exer2

Christopher Huong

2023-06-18

Due 11:59 PM CT 06/18/2023

The UC Irvine Machine Learning Repository contains a data set related to glass identification. The data consist of 214 glass samples labeled as one of seven class categories. There are nine predictors, including the refractive index and percentages of eight elements: Na, Mg, Al, Si, K, Ca, Ba, and Fe.

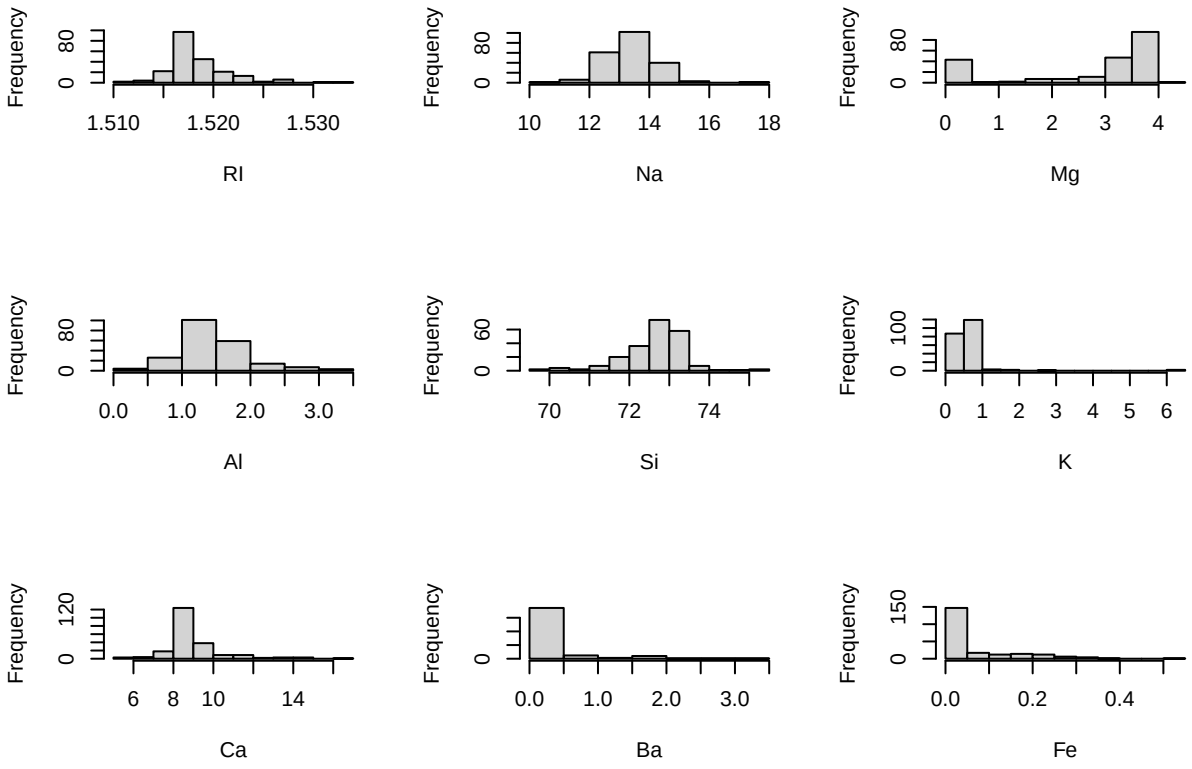
```
library(mlbench)
library(tidyverse)
library(psych)
library(caret)
data(Glass)
str(Glass)
```

```
## 'data.frame':    214 obs. of  10 variables:
##  $ RI  : num  1.52 1.52 1.52 1.52 1.52 ...
##  $ Na  : num  13.6 13.9 13.5 13.2 13.3 ...
##  $ Mg  : num  4.49 3.6 3.55 3.69 3.62 3.61 3.6 3.61 3.58 3.6 ...
##  $ Al  : num  1.1 1.36 1.54 1.29 1.24 1.62 1.14 1.05 1.37 1.36 ...
##  $ Si  : num  71.8 72.7 73 72.6 73.1 ...
##  $ K   : num  0.06 0.48 0.39 0.57 0.55 0.64 0.58 0.57 0.56 0.57 ...
##  $ Ca  : num  8.75 7.83 7.78 8.22 8.07 8.07 8.17 8.24 8.3 8.4 ...
##  $ Ba  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Fe  : num  0 0 0 0 0 0.26 0 0 0 0.11 ...
##  $ Type: Factor w/ 6 levels "1","2","3","5",..: 1 1 1 1 1 1 1 1 1 1 ...
```

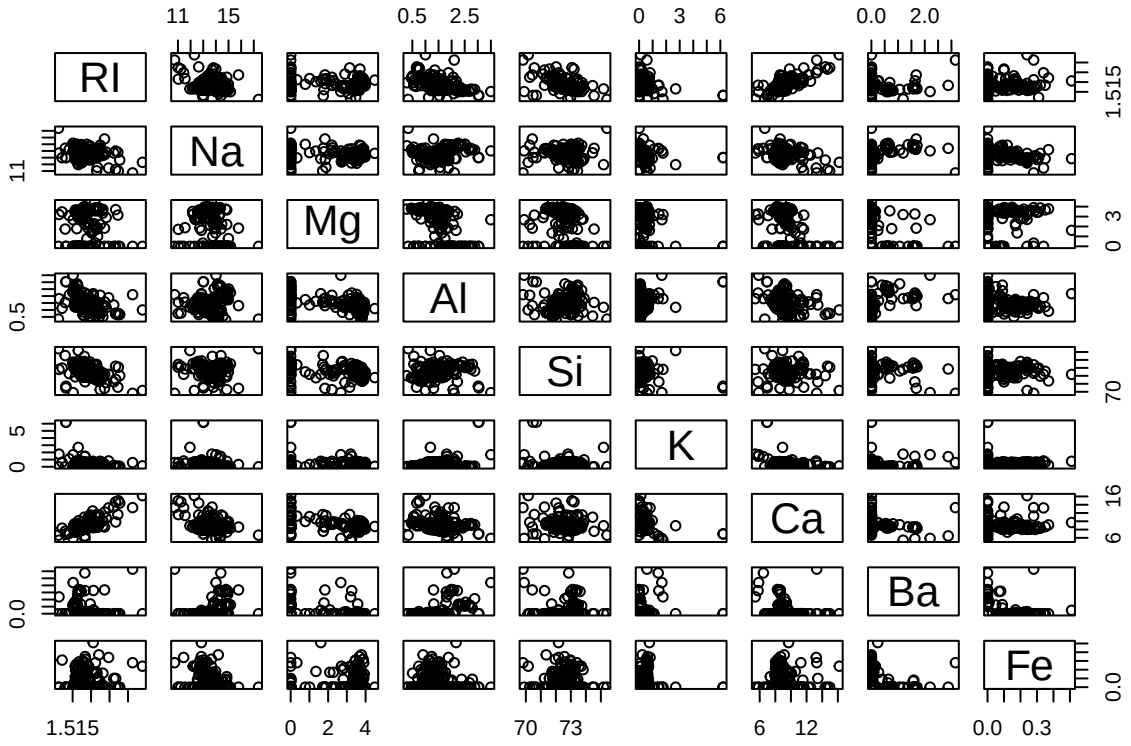
a) Using visualizations, explore the predictor variables to understand their distributions as well as the relationships between predictors.

```
vars_list <- as.list(colnames(select(Glass,-Type)))

par(mfrow=c(3,3))
for(i in vars_list){hist(select(Glass,-Type)[,i],xlab=i,main="")}
```




```
plot(select(Glass, -Type))
```



b) Do there appear to be any outliers in the data? Are any predictors skewed? Show all work!

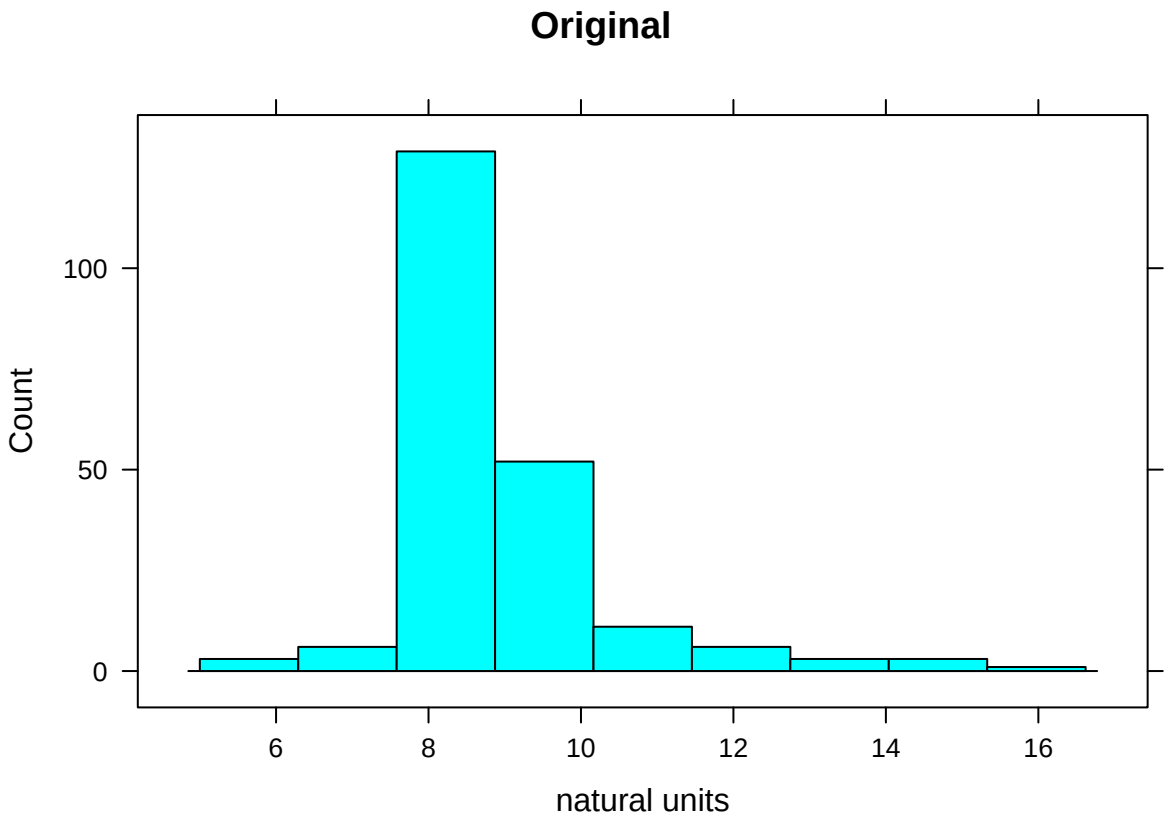
```
describe(Glass)
```

```
##       vars   n  mean   sd median trimmed  mad   min   max range  skew kurtosis
## RI       1 214  1.52 0.00   1.52    1.52 0.00  1.51  1.53  0.02  1.60     4.72
## Na       2 214 13.41 0.82  13.30   13.38 0.64 10.73 17.38  6.65  0.45     2.90
## Mg       3 214  2.68 1.44   3.48    2.87 0.30  0.00  4.49  4.49 -1.14    -0.45
## Al       4 214  1.44 0.50   1.36    1.41 0.31  0.29  3.50  3.21  0.89     1.94
## Si       5 214 72.65 0.77  72.79   72.71 0.57 69.81 75.41  5.60 -0.72     2.82
## K        6 214  0.50 0.65   0.56    0.43 0.17  0.00  6.21  6.21  6.46    52.87
## Ca       7 214  8.96 1.42   8.60    8.74 0.66  5.43 16.19 10.76  2.02     6.41
## Ba       8 214  0.18 0.50   0.00    0.03 0.00  0.00  3.15  3.15  3.37    12.08
## Fe       9 214  0.06 0.10   0.00    0.04 0.00  0.00  0.51  0.51  1.73     2.52
## Type*   10 214  2.54 1.71   2.00    2.31 1.48  1.00  6.00  5.00  1.04    -0.29
##           se
## RI      0.00
## Na      0.06
## Mg      0.10
## Al      0.03
## Si      0.05
## K       0.04
## Ca      0.10
## Ba      0.03
## Fe      0.01
## Type*   0.12
```

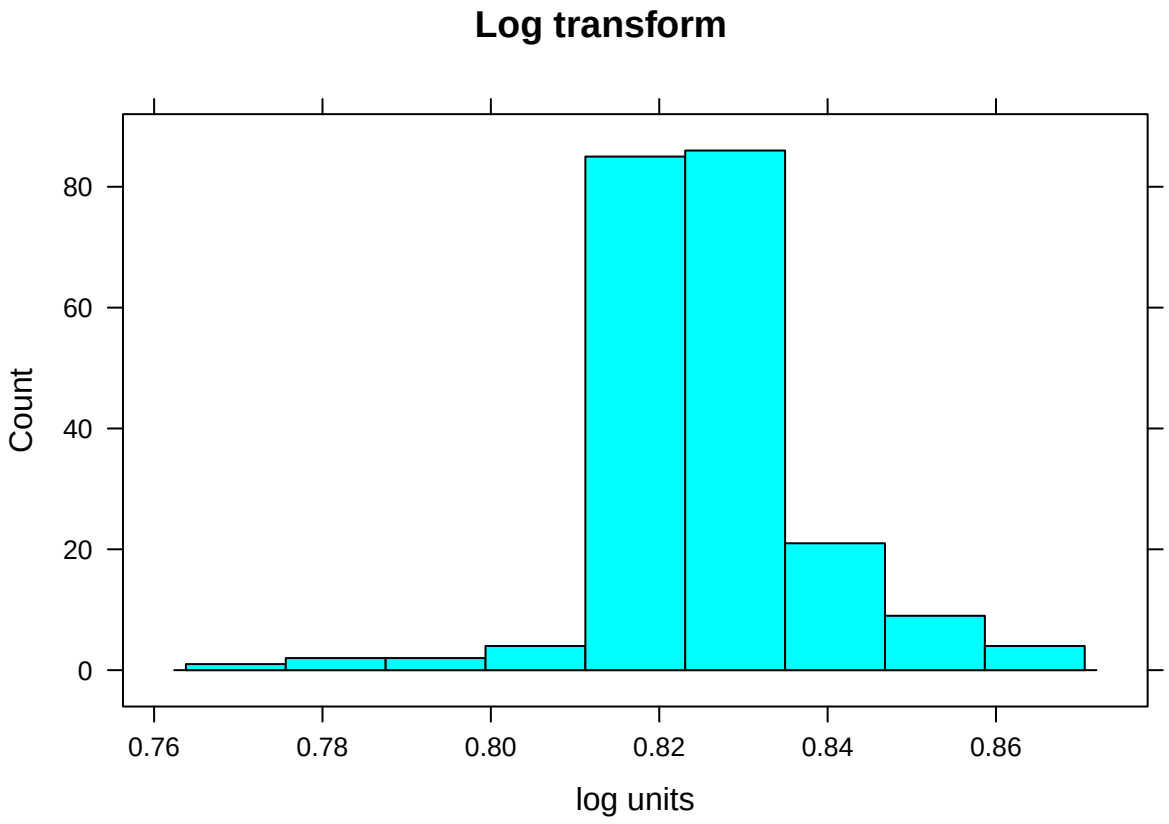
#Based on the histograms, Fe and K seem to have outliers (0.51, and 6.21 respectively). #Based on the skewness and kurtosis statistics, RI, K, Ca, Ba, and Fe are significantly right-skewed, and Mg is significantly left-skewed.

c) Are there any relevant transformations of one or more predictors that might improve the classification model? Show all work!

```
trans <- BoxCoxTrans(Glass$Ca)
b <- predict(trans, Glass$Ca)
par(mfrow = c(1, 2))
histogram(Glass$Ca, xlab='natural units', type='count', main='Original')
```




```
histogram(b, xlab='log units', type='count', main='Log transform')
```




```
describe(b)
```

```
##    vars   n mean   sd median trimmed  mad  min  max range  skew kurtosis se
## X1    1 214 0.83 0.01   0.82    0.82 0.01 0.77 0.87   0.1 -0.19     4.29  0
```

#The BoxCox transformation reduced the skewness of Ca from 2.02 to -0.19

d) Fit SVM model (You may refer to Chapter 4 material for details) using the following codes:

```
# install.packages('kernlab')
library(kernlab)
set.seed(231)
sigDist <- sigest(Type~ ., data = Glass, frac = 1)
sigDist
```

```
##        90%        50%        10%
## 0.03407935 0.11297847 0.62767315
```

```
#   90%        50%        10%
# 0.03407935 0.11297847 0.62767315

svmTuneGrid <- data.frame(sigma = as.vector(sigDist)[1], C = 2^(-2:10))
svmTuneGrid
```

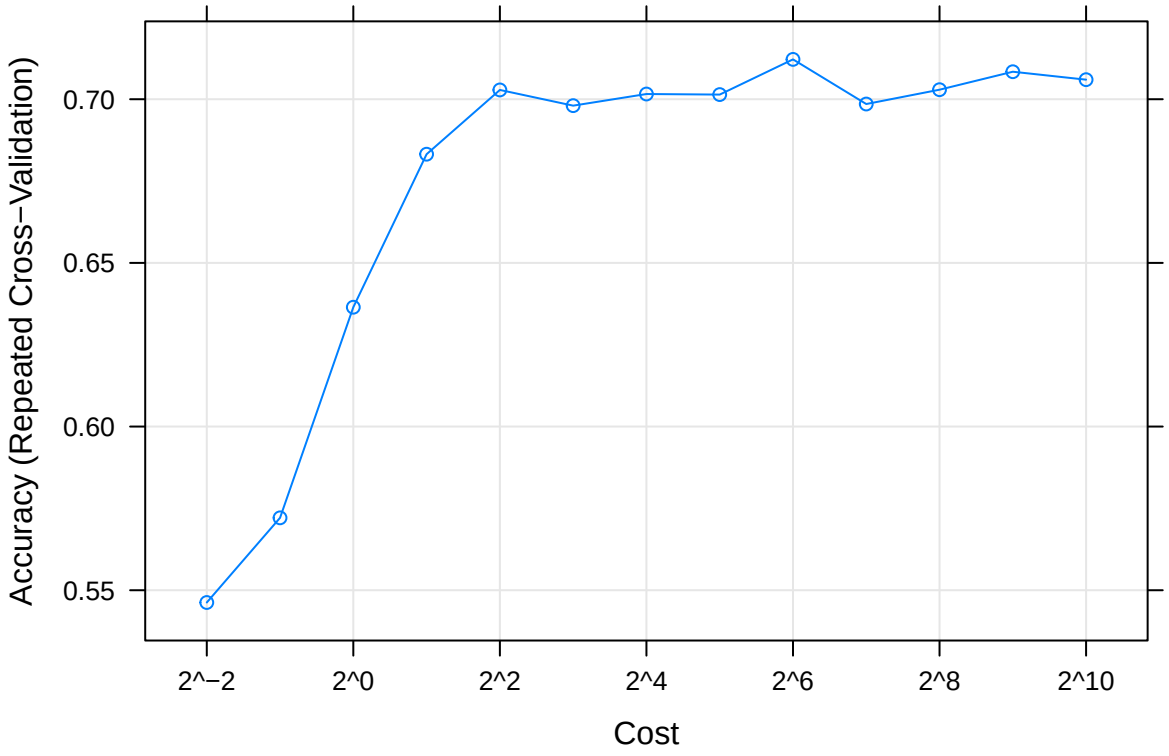
```
##         sigma       C
## 1  0.03407935    0.25
## 2  0.03407935    0.50
## 3  0.03407935    1.00
## 4  0.03407935    2.00
## 5  0.03407935    4.00
## 6  0.03407935    8.00
## 7  0.03407935   16.00
## 8  0.03407935   32.00
## 9  0.03407935   64.00
## 10 0.03407935  128.00
## 11 0.03407935  256.00
## 12 0.03407935  512.00
## 13 0.03407935 1024.00
```

```
#      sigma       C
# 1  0.03407935    0.25
# 2  0.03407935    0.50
# 3  0.03407935    1.00
# 4  0.03407935    2.00
# 5  0.03407935    4.00
# 6  0.03407935    8.00
# 7  0.03407935   16.00
# 8  0.03407935   32.00
# 9  0.03407935   64.00
# 10 0.03407935  128.00
# 11 0.03407935  256.00
# 12 0.03407935  512.00
# 13 0.03407935 1024.00
```

```
library(AppliedPredictiveModeling)
library(caret) #access the train function

set.seed(1056)
#It may take a while to run
svmFit <- train(Type~ ., data = Glass, method = "svmRadial",
preProc = c("center", "scale"),tuneGrid = svmTuneGrid,
trControl = trainControl(method = "repeatedcv", repeats = 5))

plot(svmFit, scales = list(x = list(log = 2)))
```


Accuracy (Repeated Cross−Validation)
0.70
0.65
0.60
0.55
2^−2
2^0
2^2
2^4
2^6
2^8
2^10
Cost