Homework-10

Question 14.1

The breast cancer data set breast-cancer-wisconsin.data.txt from http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/ (description at http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29) has missing values.

1. **Use the mean/mode imputation method to impute values for the missing data.**

```
rm(list = ls())

set.seed(42069)

data = read.table("C:/Users/Admin/Desktop/MM/Homework 10/breast-cancer-wisconsin.data.txt",
                  stringsAsFactors = FALSE,
                  header = FALSE,
                  sep = ",")
head(data)

# find columns with missing rows
for (i in 1:11) {
  print(paste0("V",i))
  print(which(data[,i] == "?"))
}
plot(data$V7) # only v7 has missing values, categorical data

getmode = function(v) {
  uniqv = unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}
mode = getmode(data$V7)

data_mode = data
index = (data$V7 == "?")
data_mode$V7[index] = mode
```

2. **Use regression to impute values for the missing data.**

```
library(DAAG)

# regression imputation
data_clean = data[-index,2:10]
data_clean$V7 = as.integer(data_clean$V7) # convert to integer for cv
```

```r
# linear regression
model = lm(V7~., data_clean)
step(model) # step-wise variable selection

# model with significant variables
model = lm(V7~V2+V4+V5+V8, data_clean)
summary(model)

# cross validation
model_cv = cv.lm(data_clean,
                 model,
                 m = 5)
SST = sum((as.numeric(data[-index,]$V7) - mean(as.numeric(data[-index,]$V7)))^2)
r =  1 - attr(model_cv,"ms")*nrow(data[-index,])/SST
r

# prediction for missing value
pred = round(predict(model, data[index,]))
data_reg = data
data_reg[index,]$V7 = pred
```

3. **Use regression with perturbation to impute values for the missing data.**

```r
pert = round(rnorm(nrow(data[index,]),
             mean = mean(pred),
             sd = sd(pred)))
pert
data_reg_pert = data
data_reg_pert[index,]$V7 = pert
data_reg_pert$V7[data_reg_pert$V7 < 1] = 1
data_reg_pert$V7[data_reg_pert$V7 > 10] = 10
```

4. **(Optional) Compare the results and quality of classification models (e.g., SVM, KNN) build using**

    1. **the data sets from questions 1,2,3;**

```r
library(kknn)

train.idx = sample(nrow(data), size = floor(nrow(data) * 0.7))
test.idx = setdiff(1:nrow(data), train.idx)

acc = matrix(ncol=5, nrow=5)
kknn = function(dataset, number) {
  for (k in 1:5) {
    knn_model = kknn(V11~.,
                     data[train.idx,],
                     data[test.idx,],
                     k=k)
    pred = as.integer(round(fitted(knn_model)))
    acc[k,i] = sum(pred==data[test.idx,]$V11)/nrow(data[test.idx,])*100
  }
}
```

```
kknn(data_mode, 1)
kknn(data_reg, 2)
kknn(data_reg_pert, 3)
acc
```

2. **the data that remains after data points with missing values are removed; and**

```
data_clean = data[-index,]
train.idx = sample(nrow(data_clean), size = floor(nrow(data_clean) * 0.7))
test.idx = setdiff(1:nrow(data_clean), train.idx)

kknn(data_clean, 4)
acc
```

3. **the data set when a binary variable is introduced to indicate missing values.**

```
# cluster on data for replaced binary missing category
data_binary = data
data_binary$V7[data_binary$V7 == "?"] = 0 # interaction term
data_binary$V12[data_binary$V7 == 0] = 0 # binary term
data_binary$V12[data_binary$V7 != 0] = 1

train.idx = sample(nrow(data_binary), size = floor(nrow(data_binary) * 0.7))
test.idx = setdiff(1:nrow(data_binary), train.idx)

kknn(data_binary, 5)
acc

# there isn't a significant difference in model performance between how missing
# data is being handled. Clusters might be more important.
```

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?

Price optimization is one of the most commonly known use case especially in retail industries.

- The objective function will be to maximize revenue(or maybe reduce loss/wastage) by having
- some constraints on each product based on competitors' prices, expiry for perishable items, likelihood of purchase must be above certain "predicted" levels using modeling techniques, and even having positive net profits.
- Variables that are important may include supply from manufacturers or in inventory, demand based on seasonality or trends, location, or even product brand.