KJ3

Joseph Elikishvili

October 29, 2017

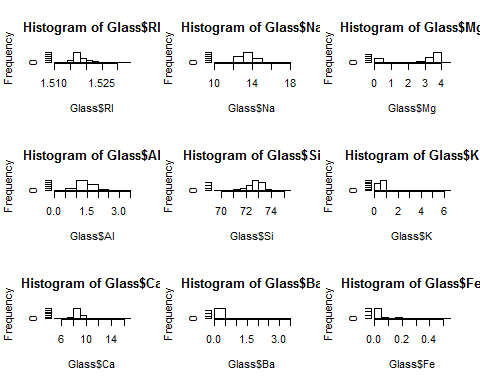
## KJ3.1

#### A. Using visualizations, explore the predictor variables to understand their

#load the data  
data(segmentationOriginal)  
library(mlbench)  
data(Glass)  
  
head(Glass)

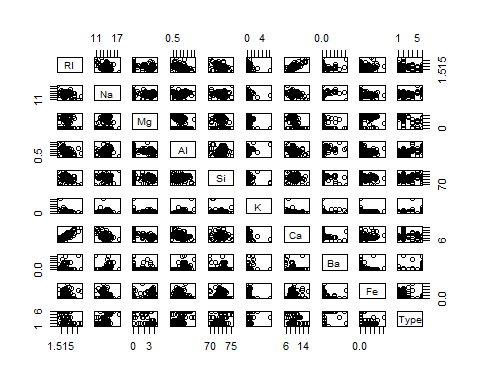
## RI Na Mg Al Si K Ca Ba Fe Type  
## 1 1.52101 13.64 4.49 1.10 71.78 0.06 8.75 0 0.00 1  
## 2 1.51761 13.89 3.60 1.36 72.73 0.48 7.83 0 0.00 1  
## 3 1.51618 13.53 3.55 1.54 72.99 0.39 7.78 0 0.00 1  
## 4 1.51766 13.21 3.69 1.29 72.61 0.57 8.22 0 0.00 1  
## 5 1.51742 13.27 3.62 1.24 73.08 0.55 8.07 0 0.00 1  
## 6 1.51596 12.79 3.61 1.62 72.97 0.64 8.07 0 0.26 1

par(mfrow=c(3,3))  
hist(Glass$RI)  
hist(Glass$Na)  
hist(Glass$Mg)  
hist(Glass$Al)  
hist(Glass$Si)  
hist(Glass$K)  
hist(Glass$Ca)  
hist(Glass$Ba)  
hist(Glass$Fe)



lets look at the distribution of each predictor, we can see that some of the predictors are close to be normally distributed. RI, Si, Na, AI, they show somew skew, but are close to normal. Ba, Fe, K show very strong skew in data. Mg is not unimodal.

pairs(Glass)

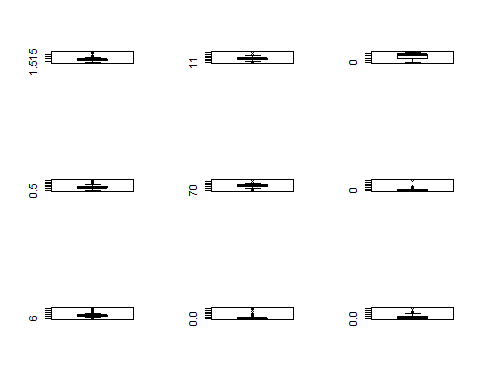


Next we will take a look at predictors and see if there is any kind of corelation between them. We can see that most of the predictors are not corelated to each other the only pair that shows strong corelation is RI/Ca

#### B. Do there appear to be any outliers in the data? Are any predictors skewed?

Lets take a look at boxplots

par(mfrow=c(3,3))  
boxplot(Glass$RI)  
boxplot(Glass$Na)  
boxplot(Glass$Mg)  
boxplot(Glass$Al)  
boxplot(Glass$Si)  
boxplot(Glass$K)  
boxplot(Glass$Ca)  
boxplot(Glass$Ba)  
boxplot(Glass$Fe)



We can also count the number of outliers

par(mfrow=c(3,3))  
length(boxplot.stats(Glass$RI)$out)

## [1] 17

length(boxplot.stats(Glass$Na)$out)

## [1] 7

length(boxplot.stats(Glass$Mg)$out)

## [1] 0

length(boxplot.stats(Glass$Al)$out)

## [1] 18

length(boxplot.stats(Glass$Si)$out)

## [1] 12

length(boxplot.stats(Glass$K)$out)

## [1] 7

length(boxplot.stats(Glass$Ca)$out)

## [1] 26

length(boxplot.stats(Glass$Ba)$out)

## [1] 38

length(boxplot.stats(Glass$Fe)$out)

## [1] 12

It appears that Mg has no outliers and Ba has 38 outliers. Also we can see froom boxpplots that Ba outliers are all above the mean. Also K has one outlier that is really away from the rest of the data points

#### C. Are there any relevant transformations of one or more predictors that might improve the classification model?

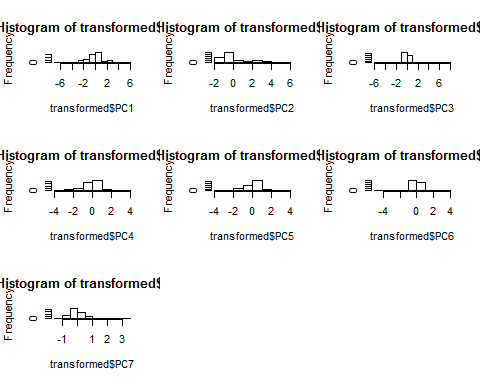
Since classification model will be used with this data, it is normally resistant to outliers, so they do not pose a big threat to the process.

We can try boxcox transformation and see how it affects the data, it is the most broad transfromation and includes log, power and invrse operations.

library(caret)  
trans <- preProcess(Glass, method = c("BoxCox", "center", "scale", "pca"))  
trans

## Created from 214 samples and 10 variables  
##   
## Pre-processing:  
## - Box-Cox transformation (5)  
## - centered (9)  
## - ignored (1)  
## - principal component signal extraction (9)  
## - scaled (9)  
##   
## Lambda estimates for Box-Cox transformation:  
## -2, -0.1, 0.5, 2, -1.1  
## PCA needed 7 components to capture 95 percent of the variance

transformed <- predict(trans, Glass)  
par(mfrow=c(3,3))  
hist(transformed$PC1)  
hist(transformed$PC2)  
hist(transformed$PC3)  
hist(transformed$PC4)  
hist(transformed$PC5)  
hist(transformed$PC6)  
hist(transformed$PC7)



We can see that transformation improved data distribution a lot, it also removed 2 predictors, we saw that 1 was heavily corelated so was not needed and it acutally needed 7 predictors to to capture 95% of variance, so overal there is definately an improvement after the transformation.

## KJ3.2

#### A. Investigate the frequency distributions for the categorical predictors. Are any of the distributions degenerate in the ways discussed earlier in this chapter?

Lets use nearZeroVar function to filter for near-zero variance predictors

library(mlbench)  
data(Soybean)  
nearZeroVar(Soybean)

## [1] 19 26 28

It appears there are 3 columns that need to be removed. Lets remove them

Soy=Soybean[,-c(19,26,28)]

#### B. Roughly 18% of the data are missing. Are there particular predictors that are more likely to be missing? Is the pattern of missing data related to the classes?

We can run summary on the dataset ans see how many NaNs we have per column. Lets do that

summary(Soy)

## Class date plant.stand precip temp   
## brown-spot : 92 5 :149 0 :354 0 : 74 0 : 80   
## alternarialeaf-spot: 91 4 :131 1 :293 1 :112 1 :374   
## frog-eye-leaf-spot : 91 3 :118 NA's: 36 2 :459 2 :199   
## phytophthora-rot : 88 2 : 93 NA's: 38 NA's: 30   
## anthracnose : 44 6 : 90   
## brown-stem-rot : 44 (Other):101   
## (Other) :233 NA's : 1   
## hail crop.hist area.dam sever seed.tmt germ   
## 0 :435 0 : 65 0 :123 0 :195 0 :305 0 :165   
## 1 :127 1 :165 1 :227 1 :322 1 :222 1 :213   
## NA's:121 2 :219 2 :145 2 : 45 2 : 35 2 :193   
## 3 :218 3 :187 NA's:121 NA's:121 NA's:112   
## NA's: 16 NA's: 1   
##   
##   
## plant.growth leaves leaf.halo leaf.marg leaf.size leaf.shread  
## 0 :441 0: 77 0 :221 0 :357 0 : 51 0 :487   
## 1 :226 1:606 1 : 36 1 : 21 1 :327 1 : 96   
## NA's: 16 2 :342 2 :221 2 :221 NA's:100   
## NA's: 84 NA's: 84 NA's: 84   
##   
##   
##   
## leaf.malf stem lodging stem.cankers canker.lesion  
## 0 :554 0 :296 0 :520 0 :379 0 :320   
## 1 : 45 1 :371 1 : 42 1 : 39 1 : 83   
## NA's: 84 NA's: 16 NA's:121 2 : 36 2 :177   
## 3 :191 3 : 65   
## NA's: 38 NA's: 38   
##   
##   
## fruiting.bodies ext.decay int.discolor fruit.pods fruit.spots seed   
## 0 :473 0 :497 0 :581 0 :407 0 :345 0 :476   
## 1 :104 1 :135 1 : 44 1 :130 1 : 75 1 :115   
## NA's:106 2 : 13 2 : 20 2 : 14 2 : 57 NA's: 92   
## NA's: 38 NA's: 38 3 : 48 4 :100   
## NA's: 84 NA's:106   
##   
##   
## mold.growth seed.discolor seed.size shriveling roots   
## 0 :524 0 :513 0 :532 0 :539 0 :551   
## 1 : 67 1 : 64 1 : 59 1 : 38 1 : 86   
## NA's: 92 NA's:106 NA's: 92 NA's:106 2 : 15   
## NA's: 31   
##   
##   
##

We can see from above that some have more data missing then the others, for example sever, seed.tmt, germ, leaf.shred and other have over 100 na values, while some have only a few missing and some have 30-80 missing. So this really from predictor to predictor.

#### C. Develop a strategy for handling missing data, either by eliminating predictors or imputation.

I normally like to impute the missing values by taking the average rather then eliminating, but this is categorical data and besides we have a lot of data missing, and in general we have a lot of predictors, so elimination seems reasonable in this case.

Another option is to encode the missing variables with a special character which means there is no data, but I am not sure this is the case where missing data means anything significant. So my strategy would be to drop the missing values and try to work with whats left.