = "both", trace = FALSE)

df2[,7] <- as.integer(replace(df2[,7], is.na(df2[,7]),
                             as.integer(predict(AIC_fit,df2[is.na(df2$V7),])+0.5)))
df2[missing_value,]
```

4 Question 14.1.3 Regression With Perturbation

I added perturbation terms that are random numbers from 0.7 standard deviation from the mean (0) of 16 numbers, since there are 16 missing values in Bare Nuclei. The reason to choose 0.7 sd instead of 1 sd is that the range for Bare Nuclei is from 1-10, choosing 1 sd will cause some of the predicted value to be 0. This method does give a different set of values from the regular regression imputation.

```
Result for predicted missing value:
5 8 3 1 1 3 2 2 2 6 1 2 5 2 1 1
```

4.1 My R code is:

```
set.seed(416)
df3 <- read.table("breast-cancer-wisconsin.data.txt", sep = ',',
                   stringsAsFactors = FALSE, header = FALSE)
```

```
missing_value3 <- which(df3$V7 == "?")
df3[df3 == "?"] <- NA
fit3 <- lm(V7~V2+V3+V4+V5+V6+V8+V9+V10,df3)
AIC_fit3 <- stepAIC(fit3, direction = "both", trace = FALSE)
perturb <- rnorm(nrow(df3[missing_value3,]),0,0.7)
df3[,7] <- as.integer(replace(df3[,7], is.na(df3[,7]),
                              predict(AIC_fit3,df3[is.na(df3$V7),])+perturb+0.5))
df3[missing_value3,]
```

5 Question 15.1

An example for optimization is associate scheduling for a Gym. Given the Gym operating hours, different positions (coach, front desk, cleaner, life guard, etc.), and restrictions on duty periods, allocate associates most effectively to operate the gym.