Homework 10

Anonymous

2024-03-21

Question 14.1

The breast cancer data set 'breast-cancer-wisconsin.data.txt' has missing values.

1. *Use the mean/mode imputation method to impute values for the missing data.*
2. *Use regression to impute the values for the missing data.*
3. *Use regression with perturbation to impute values for the missing data.*
4. *(Optional) Compare the results and quality of the classification models (e.g., SVM, KNN) build using*
    1. *the data sets from questions 1-3*
    2. *the data that remains after data points with missing values are removed*
    3. *the data set when a binary variable is introduced to indicate missing values*

Answer 14.1.1 Mean/Mode Imputation Method

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
cancer_data <- read.table(file= "C:\\Users\\sheya\\OneDrive\\Desktop\\breast-cancer-wisconsin.data.txt"
                          header = TRUE,
                          sep = ",",
                          stringsAsFactors = FALSE,
                          na.strings = "?")

head(cancer_data, 4)
```

```
##   X1000025 X5 X1 X1.1 X1.2 X2 X1.3 X3 X1.4 X1.5 X2.1
## 1  1002945  5  4    4    5  7   10  3    2    1    2
## 2  1015425  3  1    1    1  2    2  3    1    1    2
## 3  1016277  6  8    8    1  3    4  3    7    1    2
## 4  1017023  4  1    1    3  2    1  3    1    1    2
```

```r
#Give column names
colnames(cancer_data) <- c("ID", "Clump_Thickness", "Cell_Size",
                           "Cell_Shape", "Marginal_Adhesion",
                           "Single_Epith_Cell_Size", "Bare_Nuclei",
                           "Bland_Chromatin", "Normal_Nucleoli",
                           "Mitoses", "Class")
cancer_data$Class <- as.factor(cancer_data$Class)
levels(cancer_data$Class) <- c(0,1)
#Summary
summary(cancer_data)
```

```
##        ID           Clump_Thickness    Cell_Size        Cell_Shape
##  Min.   :   61634   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
##  1st Qu.:  870258   1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 1.000
##  Median : 1171710   Median : 4.000   Median : 1.000   Median : 1.000
##  Mean   : 1071807   Mean   : 4.417   Mean   : 3.138   Mean   : 3.211
##  3rd Qu.: 1238354   3rd Qu.: 6.000   3rd Qu.: 5.000   3rd Qu.: 5.000
##  Max.   :13454352   Max.   :10.000   Max.   :10.000   Max.   :10.000
##
##  Marginal_Adhesion Single_Epith_Cell_Size  Bare_Nuclei     Bland_Chromatin
##  Min.   : 1.000    Min.   : 1.000         Min.   : 1.000   Min.   : 1.000
##  1st Qu.: 1.000    1st Qu.: 2.000         1st Qu.: 1.000   1st Qu.: 2.000
##  Median : 1.000    Median : 2.000         Median : 1.000   Median : 3.000
##  Mean   : 2.809    Mean   : 3.218         Mean   : 3.548   Mean   : 3.438
##  3rd Qu.: 4.000    3rd Qu.: 4.000         3rd Qu.: 6.000   3rd Qu.: 5.000
##  Max.   :10.000    Max.   :10.000         Max.   :10.000   Max.   :10.000
##                                           NA's   :16
##  Normal_Nucleoli    Mitoses      Class
##  Min.   : 1.00   Min.   : 1.00   0:457
##  1st Qu.: 1.00   1st Qu.: 1.00   1:241
##  Median : 1.00   Median : 1.00
##  Mean   : 2.87   Mean   : 1.59
##  3rd Qu.: 4.00   3rd Qu.: 1.00
##  Max.   :10.00   Max.   :10.00
##
```

```r
#Find missing data
cancer_data[is.na(cancer_data$Bare_Nuclei),]
```

```
##          ID Clump_Thickness Cell_Size Cell_Shape Marginal_Adhesion
## 23  1057013               8         4          5                 1
## 40  1096800               6         6          6                 9
## 139 1183246               1         1          1                 1
## 145 1184840               1         1          3                 1
## 158 1193683               1         1          2                 1
## 164 1197510               5         1          1                 1
## 235 1241232               3         1          4                 1
## 249  169356               3         1          1                 1
## 275  432809               3         1          3                 1
## 292  563649               8         8          8                 1
## 294  606140               1         1          1                 1
## 297   61634               5         4          3                 1
## 315  704168               4         6          5                 6
```

```
## 321  733639               3         1          1                1
## 411 1238464               1         1          1                1
## 617 1057067               1         1          1                1
##     Single_Epith_Cell_Size Bare_Nuclei Bland_Chromatin Normal_Nucleoli Mitoses
## 23                       2          NA               7               3       1
## 40                       6          NA               7               8       1
## 139                      1          NA               2               1       1
## 145                      2          NA               2               1       1
## 158                      3          NA               1               1       1
## 164                      2          NA               3               1       1
## 235                      2          NA               3               1       1
## 249                      2          NA               3               1       1
## 275                      2          NA               2               1       1
## 292                      2          NA               6              10       1
## 294                      2          NA               2               1       1
## 297                      2          NA               2               3       1
## 315                      7          NA               4               9       1
## 321                      2          NA               3               1       1
## 411                      1          NA               2               1       1
## 617                      1          NA               1               1       1
##     Class
## 23      1
## 40      0
## 139     0
## 145     0
## 158     0
## 164     0
## 235     0
## 249     0
## 275     0
## 292     1
## 294     0
## 297     0
## 315     0
## 321     0
## 411     0
## 617     0
```

```r
#Check for the percentage of missing data with the Rule of Thumb >5% of all data
print(16/nrow(cancer_data)*100)
```

```
## [1] 2.292264
```

```r
#The missing data is nearly 2.3% of the total data which is lower than 5%

#After identifying the missing data, We can move onto the mean/mode imputation method

#14.1.1 Using mean/mode imputation method
#Run mean without NA values
Mean <- round(mean(as.integer(cancer_data$Bare_Nuclei), na.rm = TRUE))
Mean
```

```
## [1] 4
```

```
#Replace the missing data with the Mean
New_Mean <- cancer_data
New_Mean[is.na(New_Mean)] <- Mean
New_Mean[c(23,40,139,145),]
```

```
##          ID Clump_Thickness Cell_Size Cell_Shape Marginal_Adhesion
## 23  1057013               8         4          5                 1
## 40  1096800               6         6          6                 9
## 139 1183246               1         1          1                 1
## 145 1184840               1         1          3                 1
##     Single_Epith_Cell_Size Bare_Nuclei Bland_Chromatin Normal_Nucleoli Mitoses
## 23                       2           4               7               3       1
## 40                       6           4               7               8       1
## 139                      1           4               2               1       1
## 145                      2           4               2               1       1
##     Class
## 23      1
## 40      0
## 139     0
## 145     0
```

```
#Using mode imputation

calc_mode <- function(x){
  u <- unique(x)
  tab <- tabulate(match(x,u))
  u[which.max(tab)]
}
d <- cancer_data$Bare_Nuclei
calc_mode(d)
```

```
## [1] 1
```

```
#Below is inputting the mode into the Bare Nuclei column
#I did not show the entire column for simplicity visuals.
#df <- data.frame(cancer_data$Bare_Nuclei)
#df %>%
#  mutate(cancer_data.Bare_Nuclei = if_else(is.na(cancer_data.Bare_Nuclei),
#                                           calc_mode(d)),
#                                           cancer_data.Bare_Nuclei))
```

Conclusion:

In my findings above, I was able to find which column in the Breast Cancer data that contained missing data which was named Bare Nuclei. Furthermore, I checked the percentage of missing data to see if the missing data was within the 5% of all data to continue (as rule of thumb).

Next, I was able to use the Mean imputation method to calculate the mean and also replace the NA's with the mean for four rows (for simplicity visuals).

I then used the Mode imputation method and replaced the NA's for the Bare Nuclei column as well. However, I did not run the code considering it showed the entire row and for simplicity, I decided to hash-tag it out.

Question 14.1.2 Regression Imputation Method

```
library(tidyverse)
library(caret)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
cancer_data <- read.table(file= "C:\\Users\\sheya\\OneDrive\\Desktop\\breast-cancer-wisconsin.data.txt"
                          header = TRUE, sep = ",",
                          stringsAsFactors = FALSE,
                          na.strings = "?")

#Give column names
colnames(cancer_data) <- c("ID", "Clump_Thickness", "Cell_Size",
                           "Cell_Shape", "Marginal_Adhesion",
                           "Single_Epith_Cell_Size", "Bare_Nuclei",
                           "Bland_Chromatin", "Normal_Nucleoli",
                           "Mitoses", "Class")
#All other predictors data points except for the mising values & response variable
New_cancer <- cancer_data
missing.index <- which(is.na(New_cancer$Bare_Nuclei), arr.ind = TRUE)
New_cancer.2 <- New_cancer[-missing.index, 2:10]

#Split data into 70% training and 30% testing
set.seed(123)
random <- sample(2, nrow(New_cancer.2), replace = TRUE, prob = c(0.7,0.3))
c.train <- New_cancer.2[random == 1,]
c.test <- New_cancer.2[random == 2,]

#Regression
R.Model <- lm(formula = Bare_Nuclei~
                Clump_Thickness+
                Cell_Size+
                Cell_Shape+
                Marginal_Adhesion+
                Single_Epith_Cell_Size+
                Bland_Chromatin+
                Normal_Nucleoli+
                Mitoses, data = c.train, na.action = na.exclude)
summary(R.Model)
```

```
##
## Call:
## lm(formula = Bare_Nuclei ~ Clump_Thickness + Cell_Size + Cell_Shape +
```

```
##     Marginal_Adhesion + Single_Epith_Cell_Size + Bland_Chromatin +
##     Normal_Nucleoli + Mitoses, data = c.train, na.action = na.exclude)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.2573  -0.9004  -0.3415   0.7270   8.5754
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -0.64779    0.23810  -2.721  0.00676 **
## Clump_Thickness         0.20144    0.04956   4.064 5.66e-05 ***
## Cell_Size               0.08076    0.09086   0.889  0.37455
## Cell_Shape              0.23926    0.09024   2.651  0.00829 **
## Marginal_Adhesion       0.35135    0.05575   6.302 6.83e-10 ***
## Single_Epith_Cell_Size  0.08926    0.07598   1.175  0.24072
## Bland_Chromatin         0.36777    0.07217   5.096 5.06e-07 ***
## Normal_Nucleoli        -0.06775    0.05712  -1.186  0.23621
## Mitoses                -0.01445    0.07197  -0.201  0.84093
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.284 on 465 degrees of freedom
## Multiple R-squared:  0.6154, Adjusted R-squared:  0.6088
## F-statistic: 93.02 on 8 and 465 DF,  p-value: < 2.2e-16
```

```
#From the summary, the significant values will generate a new model
R2.Model <- lm(formula = Bare_Nuclei~
                 Clump_Thickness+
                 Cell_Shape+
                 Marginal_Adhesion+
                 Bland_Chromatin, data = c.train)
summary(R2.Model)
```

```
##
## Call:
## lm(formula = Bare_Nuclei ~ Clump_Thickness + Cell_Shape + Marginal_Adhesion +
##     Bland_Chromatin, data = c.train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.0016  -0.9659  -0.3441   0.7335   8.5955
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -0.55001    0.21113  -2.605  0.00948 **
## Clump_Thickness    0.20726    0.04877   4.250 2.58e-05 ***
## Cell_Shape         0.30294    0.05950   5.091 5.16e-07 ***
## Marginal_Adhesion  0.35427    0.05478   6.467 2.52e-10 ***
## Bland_Chromatin    0.36336    0.06817   5.330 1.53e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.283 on 469 degrees of freedom
## Multiple R-squared:  0.6126, Adjusted R-squared:  0.6093
```

```
## F-statistic: 185.4 on 4 and 469 DF,  p-value: < 2.2e-16
```

```
#Cross Validation Prediction
pred.Cmodel <-R2.Model
C.trainControl <- trainControl(method = "cv", number = 10)
pred.Cmodel.2 <- train(Bare_Nuclei~
                         Clump_Thickness+
                         Cell_Shape+
                         Marginal_Adhesion+
                         Bland_Chromatin, c.train, method = 'lm',
                       trControl = C.trainControl)
summary(pred.Cmodel.2)
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.0016  -0.9659  -0.3441   0.7335   8.5955
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -0.55001    0.21113  -2.605  0.00948 **
## Clump_Thickness    0.20726    0.04877   4.250 2.58e-05 ***
## Cell_Shape         0.30294    0.05950   5.091 5.16e-07 ***
## Marginal_Adhesion  0.35427    0.05478   6.467 2.52e-10 ***
## Bland_Chromatin    0.36336    0.06817   5.330 1.53e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.283 on 469 degrees of freedom
## Multiple R-squared:  0.6126, Adjusted R-squared:  0.6093
## F-statistic: 185.4 on 4 and 469 DF,  p-value: < 2.2e-16
```

```
#The results of the cross validation has the same value as the second traing model.
#I will do the prediction of the second model and run R^2
pred.train <- predict(pred.Cmodel.2, c.train)
SSE.train <- sum((pred.train - c.train[,7])^2)
SST.train <- sum((c.train[,7] - mean(c.train[,7]))^2)
R2.train <- 1 - SSE.train / SST.train
R2.train
```

```
## [1] 0.6423789
```

```
#The R2 of the training model is 64%, the performance is good.

#Testing data for regression
pred.test <- predict(pred.Cmodel.2, c.test)
SSE.test <- sum((pred.test - c.test[,7])^2)
SST.test <- sum((c.test[,7] - mean(c.test[,7]))^2)
R2.test <- 1 - SSE.test / SST.test
R2.test
```

```
## [1] 0.7231125
```

```
#The R2 of the testing model is 72%, this is slightly better than the training model.

#Predicting the missing values
regression <- cancer_data
pred.missing <- predict(pred.Cmodel.2, New_cancer[missing.index,])
regression[missing.index,]$Bare_Nuclei <- as.integer(pred.missing)
regression[c(23,40,139,145),]
```

```
##          ID Clump_Thickness Cell_Size Cell_Shape Marginal_Adhesion
## 23  1057013               8         4          5                 1
## 40  1096800               6         6          6                 9
## 139 1183246               1         1          1                 1
## 145 1184840               1         1          3                 1
##     Single_Epith_Cell_Size Bare_Nuclei Bland_Chromatin Normal_Nucleoli Mitoses
## 23                       2           5               7               3       1
## 40                       6           8               7               8       1
## 139                      1           1               2               1       1
## 145                      2           1               2               1       1
##     Class
## 23      4
## 40      2
## 139     2
## 145     2
```

Conclusion:

In my findings using Regression, the first model showed 'Clump_Thickness', 'Cell Shape', 'Marginal Adhesion', and 'Bland Chromatin' were significant therefore, I used them to create a new model. From that model, the R^2 was at 61% which led me to cross validation. After cross validation, the summary revealed an R^2 fo 61% as well. Therefore, I used that model for prediction. The result came to 64% on the training set and 72% on the testing set. The testing set was slightly better than the training set. Finally, I predicted the missing values which are shown above.

Question 14.1.3 Using regression with perturbation

```
cancer_data <- read.table(file= "C:\\Users\\sheya\\OneDrive\\Desktop\\breast-cancer-wisconsin.data.txt"
                          header = TRUE, sep = ",",
                          stringsAsFactors = FALSE,
                          na.strings = "?")

#Give column names
colnames(cancer_data) <- c("ID", "Clump_Thickness", "Cell_Size",
                           "Cell_Shape", "Marginal_Adhesion",
                           "Single_Epith_Cell_Size", "Bare_Nuclei",
                           "Bland_Chromatin", "Normal_Nucleoli",
                           "Mitoses", "Class")
#Regression with pertubation to imput values for the missing ones.
set.seed(123)
n <- rnorm(16, mean = pred.missing, sd = sd(pred.missing))
n
```

```
##  [1]  4.2619729  7.7262477  4.5413660  1.8053878  1.2710788  6.0848735
##  [7]  3.7628953 -1.0217252  0.5191977  5.0652500  3.7899388  3.2840782
## [13]  6.2727378  2.0676038 -0.2070023  4.6904530
```

```
#Combining the predicted data together.
P.reg <- cancer_data
P.reg[missing.index,]$Bare_Nuclei <- as.integer(abs(n))
P.reg[c(23,40,139,145),]
```

```
##          ID Clump_Thickness Cell_Size Cell_Shape Marginal_Adhesion
## 23  1057013               8         4          5                 1
## 40  1096800               6         6          6                 9
## 139 1183246               1         1          1                 1
## 145 1184840               1         1          3                 1
##     Single_Epith_Cell_Size Bare_Nuclei Bland_Chromatin Normal_Nucleoli Mitoses
## 23                       2           4               7               3       1
## 40                       6           7               7               8       1
## 139                      1           4               2               1       1
## 145                      2           1               2               1       1
##     Class
## 23      4
## 40      2
## 139     2
## 145     2
```

Conclusion:

From the accuracy calculated with and without the imputed data, it appears to be in the same range which doesn't give enough evidence that imputation was helpful.

Question 15.1

Describe a situation or problem from your job, everyday life, current events, etc., for which optimization would be appropriate. What data would you need?

A situation in which optimization would be appropriate is in a business analytics perspective such as banking: fraud detection. An algorithm can be used for detection and flagging of potential bank fraud. Depending on the banks stored data considering its high volume, it is extremely difficult for a person to manually detect any suspicious activity within a single account. Therefore, for this example, consider a persons account usually only spends $3,000 in a month with their credit card but, this month, there is a $30,000 charge on the credit card. The algorithm analyzes the pattern and alerts the bank. The course of action, or optimization approach, can recommend a course of action. The algorithm can range from cancelling the card to sending a text message to the account holders phone to authorize the transaction and call the bank to allow for a higher spending amount.