HW2: Classification

Your Name

due on 10/25 (Tue) 9am

Your analysis should include the following:

- Exploratory data analysis: simple summaries with plots or tables
- Performing the classification task based on
  - logistic regression
  - discriminate analysis
  - nearest neighbor
- Report the performance of your classifiers and compare them among different methods.
- Make your conclusions on data contents.

Data Source

```
library(mlbench) #install package first!!
library(corrplot)
```

Problem 1: Wisconsin Breast Cancer Data

These data consist of 699 observations on 11 variables, one being "ID" variable, 9 being ordered or nominal variables, and 1 target class (benign or malignant). The objective is to classify each example into benign class or malignant class. More data descriptions can be found by typing `help(BreastCancer)` in r command.

Note that:

- There are 16 NA's in variable **Bare.nuclei**. If you don't know how to impute data, you may only use the observations with complete data, or you may drop this variable from your analysis.
- There are high correlations between all input variables, which should be aware of during modeling and data interpretation.

```
data(BreastCancer)
head(BreastCancer)
```

```
##        Id Cl.thickness Cell.size Cell.shape Marg.adhesion Epith.c.size
## 1 1000025            5         1          1             1            2
## 2 1002945            5         4          4             5            7
## 3 1015425            3         1          1             1            2
## 4 1016277            6         8          8             1            3
## 5 1017023            4         1          1             3            2
## 6 1017122            8        10         10             8            7
##   Bare.nuclei Bl.cromatin Normal.nucleoli Mitoses     Class
## 1           1           3               1       1    benign
## 2          10           3               2       1    benign
## 3           2           3               1       1    benign
## 4           4           3               7       1    benign
## 5           1           3               1       1    benign
## 6          10           9               7       1 malignant
```

```
dim(BreastCancer)
```

```
## [1] 699  11
```

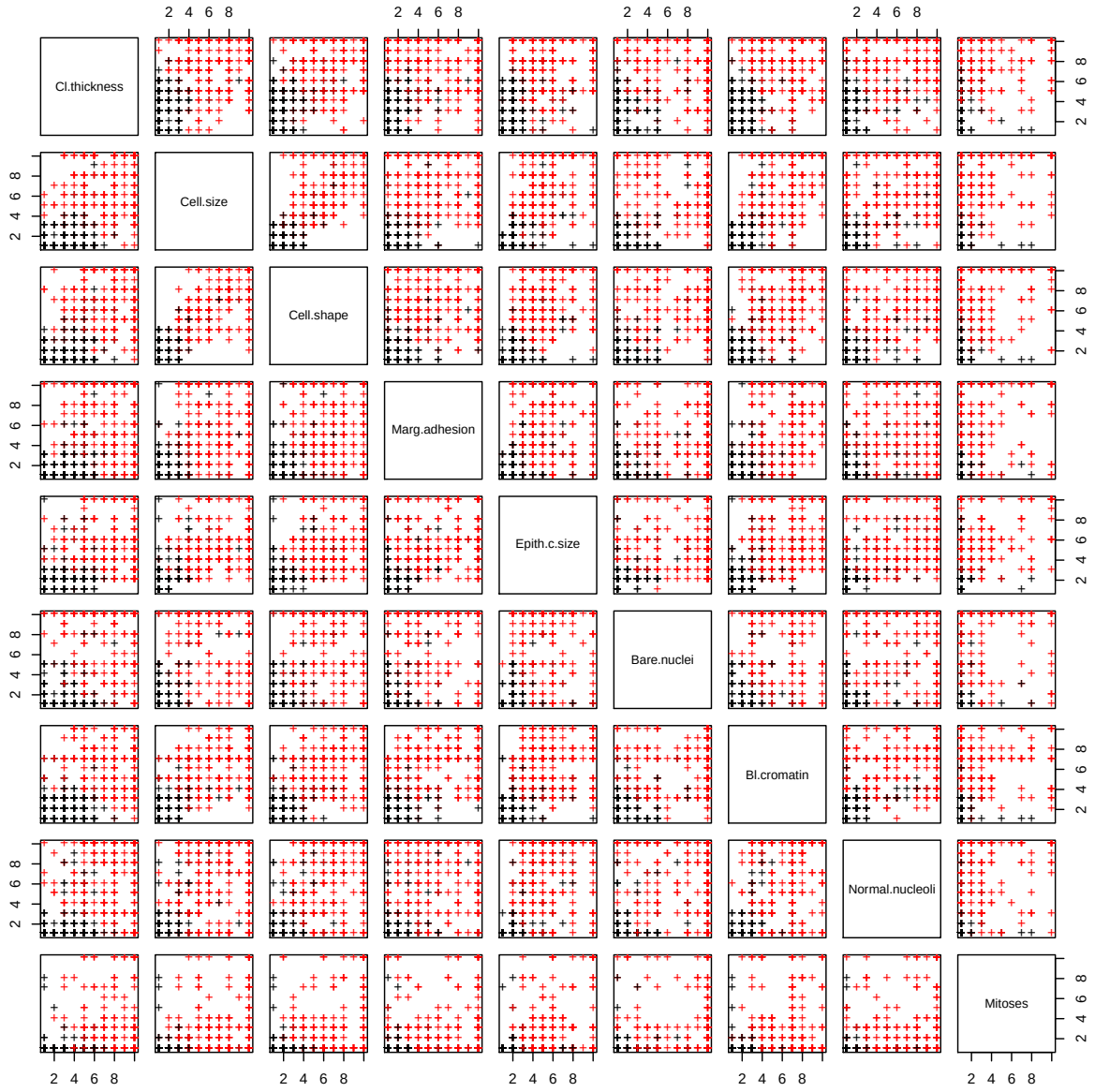
```r
#help(BreastCancer)

#make variables numeric (remove variable:ID) and save the data as a dataframe object
dat1 = matrix(as.numeric(as.matrix(BreastCancer[,2:10])), 699, 9)
dat1 = data.frame(dat1)
colnames(dat1) <- colnames(BreastCancer)[2:10]
head(dat1)
```

```
##   Cl.thickness Cell.size Cell.shape Marg.adhesion Epith.c.size Bare.nuclei
## 1            5         1          1             1            2           1
## 2            5         4          4             5            7          10
## 3            3         1          1             1            2           2
## 4            6         8          8             1            3           4
## 5            4         1          1             3            2           1
## 6            8        10         10             8            7          10
##   Bl.cromatin Normal.nucleoli Mitoses
## 1           3               1       1
## 2           3               2       1
## 3           3               1       1
## 4           3               7       1
## 5           3               1       1
## 6           9               7       1
```

```r
dat1$case = as.numeric(BreastCancer$Class=="malignant")

pairs(dat1[,1:9], col=as.factor(dat1[,10]), pch="+")
```




```
round(cor(dat1, use="pairwise.complete.obs"),2) #handling data with NA
```

```
##                 Cl.thickness Cell.size Cell.shape Marg.adhesion
## Cl.thickness            1.00      0.64       0.65          0.49
## Cell.size               0.64      1.00       0.91          0.71
## Cell.shape              0.65      0.91       1.00          0.68
## Marg.adhesion           0.49      0.71       0.68          1.00
## Epith.c.size            0.52      0.75       0.72          0.60
## Bare.nuclei             0.59      0.69       0.71          0.67
## Bl.cromatin             0.56      0.76       0.74          0.67
## Normal.nucleoli         0.54      0.72       0.72          0.60
## Mitoses                 0.35      0.46       0.44          0.42
## case                    0.72      0.82       0.82          0.70
##                 Epith.c.size Bare.nuclei Bl.cromatin Normal.nucleoli
```

```
## Cl.thickness            0.52        0.59        0.56            0.54
## Cell.size               0.75        0.69        0.76            0.72
## Cell.shape              0.72        0.71        0.74            0.72
## Marg.adhesion           0.60        0.67        0.67            0.60
## Epith.c.size            1.00        0.59        0.62            0.63
## Bare.nuclei             0.59        1.00        0.68            0.58
## Bl.cromatin             0.62        0.68        1.00            0.67
## Normal.nucleoli         0.63        0.58        0.67            1.00
## Mitoses                 0.48        0.34        0.34            0.43
## case                    0.68        0.82        0.76            0.71
##                 Mitoses case
## Cl.thickness       0.35 0.72
## Cell.size          0.46 0.82
## Cell.shape         0.44 0.82
## Marg.adhesion      0.42 0.70
## Epith.c.size       0.48 0.68
## Bare.nuclei        0.34 0.82
## Bl.cromatin        0.34 0.76
## Normal.nucleoli    0.43 0.71
## Mitoses            1.00 0.42
## case               0.42 1.00
```

```
#remove missing data (NA)
dat1 = na.omit(dat1)
dim(dat1) #check data dimension
```

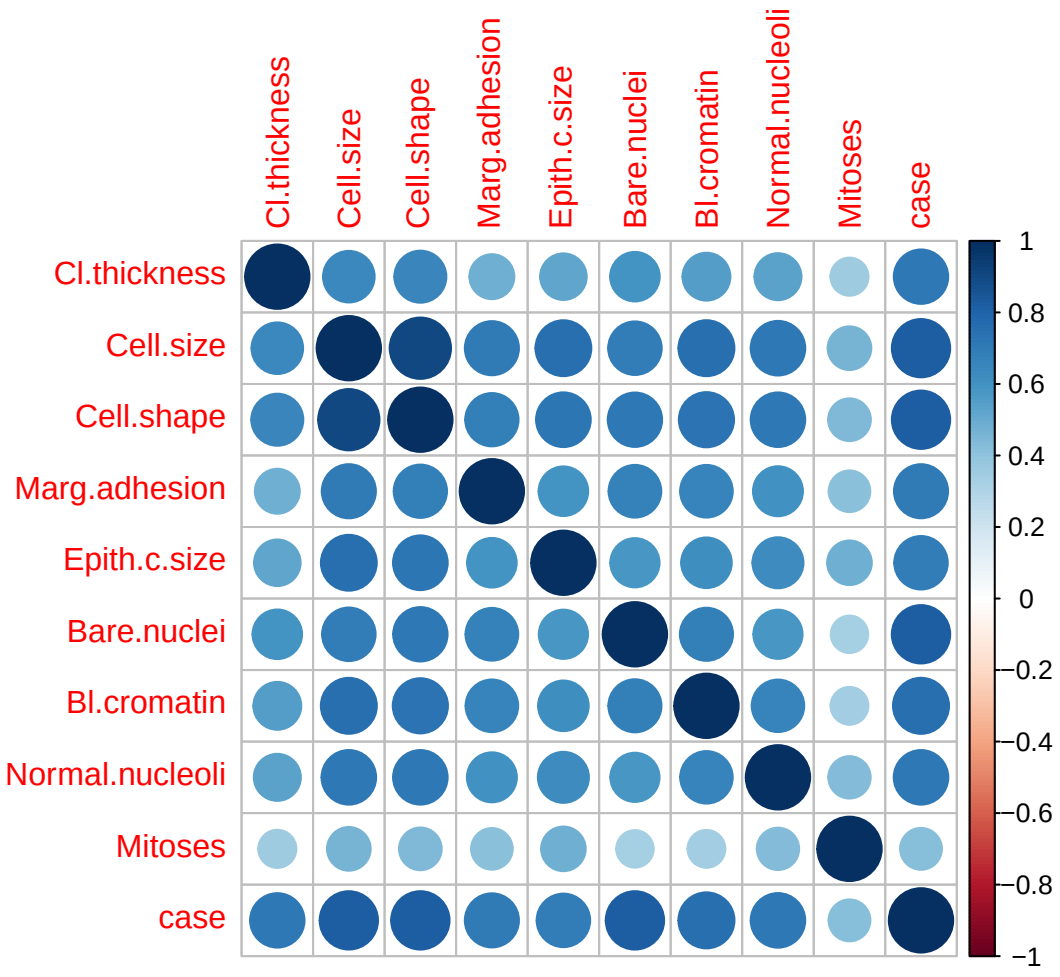
```
## [1] 683  10
```

```
#view variable correlations:
round(cor(dat1),2)
```

```
##                 Cl.thickness Cell.size Cell.shape Marg.adhesion
## Cl.thickness            1.00      0.64       0.65          0.49
## Cell.size               0.64      1.00       0.91          0.71
## Cell.shape              0.65      0.91       1.00          0.69
## Marg.adhesion           0.49      0.71       0.69          1.00
## Epith.c.size            0.52      0.75       0.72          0.59
## Bare.nuclei             0.59      0.69       0.71          0.67
## Bl.cromatin             0.55      0.76       0.74          0.67
## Normal.nucleoli         0.53      0.72       0.72          0.60
## Mitoses                 0.35      0.46       0.44          0.42
## case                    0.71      0.82       0.82          0.71
##                 Epith.c.size Bare.nuclei Bl.cromatin Normal.nucleoli
## Cl.thickness            0.52        0.59        0.55            0.53
## Cell.size               0.75        0.69        0.76            0.72
## Cell.shape              0.72        0.71        0.74            0.72
## Marg.adhesion           0.59        0.67        0.67            0.60
## Epith.c.size            1.00        0.59        0.62            0.63
## Bare.nuclei             0.59        1.00        0.68            0.58
## Bl.cromatin             0.62        0.68        1.00            0.67
## Normal.nucleoli         0.63        0.58        0.67            1.00
## Mitoses                 0.48        0.34        0.35            0.43
## case                    0.69        0.82        0.76            0.72
##                 Mitoses case
## Cl.thickness       0.35 0.71
## Cell.size          0.46 0.82
```

```
## Cell.shape          0.44 0.82
## Marg.adhesion       0.42 0.71
## Epith.c.size        0.48 0.69
## Bare.nuclei         0.34 0.82
## Bl.cromatin         0.35 0.76
## Normal.nucleoli     0.43 0.72
## Mitoses             1.00 0.42
## case                0.42 1.00
```

```
corrplot(cor(dat1))
```



Problem 2: Glass Data

These data consist of 214 examples of the chemical analysis of 6 different types of glass (the target class to be predicted). There are 9 chemical variables for glass classification. More data descriptions can be found by typing `help(Glass)` in r command.

This problem concerns about multi-class classification. You may use multi-class classification methods for all

class simultaneously, or use 1-vs-1 or 1-vs-others schemes to build up your ensemble classifier.

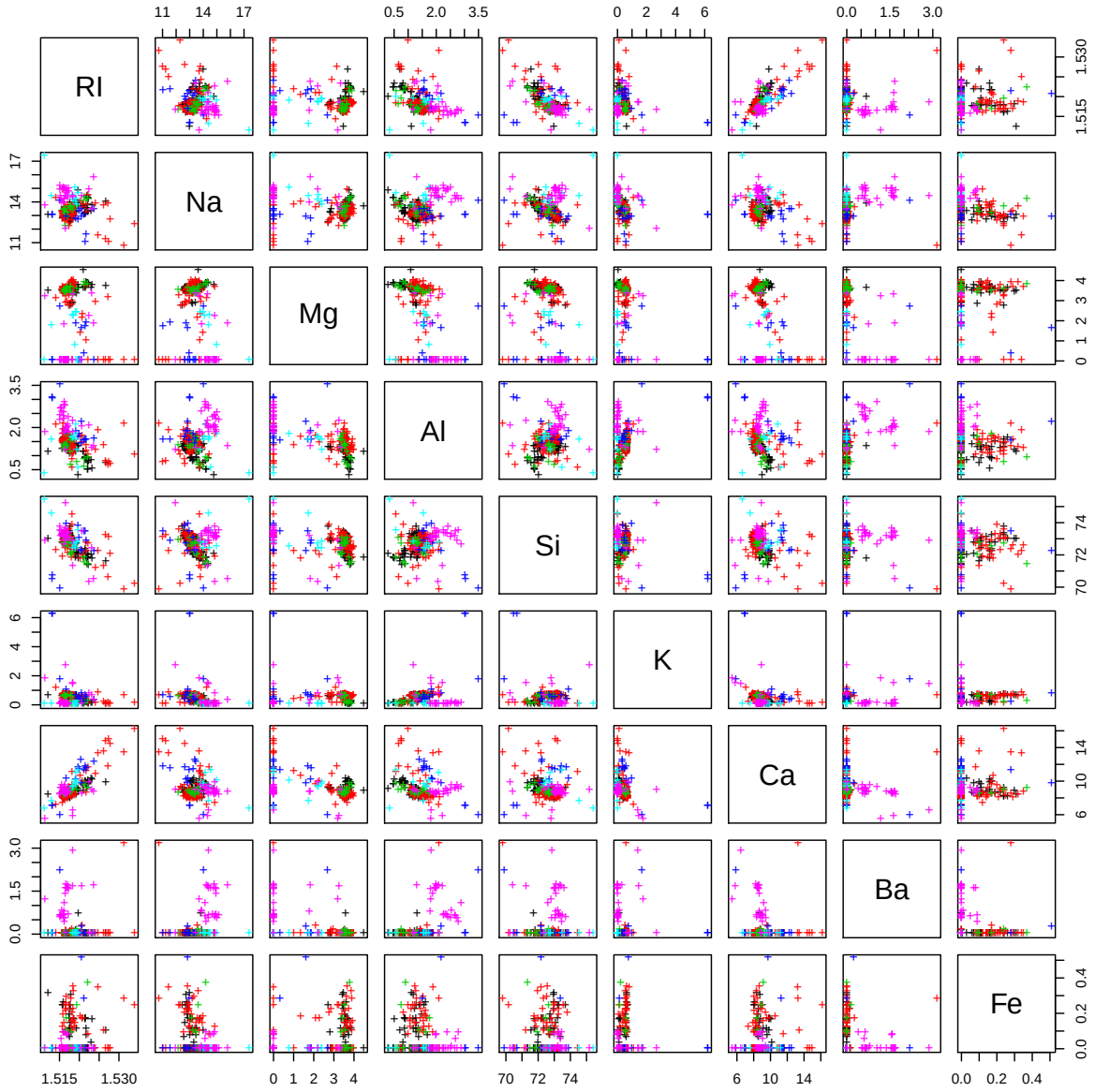
```
data(Glass)
head(Glass)
```

```
##        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
```

```
#View(Glass)
summary(Glass)
```

```
##        RI              Na              Mg              Al
##  Min.   :1.511   Min.   :10.73   Min.   :0.000   Min.   :0.290
##  1st Qu.:1.517   1st Qu.:12.91   1st Qu.:2.115   1st Qu.:1.190
##  Median :1.518   Median :13.30   Median :3.480   Median :1.360
##  Mean   :1.518   Mean   :13.41   Mean   :2.685   Mean   :1.445
##  3rd Qu.:1.519   3rd Qu.:13.82   3rd Qu.:3.600   3rd Qu.:1.630
##  Max.   :1.534   Max.   :17.38   Max.   :4.490   Max.   :3.500
##        Si              K               Ca              Ba
##  Min.   :69.81   Min.   :0.0000   Min.   : 5.430   Min.   :0.000
##  1st Qu.:72.28   1st Qu.:0.1225   1st Qu.: 8.240   1st Qu.:0.000
##  Median :72.79   Median :0.5550   Median : 8.600   Median :0.000
##  Mean   :72.65   Mean   :0.4971   Mean   : 8.957   Mean   :0.175
##  3rd Qu.:73.09   3rd Qu.:0.6100   3rd Qu.: 9.172   3rd Qu.:0.000
##  Max.   :75.41   Max.   :6.2100   Max.   :16.190   Max.   :3.150
##        Fe          Type
##  Min.   :0.00000   1:70
##  1st Qu.:0.00000   2:76
##  Median :0.00000   3:17
##  Mean   :0.05701   5:13
##  3rd Qu.:0.10000   6: 9
##  Max.   :0.51000   7:29
```

```
pairs(Glass[,1:9], col=Glass[,10], pch="+") #view data (colored by glass type)
```




```
dat2 = data.frame(Glass)
round(cor(dat2[,1:9]),2) #only for numeric variables
```

```
##       RI    Na    Mg    Al    Si     K    Ca    Ba    Fe
## RI  1.00 -0.19 -0.12 -0.41 -0.54 -0.29  0.81  0.00  0.14
## Na -0.19  1.00 -0.27  0.16 -0.07 -0.27 -0.28  0.33 -0.24
## Mg -0.12 -0.27  1.00 -0.48 -0.17  0.01 -0.44 -0.49  0.08
## Al -0.41  0.16 -0.48  1.00 -0.01  0.33 -0.26  0.48 -0.07
## Si -0.54 -0.07 -0.17 -0.01  1.00 -0.19 -0.21 -0.10 -0.09
## K  -0.29 -0.27  0.01  0.33 -0.19  1.00 -0.32 -0.04 -0.01
## Ca  0.81 -0.28 -0.44 -0.26 -0.21 -0.32  1.00 -0.11  0.12
## Ba  0.00  0.33 -0.49  0.48 -0.10 -0.04 -0.11  1.00 -0.06
## Fe  0.14 -0.24  0.08 -0.07 -0.09 -0.01  0.12 -0.06  1.00
```

```
corrplot(cor(dat2[,1:9]))
```