Homework 3

Tianyu Liu

2023-11-11

Problem Set 13, Applications

1

(a)

```
vehdata <- read.csv("vehicle.csv")
summary(vehdata)
```

```
##   Compactness      Circularity    Distance.Circularity  Radius.Ratio
##  Min.   : 73.00   Min.   :33.00   Min.   : 40.00       Min.   :104.0
##  1st Qu.: 87.00   1st Qu.:40.00   1st Qu.: 70.00       1st Qu.:141.0
##  Median : 93.00   Median :44.00   Median : 80.00       Median :167.0
##  Mean   : 93.68   Mean   :44.86   Mean   : 82.09       Mean   :168.9
##  3rd Qu.:100.00   3rd Qu.:49.00   3rd Qu.: 98.00       3rd Qu.:195.0
##  Max.   :119.00   Max.   :59.00   Max.   :112.00       Max.   :333.0
##  Pr.Axis.Aspect.Ratio Max.Length.Aspect.Ratio Scatter.Ratio   Elongatedness
##  Min.   : 47.00       Min.   : 2.000          Min.   :112.0   Min.   :26.00
##  1st Qu.: 57.00       1st Qu.: 7.000          1st Qu.:146.2   1st Qu.:33.00
##  Median : 61.00       Median : 8.000          Median :157.0   Median :43.00
##  Mean   : 61.69       Mean   : 8.567          Mean   :168.8   Mean   :40.93
##  3rd Qu.: 65.00       3rd Qu.:10.000          3rd Qu.:198.0   3rd Qu.:46.00
##  Max.   :138.00       Max.   :55.000          Max.   :265.0   Max.   :61.00
##  Pr.Axis.Rectangularity Max.Length.Rectangularity
##  Min.   :17.00          Min.   :118
##  1st Qu.:19.00          1st Qu.:137
##  Median :20.00          Median :146
##  Mean   :20.58          Mean   :148
##  3rd Qu.:23.00          3rd Qu.:159
##  Max.   :29.00          Max.   :188
##  Scaled.Variance.Along.Major.Axis Scaled.Variance.Along.Minor.Axis
##  Min.   :130.0                    Min.   : 184.0
##  1st Qu.:167.0                    1st Qu.: 318.2
##  Median :178.5                    Median : 364.0
##  Mean   :188.6                    Mean   : 439.9
##  3rd Qu.:217.0                    3rd Qu.: 587.0
##  Max.   :320.0                    Max.   :1018.0
##  Scaled.Radius.of.Gyration Skewness.About.Major.Axis Skewness.About.Minor.Axis
##  Min.   :109.0             Min.   : 59.00            Min.   : 0.000
##  1st Qu.:149.0             1st Qu.: 67.00            1st Qu.: 2.000
##  Median :173.0             Median : 71.50            Median : 6.000
##  Mean   :174.7             Mean   : 72.46            Mean   : 6.377
##  3rd Qu.:198.0             3rd Qu.: 75.00            3rd Qu.: 9.000
##  Max.   :268.0             Max.   :135.00            Max.   :22.000
##  Kurtosis.About.Minor.Axis Kurtosis.About.Major.Axis Hollows.Ratio
##  Min.   : 0.0              Min.   :176.0             Min.   :181.0
##  1st Qu.: 5.0              1st Qu.:184.0             1st Qu.:190.2
##  Median :11.0              Median :188.0             Median :197.0
##  Mean   :12.6              Mean   :188.9             Mean   :195.6
##  3rd Qu.:19.0              3rd Qu.:193.0             3rd Qu.:201.0
##  Max.   :41.0              Max.   :206.0             Max.   :211.0
##      class
##  Min.   :1.000
##  1st Qu.:1.250
##  Median :2.000
##  Mean   :2.478
##  3rd Qu.:3.000
##  Max.   :4.000
```

(b)

```
vehdata$class = factor(vehdata$class, labels=c('2D', '4D', 'BUS', 'VAN'))
summary(vehdata$class)
```

```
##  2D  4D BUS VAN
## 212 217 218 199
```

(c)

```
cor(vehdata[,1:18])
```

```
##                                Compactness Circularity Distance.Circularity
## Compactness                     1.00000000  0.69286923            0.7924440
## Circularity                     0.69286923  1.00000000            0.7984920
## Distance.Circularity            0.79244402  0.79849200            1.0000000
## Radius.Ratio                    0.69165856  0.62277837            0.7716439
## Pr.Axis.Aspect.Ratio            0.09322213  0.14969187            0.1615292
## Max.Length.Aspect.Ratio         0.14824919  0.24746673            0.2643086
## Scatter.Ratio                   0.81300326  0.86036714            0.9072801
## Elongatedness                  -0.78864702 -0.82875480           -0.9123072
## Pr.Axis.Rectangularity          0.81343702  0.85792532            0.8953261
## Max.Length.Rectangularity       0.67614317  0.96577578            0.7745238
## Scaled.Variance.Along.Major.Axis  0.76441546  0.80849631            0.8644323
## Scaled.Variance.Along.Minor.Axis  0.81863161  0.85267941            0.8896611
## Scaled.Radius.of.Gyration       0.58534709  0.93608041            0.7058181
## Skewness.About.Major.Axis      -0.25298400  0.05866929           -0.2316598
## Skewness.About.Minor.Axis       0.23369282  0.14843285            0.1164777
## Kurtosis.About.Minor.Axis       0.15677928 -0.01548177            0.2645222
## Kurtosis.About.Major.Axis       0.29778044 -0.11304723            0.1479800
## Hollows.Ratio                   0.36555185  0.03867702            0.3354525
##                                Radius.Ratio Pr.Axis.Aspect.Ratio
## Compactness                       0.69165856           0.09322213
## Circularity                       0.62277837           0.14969187
## Distance.Circularity              0.77164394           0.16152916
## Radius.Ratio                      1.00000000           0.66540656
## Pr.Axis.Aspect.Ratio              0.66540656           1.00000000
## Max.Length.Aspect.Ratio           0.44804838           0.64809643
## Scatter.Ratio                     0.73846048           0.10606364
## Elongatedness                    -0.79255741          -0.18508583
## Pr.Axis.Rectangularity            0.71149706           0.07998168
## Max.Length.Rectangularity         0.57015440           0.12916829
## Scaled.Variance.Along.Major.Axis  0.79758793           0.27470468
## Scaled.Variance.Along.Minor.Axis  0.72578142           0.09218201
## Scaled.Radius.of.Gyration         0.53906903           0.12440783
## Skewness.About.Major.Axis        -0.18241222           0.15230355
## Skewness.About.Minor.Axis         0.05053425          -0.05680698
## Kurtosis.About.Minor.Axis         0.17429937          -0.03417976
## Kurtosis.About.Major.Axis         0.38075719           0.23819525
## Hollows.Ratio                     0.46927820           0.26738123
##                                Max.Length.Aspect.Ratio Scatter.Ratio
## Compactness                                 0.14824919   0.813003257
## Circularity                                 0.24746673   0.860367138
## Distance.Circularity                        0.26430861   0.907280057
## Radius.Ratio                                0.44804838   0.738460476
## Pr.Axis.Aspect.Ratio                        0.64809643   0.106063643
## Max.Length.Aspect.Ratio                     1.00000000   0.166769266
## Scatter.Ratio                               0.16676927   1.000000000
## Elongatedness                              -0.18004430  -0.973385344
## Pr.Axis.Rectangularity                      0.16186089   0.992088347
## Max.Length.Rectangularity                   0.30594269   0.810647613
## Scaled.Variance.Along.Major.Axis            0.31942961   0.951862122
## Scaled.Variance.Along.Minor.Axis            0.14460641   0.996318013
## Scaled.Radius.of.Gyration                   0.18979054   0.800931523
```

```
## Skewness.About.Major.Axis                   0.29454772  -0.028618420
## Skewness.About.Minor.Axis                   0.01521789   0.074347173
## Kurtosis.About.Minor.Axis                   0.04337924   0.210706455
## Kurtosis.About.Major.Axis                  -0.02731663   0.004053166
## Hollows.Ratio                               0.14391873   0.119949827
##                                Elongatedness Pr.Axis.Rectangularity
## Compactness                      -0.78864702             0.81343702
## Circularity                      -0.82875480             0.85792532
## Distance.Circularity             -0.91230719             0.89532606
## Radius.Ratio                     -0.79255741             0.71149706
## Pr.Axis.Aspect.Ratio             -0.18508583             0.07998168
## Max.Length.Aspect.Ratio          -0.18004430             0.16186089
## Scatter.Ratio                    -0.97338534             0.99208835
## Elongatedness                     1.00000000            -0.95051244
## Pr.Axis.Rectangularity           -0.95051244             1.00000000
## Max.Length.Rectangularity        -0.77553091             0.81330473
## Scaled.Variance.Along.Major.Axis -0.93839190             0.93826639
## Scaled.Variance.Along.Minor.Axis -0.95652176             0.99234619
## Scaled.Radius.of.Gyration        -0.76614632             0.79828199
## Skewness.About.Major.Axis         0.10487465            -0.01633331
## Skewness.About.Minor.Axis        -0.05334635             0.08234563
## Kurtosis.About.Minor.Axis        -0.18518873             0.21308767
## Kurtosis.About.Major.Axis        -0.11204551            -0.02205601
## Hollows.Ratio                    -0.21672508             0.09851910
##                                Max.Length.Rectangularity
## Compactness                                  0.676143173
## Circularity                                  0.965775776
## Distance.Circularity                         0.774523840
## Radius.Ratio                                 0.570154405
## Pr.Axis.Aspect.Ratio                         0.129168289
## Max.Length.Aspect.Ratio                      0.305942689
## Scatter.Ratio                                0.810647613
## Elongatedness                               -0.775530913
## Pr.Axis.Rectangularity                       0.813304735
## Max.Length.Rectangularity                    1.000000000
## Scaled.Variance.Along.Major.Axis             0.747151993
## Scaled.Variance.Along.Minor.Axis             0.798071495
## Scaled.Radius.of.Gyration                    0.866478540
## Skewness.About.Major.Axis                    0.040339043
## Skewness.About.Minor.Axis                    0.137473338
## Kurtosis.About.Minor.Axis                    0.001183035
## Kurtosis.About.Major.Axis                   -0.107718191
## Hollows.Ratio                                0.076769617
##                                Scaled.Variance.Along.Major.Axis
## Compactness                                         0.76441546
## Circularity                                         0.80849631
## Distance.Circularity                                0.86443228
## Radius.Ratio                                        0.79758793
## Pr.Axis.Aspect.Ratio                                0.27470468
## Max.Length.Aspect.Ratio                             0.31942961
## Scatter.Ratio                                       0.95186212
## Elongatedness                                      -0.93839190
```

```
## Pr.Axis.Rectangularity                          0.93826639
## Max.Length.Rectangularity                       0.74715199
## Scaled.Variance.Along.Major.Axis                1.00000000
## Scaled.Variance.Along.Minor.Axis                0.94988787
## Scaled.Radius.of.Gyration                       0.78096248
## Skewness.About.Major.Axis                       0.11129593
## Skewness.About.Minor.Axis                       0.03913734
## Kurtosis.About.Minor.Axis                       0.19194816
## Kurtosis.About.Major.Axis                       0.01118902
## Hollows.Ratio                                   0.08553952
##                              Scaled.Variance.Along.Minor.Axis
## Compactness                                      0.818631612
## Circularity                                      0.852679411
## Distance.Circularity                             0.889661065
## Radius.Ratio                                     0.725781417
## Pr.Axis.Aspect.Ratio                             0.092182007
## Max.Length.Aspect.Ratio                          0.144606405
## Scatter.Ratio                                    0.996318013
## Elongatedness                                   -0.956521755
## Pr.Axis.Rectangularity                           0.992346188
## Max.Length.Rectangularity                        0.798071495
## Scaled.Variance.Along.Major.Axis                 0.949887874
## Scaled.Variance.Along.Minor.Axis                 1.000000000
## Scaled.Radius.of.Gyration                        0.797539612
## Skewness.About.Major.Axis                       -0.019872473
## Skewness.About.Minor.Axis                        0.076435424
## Kurtosis.About.Minor.Axis                        0.202893719
## Kurtosis.About.Major.Axis                        0.005379651
## Hollows.Ratio                                    0.105432191
##                              Scaled.Radius.of.Gyration
## Compactness                                 0.58534709
## Circularity                                 0.93608041
## Distance.Circularity                        0.70581805
## Radius.Ratio                                0.53906903
## Pr.Axis.Aspect.Ratio                        0.12440783
## Max.Length.Aspect.Ratio                     0.18979054
## Scatter.Ratio                               0.80093152
## Elongatedness                              -0.76614632
## Pr.Axis.Rectangularity                      0.79828199
## Max.Length.Rectangularity                   0.86647854
## Scaled.Variance.Along.Major.Axis            0.78096248
## Scaled.Variance.Along.Minor.Axis            0.79753961
## Scaled.Radius.of.Gyration                   1.00000000
## Skewness.About.Major.Axis                   0.19120943
## Skewness.About.Minor.Axis                   0.16800281
## Kurtosis.About.Minor.Axis                  -0.05621812
## Kurtosis.About.Major.Axis                  -0.22736551
## Hollows.Ratio                              -0.11780842
##                              Skewness.About.Major.Axis
## Compactness                                -0.25298400
## Circularity                                 0.05866929
## Distance.Circularity                       -0.23165977
```

```
## Radius.Ratio                                -0.18241222
## Pr.Axis.Aspect.Ratio                         0.15230355
## Max.Length.Aspect.Ratio                      0.29454772
## Scatter.Ratio                               -0.02861842
## Elongatedness                                0.10487465
## Pr.Axis.Rectangularity                      -0.01633331
## Max.Length.Rectangularity                    0.04033904
## Scaled.Variance.Along.Major.Axis             0.11129593
## Scaled.Variance.Along.Minor.Axis            -0.01987247
## Scaled.Radius.of.Gyration                    0.19120943
## Skewness.About.Major.Axis                    1.00000000
## Skewness.About.Minor.Axis                   -0.09253891
## Kurtosis.About.Minor.Axis                   -0.12615735
## Kurtosis.About.Major.Axis                   -0.75414191
## Hollows.Ratio                               -0.80539854
##                                 Skewness.About.Minor.Axis
## Compactness                                  0.23369282
## Circularity                                  0.14843285
## Distance.Circularity                         0.11647771
## Radius.Ratio                                 0.05053425
## Pr.Axis.Aspect.Ratio                        -0.05680698
## Max.Length.Aspect.Ratio                      0.01521789
## Scatter.Ratio                                0.07434717
## Elongatedness                               -0.05334635
## Pr.Axis.Rectangularity                       0.08234563
## Max.Length.Rectangularity                    0.13747334
## Scaled.Variance.Along.Major.Axis             0.03913734
## Scaled.Variance.Along.Minor.Axis             0.07643542
## Scaled.Radius.of.Gyration                    0.16800281
## Skewness.About.Major.Axis                   -0.09253891
## Skewness.About.Minor.Axis                    1.00000000
## Kurtosis.About.Minor.Axis                   -0.03936531
## Kurtosis.About.Major.Axis                    0.11693185
## Hollows.Ratio                                0.09914923
##                                 Kurtosis.About.Minor.Axis
## Compactness                                  0.156779278
## Circularity                                 -0.015481769
## Distance.Circularity                         0.264522153
## Radius.Ratio                                 0.174299370
## Pr.Axis.Aspect.Ratio                        -0.034179762
## Max.Length.Aspect.Ratio                      0.043379243
## Scatter.Ratio                                0.210706455
## Elongatedness                               -0.185188734
## Pr.Axis.Rectangularity                       0.213087671
## Max.Length.Rectangularity                    0.001183035
## Scaled.Variance.Along.Major.Axis             0.191948164
## Scaled.Variance.Along.Minor.Axis             0.202893719
## Scaled.Radius.of.Gyration                   -0.056218120
## Skewness.About.Major.Axis                   -0.126157345
## Skewness.About.Minor.Axis                   -0.039365312
## Kurtosis.About.Minor.Axis                    1.000000000
## Kurtosis.About.Major.Axis                    0.079089910
```

```
## Hollows.Ratio                                0.204923603
##                              Kurtosis.About.Major.Axis Hollows.Ratio
## Compactness                                0.297780441    0.36555185
## Circularity                               -0.113047232    0.03867702
## Distance.Circularity                       0.147979965    0.33545253
## Radius.Ratio                               0.380757189    0.46927820
## Pr.Axis.Aspect.Ratio                       0.238195253    0.26738123
## Max.Length.Aspect.Ratio                   -0.027316627    0.14391873
## Scatter.Ratio                              0.004053166    0.11994983
## Elongatedness                             -0.112045506   -0.21672508
## Pr.Axis.Rectangularity                    -0.022056010    0.09851910
## Max.Length.Rectangularity                 -0.107718191    0.07676962
## Scaled.Variance.Along.Major.Axis           0.011189021    0.08553952
## Scaled.Variance.Along.Minor.Axis           0.005379651    0.10543219
## Scaled.Radius.of.Gyration                 -0.227365512   -0.11780842
## Skewness.About.Major.Axis                 -0.754141913   -0.80539854
## Skewness.About.Minor.Axis                  0.116931854    0.09914923
## Kurtosis.About.Minor.Axis                  0.079089910    0.20492360
## Kurtosis.About.Major.Axis                  1.000000000    0.89409781
## Hollows.Ratio                              0.894097812    1.00000000
```

The following variables have strong correlations (beyond ±0.7):
Compactness and Distance.Circularity
Compactness and Scatter.Ratio
Compactness and Elongatedness
Compactness and Pr.Axis.Rectangularity
Compactness and Scaled.Variance.Along.Major.Axis
Compactness and Scaled.Variance.Along.Minor.Axis
Circularity and Distance.Circularity
Circularity and Scatter.Ratio
Circularity and Elongatedness
Circularity and Pr.Axis.Rectangularity
Circularity and Scaled.Variance.Along.Major.Axis
Circularity and Scaled.Variance.Along.Minor.Axis
Distance.Circularity and Radius.Ratio
Distance.Circularity and Pr.Axis.Rectangularity
Distance.Circularity and Max.Length.Rectangularity
Distance.Circularity and Scaled.Variance.Along.Major.Axis
Distance.Circularity and Scaled.Variance.Along.Minor.Axis
Distance.Circularity and Scaled.Radius.of.Gyration
Radius.Ratio and Pr.Axis.Aspect.Ratio
Radius.Ratio and Scatter.Ratio
Radius.Ratio and Elongatedness
Radius.Ratio and Pr.Axis.Rectangularity
Radius.Ratio and Scaled.Variance.Along.Major.Axis
Radius.Ratio and Scaled.Variance.Along.Minor.Axis
Scatter.Ratio and Max.Length.Rectangularity
Scatter.Ratio and Scaled.Radius.of.Gyration
Elongatedness and Max.Length.Rectangularity
Elongatedness and Scaled.Radius.of.Gyration

Pr.Axis.Rectangularity and Max.Length.Rectangularity
Pr.Axis.Rectangularity and Scaled.Radius.of.Gyration
Max.Length.Rectangularity and Scaled.Variance.Along.Major.Axis
Max.Length.Rectangularity and Scaled.Variance.Along.Minor.Axis
Max.Length.Rectangularity and Scaled.Radius.of.Gyration
Scaled.Variance.Along.Major.Axis and Scaled.Radius.of.Gyration
Scaled.Variance.Along.Minor.Axis and Scaled.Radius.of.Gyration
Skewness.About.Major.Axis and Kurtosis.About.Major.Axis
Skewness.About.Major.Axis and Hollows.Ratio
Kurtosis.About.Major.Axis and Hollows.Ratio

The following variables have especially strong correlations (beyond ±0.9):
Circularity and Max.Length.Rectangularity
Circularity and Scaled.Radius.of.Gyration
Distance.Circularity and Scatter.Ratio
Distance.Circularity and Elongatedness
Scatter.Ratio and Elongatedness
Scatter.Ratio and Pr.Axis.Rectangularity
Scatter.Ratio and Scaled.Variance.Along.Major.Axis
Scatter.Ratio and Scaled.Variance.Along.Minor.Axis
Elongatedness and Pr.Axis.Rectangularity
Elongatedness and Scaled.Variance.Along.Major.Axis
Elongatedness and Scaled.Variance.Along.Minor.Axis
Pr.Axis.Rectangularity and Scaled.Variance.Along.Major.Axis
Pr.Axis.Rectangularity and Scaled.Variance.Along.Minor.Axis
Scaled.Variance.Along.Major.Axis and Scaled.Variance.Along.Minor.Axis

2

```
set.seed(46685326, kind = "Mersenne-Twister")
perm <- sample(x = nrow(vehdata))
set1 <- vehdata[which(perm <= 3*nrow(vehdata)/4), ]
set2 <- vehdata[which(perm > 3*nrow(vehdata)/4), ]
head(set1)
```

```
##   Compactness Circularity Distance.Circularity Radius.Ratio
## 1          95          48                  83          178
## 2          91          41                  84          141
## 3         104          50                 106          209
## 4          93          41                  82          159
## 5          85          44                  70          205
## 7          97          43                  73          173
##   Pr.Axis.Aspect.Ratio Max.Length.Aspect.Ratio Scatter.Ratio Elongatedness
## 1                   72                      10           162            42
## 2                   57                       9           149            45
## 3                   66                      10           207            32
## 4                   63                       9           144            46
## 5                  103                      52           149            45
## 7                   65                       6           153            42
##   Pr.Axis.Rectangularity Max.Length.Rectangularity
## 1                     20                      159
## 2                     19                      143
## 3                     23                      158
## 4                     19                      143
## 5                     19                      144
## 7                     19                      143
##   Scaled.Variance.Along.Major.Axis Scaled.Variance.Along.Minor.Axis
## 1                              176                             379
## 2                              170                             330
## 3                              223                             635
## 4                              160                             309
## 5                              241                             325
## 7                              176                             361
##   Scaled.Radius.of.Gyration Skewness.About.Major.Axis Skewness.About.Minor.Axis
## 1                       184                        70                         6
## 2                       158                        72                         9
## 3                       220                        73                        14
## 4                       127                        63                         6
## 5                       188                       127                         9
## 7                       172                        66                        13
##   Kurtosis.About.Minor.Axis Kurtosis.About.Major.Axis Hollows.Ratio class
## 1                        16                       187           197   VAN
## 2                        14                       189           199   VAN
## 3                         9                       188           196    4D
## 4                        10                       199           207   VAN
## 5                        11                       180           183   BUS
## 7                         1                       200           204   BUS
```

```
head(set2)
```

```
##    Compactness Circularity Distance.Circularity Radius.Ratio
## 6          107          57                  106          172
## 11          86          36                   70          143
## 12          90          34                   66          136
## 14          89          42                   85          144
## 15          94          49                   79          203
## 16          96          55                  103          201
##    Pr.Axis.Aspect.Ratio Max.Length.Aspect.Ratio Scatter.Ratio Elongatedness
## 6                    50                       6           255            26
## 11                   61                       9           133            50
## 12                   55                       6           123            54
## 14                   58                      10           152            44
## 15                   71                       5           174            37
## 16                   65                       9           204            32
##    Pr.Axis.Rectangularity Max.Length.Rectangularity
## 6                      28                      169
## 11                     18                      130
## 12                     17                      118
## 14                     19                      144
## 15                     21                      154
## 16                     23                      166
##    Scaled.Variance.Along.Major.Axis Scaled.Variance.Along.Minor.Axis
## 6                               280                             957
## 11                              153                             266
## 12                              148                             224
## 14                              173                             345
## 15                              196                             465
## 16                              227                             624
##    Scaled.Radius.of.Gyration Skewness.About.Major.Axis
## 6                        264                        85
## 11                       127                        66
## 12                       118                        65
## 14                       161                        72
## 15                       206                        71
## 16                       246                        74
##    Skewness.About.Minor.Axis Kurtosis.About.Minor.Axis
## 6                          5                         9
## 11                         2                        10
## 12                         5                        26
## 14                         8                        13
## 15                         6                         2
## 16                         6                         2
##    Kurtosis.About.Major.Axis Hollows.Ratio class
## 6                        181           183   BUS
## 11                       194           202   VAN
## 12                       196           202    4D
## 14                       187           197   VAN
## 15                       197           199   BUS
## 16                       186           194    2D
```

3

```
library(FNN)

### Split up the predictor variables from the class labels.
X.train.raw = set1[, -19]
X.valid.raw = set2[, -19]
Y.train = set1[, 19]
Y.valid = set2[, 19]


### KNN is based on distances. If variables are measured on different
### scales, we can change which points are neighbours by measuring
### in different units.
### a function we can use to rescale the columns of
### a data frame to have mean 0 and SD 1. We can also use it to rescale
### a data frame based on the means and SDs of another (this is useful
### for scaling the validation set to match the training set).

### Rescale x1 using the means and SDs of x2
scale.1 <- function(x1, x2) {
  for (col in 1:ncol(x1)) {
    a <- mean(x2[, col])
    b <- sd(x2[, col])
    x1[, col] <- (x1[, col] - a) / b
  }
  x1
}

### Rescale our training and validation sets
X.train <- scale.1(X.train.raw, X.train.raw)
X.valid <- scale.1(X.valid.raw, X.train.raw) # Watch the order


### Now we can fit a KNN model using the knn function in the FNN
### package. The syntax of the knn function is a bit different from
### what we're used to. The first two inputs are the training and
### validation predictor matrices. The third input is the class labels
### for the training set. We can also set k to the number of
### neighbours we want. The function then outputs predicted class
### labels for the validation set. Let's use 1 neighbour.
pred.knn <- knn(X.train, X.valid, Y.train, k = 1)
```

(a)

```
### Let's make a confusion matrix. We get this using the table()
### function and providing both the predicted and true class labels
### for the validation set. We can also set the axis labels using
### the dnn input.
table(pred.knn, Y.valid, dnn = c("Predicted", "Observed"))
```

```
##          Observed
## Predicted 2D 4D BUS VAN
##       2D  25 18   1   0
##       4D  26 29   1   2
##       BUS  0  1  46   6
##       VAN  4  6   1  46
```

There are 55 vehicles of type 2D, 54 vehicles of type 4D, 59 vehicles of type BUS and 54 vehicles of type VAN. In particular, it seems to be difficult to distinguish between 2D and 4D vehicles, whereas BUS and VAN vehicles are easy to predict.

(b)

```
### Next, let's get the misclassification rate
(misclass.knn <- mean(pred.knn != Y.valid))
```

```
## [1] 0.3113208
```

```
(se.knn <- sapply(misclass.knn, function(r) {
     sqrt(r * (1 - r) / nrow(X.train))
 }))
```

```
## [1] 0.01838941
```

The test misclassification rate is about 31%, and the standard error is approximately ±1.8%

Problem Set 14, Applications

1

(a)

```
### Rescale the columns of x1 so that the columns of x2 fall between 0 and 1
rescale <- function(x1, x2) {
  for (col in 1:ncol(x1)) {
    a <- min(x2[, col])
    b <- max(x2[, col])
    x1[, col] <- (x1[, col] - a) / (b - a)
  }
  x1
}

### Create copies of our datasets and rescale
data.train.scale <- set1
data.valid.scale <- set2
data.train.scale[, -19] <- rescale(data.train.scale[, -19], set1[, -19])
data.valid.scale[, -19] <- rescale(data.valid.scale[, -19], set1[, -19])
summary(data.train.scale[,1:3])
```

```
##   Compactness      Circularity     Distance.Circularity
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.3261   1st Qu.:0.2692   1st Qu.:0.4286
##  Median :0.4348   Median :0.4231   Median :0.5714
##  Mean   :0.4483   Mean   :0.4512   Mean   :0.5973
##  3rd Qu.:0.5652   3rd Qu.:0.6154   3rd Qu.:0.8000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
```

```
summary(data.train.scale[,1:3])
```

```
##   Compactness      Circularity     Distance.Circularity
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.3261   1st Qu.:0.2692   1st Qu.:0.4286
##  Median :0.4348   Median :0.4231   Median :0.5714
##  Mean   :0.4483   Mean   :0.4512   Mean   :0.5973
##  3rd Qu.:0.5652   3rd Qu.:0.6154   3rd Qu.:0.8000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
```

It appears that the data has been correctly scaled, to lie between 0 and 1.

(b)

(i)

```
library(nnet)
library(car)
```

```
## Loading required package: carData
```

```
library(glmnet)
```

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.1-8
```

```
library(MASS)

fit.log.nnet <- multinom(class ~ ., data = data.train.scale)
```

```
## # weights:  80 (57 variable)
## initial  value 878.910625
## iter  10 value 540.126072
## iter  20 value 264.002381
## iter  30 value 222.740935
## iter  40 value 211.568230
## iter  50 value 203.944730
## iter  60 value 200.943742
## iter  70 value 199.287545
## iter  80 value 197.947131
## iter  90 value 196.809842
## iter 100 value 196.152373
## final  value 196.152373
## stopped after 100 iterations
```

```
Anova(fit.log.nnet)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: class
##                                LR Chisq Df Pr(>Chisq)
## Compactness                      55.741  3  4.770e-12 ***
## Circularity                      18.992  3  0.0002744 ***
## Distance.Circularity             19.143  3  0.0002554 ***
## Radius.Ratio                    131.680  3  < 2.2e-16 ***
## Pr.Axis.Aspect.Ratio            141.223  3  < 2.2e-16 ***
## Max.Length.Aspect.Ratio          15.429  3  0.0014845 **
## Scatter.Ratio                     2.759  3  0.4302978
## Elongatedness                     4.330  3  0.2279767
## Pr.Axis.Rectangularity            8.232  3  0.0414498 *
## Max.Length.Rectangularity        28.668  3  2.630e-06 ***
## Scaled.Variance.Along.Major.Axis  31.655  3  6.186e-07 ***
## Scaled.Variance.Along.Minor.Axis   1.058  3  0.7871357
## Scaled.Radius.of.Gyration        26.215  3  8.599e-06 ***
## Skewness.About.Major.Axis        28.790  3  2.479e-06 ***
## Skewness.About.Minor.Axis        13.571  3  0.0035514 **
## Kurtosis.About.Minor.Axis         8.339  3  0.0395014 *
## Kurtosis.About.Major.Axis        99.194  3  < 2.2e-16 ***
## Hollows.Ratio                    61.632  3  2.634e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to Anova, the following variables are important (at 95% confidence level):
Compactness
Circularity
Distance.Circularity
Radius.Ratio
Pr.Axis.Aspect.Ratio
Max.Length.Aspect.Ratio
Max.Length.Rectangularity
Scaled.Variance.Along.Major.Axis
Scaled.Radius.of.Gyration
Skewness.About.Major.Axis
Skewness.About.Minor.Axis
Kurtosis.About.Minor.Axis
Kurtosis.About.Major.Axis
Hollows.Ratio

So that’s 14 out of 18 variables. The other 4 are relatively unimportant.

(ii)

```
pred.log.nnet <- predict(fit.log.nnet, data.valid.scale)
(misclass.log.nnet <- mean(pred.log.nnet != Y.valid))
```

```
## [1] 0.2122642
```

```
(misclass.log.se <- sapply(misclass.log.nnet, function(r) {
     sqrt(r * (1 - r) / nrow(X.train))
 }))
```

```
## [1] 0.01623992
```

The misclassification rate is still just over 21%. The standard error for the multinomial classification is about 1.6%, slightly lower than KNN but not by much.

(iii)

```
table(Y.valid, pred.log.nnet, ### Confusion matrix
  dnn = c("Observed", "Predicted")
)
```

```
##          Predicted
## Observed 2D 4D BUS VAN
##      2D  34 19   1   1
##      4D  18 33   1   2
##      BUS  1  1  47   0
##      VAN  1  0   0  53
```

```
misclass.log.nnet
```

```
## [1] 0.2122642
```

Note that the misclassification rate is lower (just above 21%). We still observe that 2D and 4D vehicles are hard to distinguish and that BUS and VAN are much easier to distinguish.

2

(a)

```
### The glmnet() function uses predictor matrix/response vector syntax,
### so we need to extract these from our training and validation sets.
### We also have to convert the predictors to a matrix using the
### as.matrix() function.
X.train.scale <- as.matrix(data.train.scale[, -19])
Y.train <- data.train.scale[, 19]
X.valid.scale <- as.matrix(data.valid.scale[, -19])
Y.valid <- data.valid.scale[, 19]

### While we're looking at the glmnet package, let's do LASSO. We need to
### choose lambda using CV. Fortunately, the cv.glmnet() function does this
### for us. The syntax for cv.glmnet() is the same as for glmnet().
fit.CV.lasso <- cv.glmnet(X.train.scale, Y.train, family = "multinomial")

### The CV-min values are stored in the output from
### cv.glmnet()
lambda.min <- fit.CV.lasso$lambda.min

### Let's check which predictors are included in each "best" model. We
### can get the coefficients using the coef() function, setting s to
### the appropriate lambda value.
coef(fit.CV.lasso, s = lambda.min)
```

```
## $`2D`
## 19 x 1 sparse Matrix of class "dgCMatrix"
##                                           1
## (Intercept)                        10.21104200
## Compactness                        -6.45355109
## Circularity                         4.17746021
## Distance.Circularity               -0.26588534
## Radius.Ratio                       17.96378752
## Pr.Axis.Aspect.Ratio              -27.99934280
## Max.Length.Aspect.Ratio            -3.63129111
## Scatter.Ratio                       0.02509100
## Elongatedness                       3.08830452
## Pr.Axis.Rectangularity              .
## Max.Length.Rectangularity          -7.33835179
## Scaled.Variance.Along.Major.Axis   -1.49417719
## Scaled.Variance.Along.Minor.Axis   15.44674226
## Scaled.Radius.of.Gyration          -3.02963705
## Skewness.About.Major.Axis         -14.18581458
## Skewness.About.Minor.Axis           0.05707199
## Kurtosis.About.Minor.Axis           0.20129341
## Kurtosis.About.Major.Axis           0.97398338
## Hollows.Ratio                      -5.38678590
##
## $`4D`
## 19 x 1 sparse Matrix of class "dgCMatrix"
##                                            1
## (Intercept)                         13.1396119
## Compactness                          4.4747208
## Circularity                        -12.2757752
## Distance.Circularity                 0.2658853
## Radius.Ratio                        20.5065471
## Pr.Axis.Aspect.Ratio               -27.6999366
## Max.Length.Aspect.Ratio             -9.6494204
## Scatter.Ratio                        .
## Elongatedness                       -3.0883045
## Pr.Axis.Rectangularity               7.8058323
## Max.Length.Rectangularity           -0.8026757
## Scaled.Variance.Along.Major.Axis     .
## Scaled.Variance.Along.Minor.Axis   -10.3225408
## Scaled.Radius.of.Gyration            6.3002800
## Skewness.About.Major.Axis          -10.5404710
## Skewness.About.Minor.Axis            0.1181783
## Kurtosis.About.Minor.Axis           -0.2012934
## Kurtosis.About.Major.Axis          -12.6231148
## Hollows.Ratio                        5.3867859
##
## $BUS
## 19 x 1 sparse Matrix of class "dgCMatrix"
##                                           1
## (Intercept)                         0.2702513
## Compactness                        -4.4747208
## Circularity                         .
```

```
## Distance.Circularity              -6.2573078
## Radius.Ratio                     -62.9143391
## Pr.Axis.Aspect.Ratio             102.7336033
## Max.Length.Aspect.Ratio            3.6312911
## Scatter.Ratio                      .
## Elongatedness                    -19.3133250
## Pr.Axis.Rectangularity             .
## Max.Length.Rectangularity          0.8026757
## Scaled.Variance.Along.Major.Axis  44.7123659
## Scaled.Variance.Along.Minor.Axis   .
## Scaled.Radius.of.Gyration          3.0296370
## Skewness.About.Major.Axis         10.5404710
## Skewness.About.Minor.Axis         -4.9814126
## Kurtosis.About.Minor.Axis          3.0009161
## Kurtosis.About.Major.Axis         46.2391073
## Hollows.Ratio                    -24.6221721
##
## $VAN
## 19 x 1 sparse Matrix of class "dgCMatrix"
##                                          1
## (Intercept)                    -23.62090512
## Compactness                     12.59825716
## Circularity                      .
## Distance.Circularity            21.05765272
## Radius.Ratio                   -17.96378752
## Pr.Axis.Aspect.Ratio            27.69993664
## Max.Length.Aspect.Ratio         28.02519655
## Scatter.Ratio                  -33.56834420
## Elongatedness                    9.32075772
## Pr.Axis.Rectangularity         -18.23841429
## Max.Length.Rectangularity       36.55491373
## Scaled.Variance.Along.Major.Axis   .
## Scaled.Variance.Along.Minor.Axis   .
## Scaled.Radius.of.Gyration      -18.54739118
## Skewness.About.Major.Axis       34.48477108
## Skewness.About.Minor.Axis       -0.05707199
## Kurtosis.About.Minor.Axis       -3.14142057
## Kurtosis.About.Major.Axis       -0.97398338
## Hollows.Ratio                   12.97849013
```

```
### Now we can get predictions for both "best" models
pred.lasso.min <- predict(fit.CV.lasso, X.valid.scale,
  s = lambda.min,
  type = "class"
)

table(Y.valid, pred.lasso.min, dnn = c("Obs", "Pred"))
```

```
##      Pred
## Obs   2D 4D BUS VAN
##   2D  33 20   1   1
##   4D  18 34   1   1
##   BUS  0  1  48   0
##   VAN  1  0   2  51
```

For predicting 2D, 17 of the 18 variables are important, except for Pr.Axis.Rectangularity.
For predicting 4D, 16 of the 18 variables are important, except for Scatter.Ratio and Scaled.Variance.Along.Major.Axis.
For predicting BUS, 14 of the 18 variables are important, except for Circularity, Scatter.Ratio, Pr.Axis.Rectangularity, and Scaled.Variance.Along.Minor.Axis.
For predicting VAN, 15 of the 18 variables are important, except for Circularity, Scaled.Variance.Along.Major.Axis, and Scaled.Variance.Along.Minor.Axis.

Recall from ANOVA, that Scatter.Ratio, Elongatedness, Pr.Axis.Rectangularity, and Scaled.Variance.Along.Minor.Axis are unimportant variables, but the LASSO logit model suggests that Circularity would be unimportant for predicting BUS and VAN, and Elongatedness is important in all cases. The other variables generally match with ANOVA.

(b)

```
(miss.lasso.min <- mean(Y.valid != pred.lasso.min))
```

```
## [1] 0.2169811
```

```
(se.miss.lasso.min <- sapply(miss.lasso.min, function(r) {
     sqrt(r * (1 - r) / nrow(X.train.scale))
 }))
```

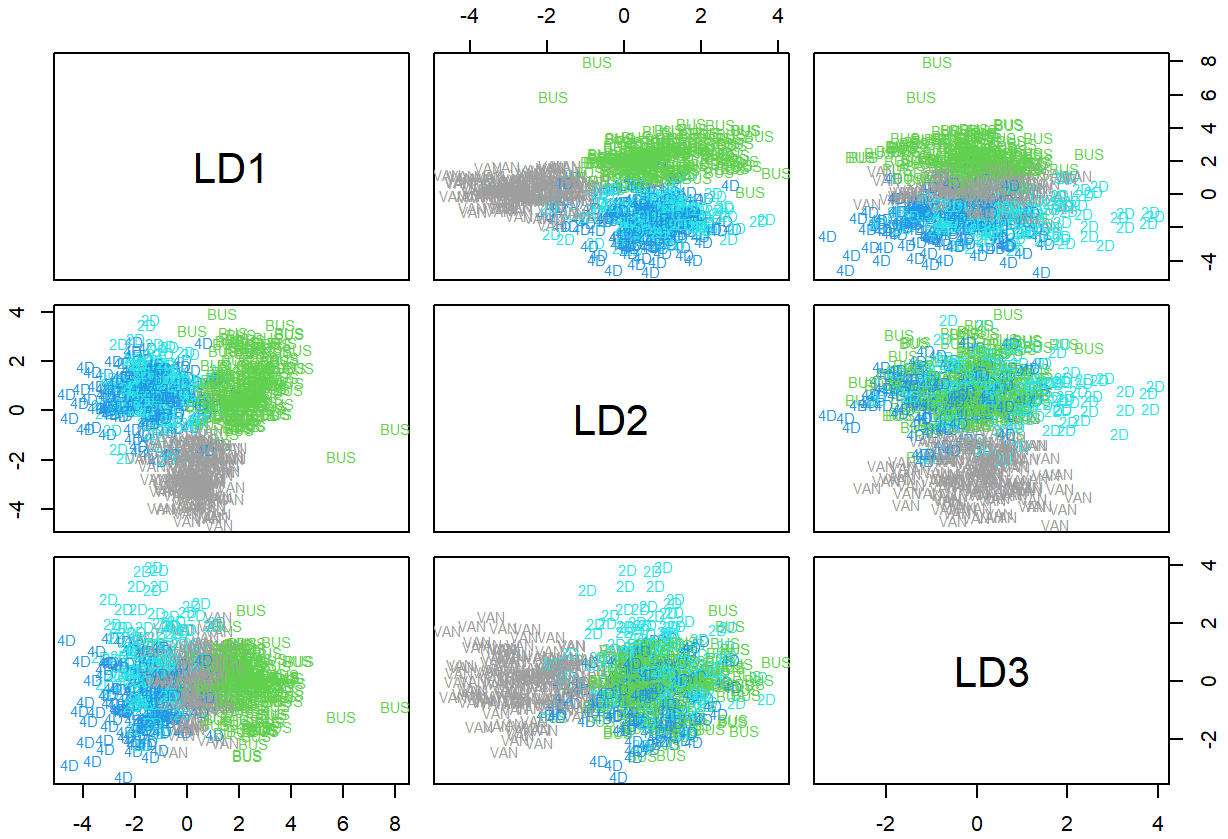
```
## [1] 0.01637014
```

The misclassification rate is now at around 22%. The error is still at 1.6%. No improvement has been made.

3

(a)

```
### For discriminant analysis, it's best to scale predictors
### to have mean 0 and SD 1 (this makes the results easier to
### interpret). We can do this using using the following function.

### Rescale x1 using the means and SDs of x2
scale.1 <- function(x1, x2) {
  for (col in 1:ncol(x1)) {
    a <- mean(x2[, col])
    b <- sd(x2[, col])
    x1[, col] <- (x1[, col] - a) / b
  }
  x1
}


X.train.DA <- scale.1(set1[, -19], set1[, -19])
X.valid.DA <- scale.1(set2[, -19], set1[, -19])

class.col <- ifelse(set1$class=='2D', y = 53,
n = ifelse(set1$class=='4D', y = 68, n = ifelse(set1$class=='BUS',y=203,n=464)))

### Fit an LDA model using the lda() funtion from the MASS package. This
### function uses predictor/response syntax.
fit.lda <- lda(X.train.DA, Y.train)

### We can plot the data using the linear discriminants. It's best to
### include colors. Let's just recycle the colors from above.
### There is no simple way to change the axis labels. Sometimes we just need
### to live with the defaults.
plot(fit.lda, col = class.col)
```



LD1 seems to be separating vehicles of class BUS from the rest.
LD2 seems to be separating vehicles of class 2D and 4D from the rest. It cannot really distinguish between the 2 classes.
LD3 seems to be separating vehicles of class VAN from LD2.

(b)

```
### We get predictions by extracting the class object from the predict()
### function's output.
pred.lda <- predict(fit.lda, X.valid.DA)$class

table(Y.valid, pred.lda, dnn = c("Obs", "Pred"))
```

```
##       Pred
## Obs    2D 4D BUS VAN
##   2D   34 19   1   1
##   4D   15 33   3   3
##   BUS   0  0  49   0
##   VAN   0  0   0  54
```

```
(miss.lda <- mean(Y.valid != pred.lda))
```

```
## [1] 0.1981132
```

```
(se.miss.lda <- sapply(miss.lda, function(r) {
     sqrt(r * (1 - r) / nrow(X.train.DA))
 }))
```

```
## [1] 0.01582955
```

The misclassification rate is just under 20%. The error is just under 1.6%, so it is a very slight improvement compare to the rest.