Tutorial on applying linear (and quadratic) discriminant analysis

In Part B of this assignment you will have to apply linear and qudratic disciminant analysis. This tutorial shows how to do this, using the Breast Cancer Wisconsin data set from Kaggle (https://www.kaggle.com/uciml/breast-cancer-wisconsin-data/data). We will load the data and select the ID, diagnosis and a selection of ten variables. You can find more information on these variables at the Kaggle site. We are going to use LDA to find the 'cutoff' or 'descision boundary' that most successfully classifies these data into malignant or benign.

```
wdbc <- read.csv("wdbc.csv", header = T)
features <- c("radius", "texture", "perimeter", "area", "smoothness", "compactness",
              "concavity", "concave_points", "symmetry", "fractal_dimension")
names(wdbc) <- c("id", "diagnosis", paste0(features,"_mean"), paste0(features,"_se"),
                 paste0(features,"_worst"))
wdbc.data <- wdbc[,c(3:32)]
row.names(wdbc.data) <- wdbc$id
wdbc$diagnosis[wdbc$diagnosis=='M'] <-1
wdbc$diagnosis[wdbc$diagnosis=='B'] <-0
wdbc_raw <- cbind(wdbc.data, as.factor(wdbc$diagnosis))
colnames(wdbc_raw)[31] <- "diagnosis"
```

Note the last part removes ID as a variable, recodes the diagnosis as either 1 or 0 (deletes a few strays!) and retains the ID in the rownames.

Once the data is split into training (75%) and testing/evaluation (25%) sets, we can use the `lda()` function in R to perform our analysis. The form is the same as the `lm()` and `glm()` functions. We will use the first three of the variables. Note that the `lda()` function is part of the `MASS` package.

```
## Call:
## lda(diagnosis ~ radius_mean + texture_mean + perimeter_mean,
##     data = wdbc.train)
##
## Prior probabilities of groups:
##         0         1
## 0.6502347 0.3497653
##
## Group means:
##   radius_mean texture_mean perimeter_mean
## 0    12.14371     17.95610       78.08097
## 1    17.38201     21.74007      114.74960
##
## Coefficients of linear discriminants:
##                         LD1
## radius_mean    -1.53394873
## texture_mean    0.08332531
## perimeter_mean  0.28153235
```

Finally we can look at the predictions and compare them to the test/evaluation dataset just as we did for other methods.

We can perform QDA in the same way, using the `qda` function (see help file for details).