hw10.R

Muhammad

2021-11-03

setwd("C:/Users/Muhammad/ISYE/10")  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(tidyr)  
data <-read.table("breast-cancer-wisconsin.data.txt", stringsAsFactor = FALSE, header = F, sep = ",", na.strings="?")  
  
#Naming columns   
colnames(data) <- c("ID", "Clump\_Thickness", "Uniform\_Cell\_Size", "Uniform\_Cell\_Shape",  
 "Marg\_Adhesion", "Single\_Epith\_Cell\_Size", "Bare\_Nuclei", "Bland\_Chromatin",  
 "Normal\_Nucleoli", "Mitoses", "Class")  
  
#for checking number of unique classes  
#unique(data$Class)  
#converting Class variable to categorical data or factors  
data$Class <- as.factor(data$Class)  
summary(data)

## ID Clump\_Thickness Uniform\_Cell\_Size Uniform\_Cell\_Shape  
## Min. : 61634 Min. : 1.000 Min. : 1.000 Min. : 1.000   
## 1st Qu.: 870688 1st Qu.: 2.000 1st Qu.: 1.000 1st Qu.: 1.000   
## Median : 1171710 Median : 4.000 Median : 1.000 Median : 1.000   
## Mean : 1071704 Mean : 4.418 Mean : 3.134 Mean : 3.207   
## 3rd Qu.: 1238298 3rd Qu.: 6.000 3rd Qu.: 5.000 3rd Qu.: 5.000   
## Max. :13454352 Max. :10.000 Max. :10.000 Max. :10.000   
##   
## Marg\_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 : 2.807 Mean : 3.216 Mean : 3.545 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. : 1.000 Min. : 1.000 2:458   
## 1st Qu.: 1.000 1st Qu.: 1.000 4:241   
## Median : 1.000 Median : 1.000   
## Mean : 2.867 Mean : 1.589   
## 3rd Qu.: 4.000 3rd Qu.: 1.000   
## Max. :10.000 Max. :10.000   
##

#missing values in Bare\_Nuclei variable  
  
#creating a temporary variable for mean imputation  
  
data1 <- data  
  
#Q14.1.1  
#Mean imputation  
data1$Bare\_Nuclei[is.na(data$Bare\_Nuclei)] <- mean(data1$Bare\_Nuclei, na.rm = TRUE)  
mean(data1$Bare\_Nuclei)

## [1] 3.544656

#mide Imputation:  
  
#Creating a commonly used mode function (used a generic function recommended)  
data2<-data  
  
Mode <- function(x) {   
 u\_var <- unique(x)  
 mode <- u\_var[which.max(tabulate(match(x, u\_var)))]  
 mode  
}  
  
data2$Bare\_Nuclei[is.na(data2$Bare\_Nuclei)] <- Mode(data2$Bare\_Nuclei[!is.na(data2$Bare\_Nuclei)])   
summary(data2$Bare\_Nuclei)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 1.000 1.000 3.486 5.000 10.000

#Q14.1.2  
#Regression imputation   
data3 <- data  
  
#removing all data with na  
data3.1 <-drop\_na(data3)  
  
#Linear Model for predicting Bare nuclie value for Na value  
model <- lm(Bare\_Nuclei~Clump\_Thickness+Uniform\_Cell\_Size+Uniform\_Cell\_Shape+Marg\_Adhesion+Single\_Epith\_Cell\_Size+Bland\_Chromatin+Normal\_Nucleoli+Mitoses  
 ,data=data3.1 )  
  
  
predicted\_Na\_value <- predict(model,newdata=data3[which(is.na(data3$Bare\_Nuclei),),])   
  
  
data\_3.1a <- data3  
#replacing Na values with predicted values in regression  
data\_3.1a[which(is.na(data3$Bare\_Nuclei)),]$Bare\_Nuclei <- as.integer(predicted\_Na\_value)  
summary(data\_3.1a)

## ID Clump\_Thickness Uniform\_Cell\_Size Uniform\_Cell\_Shape  
## Min. : 61634 Min. : 1.000 Min. : 1.000 Min. : 1.000   
## 1st Qu.: 870688 1st Qu.: 2.000 1st Qu.: 1.000 1st Qu.: 1.000   
## Median : 1171710 Median : 4.000 Median : 1.000 Median : 1.000   
## Mean : 1071704 Mean : 4.418 Mean : 3.134 Mean : 3.207   
## 3rd Qu.: 1238298 3rd Qu.: 6.000 3rd Qu.: 5.000 3rd Qu.: 5.000   
## Max. :13454352 Max. :10.000 Max. :10.000 Max. :10.000   
## Marg\_Adhesion Single\_Epith\_Cell\_Size Bare\_Nuclei Bland\_Chromatin   
## Min. : 1.000 Min. : 1.000 Min. : 0.000 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 : 2.807 Mean : 3.216 Mean : 3.514 Mean : 3.438   
## 3rd Qu.: 4.000 3rd Qu.: 4.000 3rd Qu.: 5.500 3rd Qu.: 5.000   
## Max. :10.000 Max. :10.000 Max. :10.000 Max. :10.000   
## Normal\_Nucleoli Mitoses Class   
## Min. : 1.000 Min. : 1.000 2:458   
## 1st Qu.: 1.000 1st Qu.: 1.000 4:241   
## Median : 1.000 Median : 1.000   
## Mean : 2.867 Mean : 1.589   
## 3rd Qu.: 4.000 3rd Qu.: 1.000   
## Max. :10.000 Max. :10.000

#Q14.1.3  
#Imputation with Reg and Perturbation  
#total number of Na values  
t\_val= sum(is.na(data$Bare\_Nuclei))  
t\_val

## [1] 16

#Generating random number using predicted values from regression  
P <- as.integer(abs(rnorm(t\_val, mean = predicted\_Na\_value, sd = sd(predicted\_Na\_value))))  
  
data4 <- data  
data4[which(is.na(data3$Bare\_Nuclei)),]$Bare\_Nuclei<- P  
summary(data4)

## ID Clump\_Thickness Uniform\_Cell\_Size Uniform\_Cell\_Shape  
## Min. : 61634 Min. : 1.000 Min. : 1.000 Min. : 1.000   
## 1st Qu.: 870688 1st Qu.: 2.000 1st Qu.: 1.000 1st Qu.: 1.000   
## Median : 1171710 Median : 4.000 Median : 1.000 Median : 1.000   
## Mean : 1071704 Mean : 4.418 Mean : 3.134 Mean : 3.207   
## 3rd Qu.: 1238298 3rd Qu.: 6.000 3rd Qu.: 5.000 3rd Qu.: 5.000   
## Max. :13454352 Max. :10.000 Max. :10.000 Max. :10.000   
## Marg\_Adhesion Single\_Epith\_Cell\_Size Bare\_Nuclei Bland\_Chromatin   
## Min. : 1.000 Min. : 1.000 Min. : 0.000 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 : 2.807 Mean : 3.216 Mean : 3.519 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. :11.000 Max. :10.000   
## Normal\_Nucleoli Mitoses Class   
## Min. : 1.000 Min. : 1.000 2:458   
## 1st Qu.: 1.000 1st Qu.: 1.000 4:241   
## Median : 1.000 Median : 1.000   
## Mean : 2.867 Mean : 1.589   
## 3rd Qu.: 4.000 3rd Qu.: 1.000   
## Max. :10.000 Max. :10.000