

DATA 624: PREDICTIVE ANALYTICS HW4

Gabriel Campos

Last edited February 22, 2024

```
library(mlbench)
```

Instructions

Do problems 3.1 and 3.2 in the Kuhn and Johnson book Applied Predictive Modeling. Please submit your Rpubs link along with your .pdf for your run code.

3.1

The UC Irvine Machine Learning Repository⁶ 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. The data can be accessed via:

```
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.

(b)

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

(c)

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

3.2

The soybean data can also be found at the UC Irvine Machine Learning Repository. Data were collected to predict disease in 683 soybeans. The 35 predictors are mostly categorical and include information on the environmental conditions (e.g., temperature, precipitation) and plant conditions (e.g., left spots, mold growth). The outcome labels consist of 19 distinct classes.

The data can be loaded via:

```
data(Soybean)
## See ?Soybean for details
```

(a)

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

(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? ## (c)

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