Análisis Estadístico Multivariante

Problemas Propuestos de Análisis Discriminante

Francisco Javier Mercader Martínez

Problema 1

El conjunto de datos **iris** de R contiene las variables Sepal.Length(X1, longitud de los sépalos), Sepal.Width (X2, anchura de los sépalos), Petal.Length (X3, longitud de los pétalos), Petal.Width (X4, anchura de los pétalos). medidas en centímetros, de tres especies diferentes de flor de iris (Species: *setosa, versicolor* y *virginica*). Se desea realizar un análisis discriminante para determinar la especie de las flores de iris a partir de las magnitudes de sus pétalos y sépalos. Se pide:

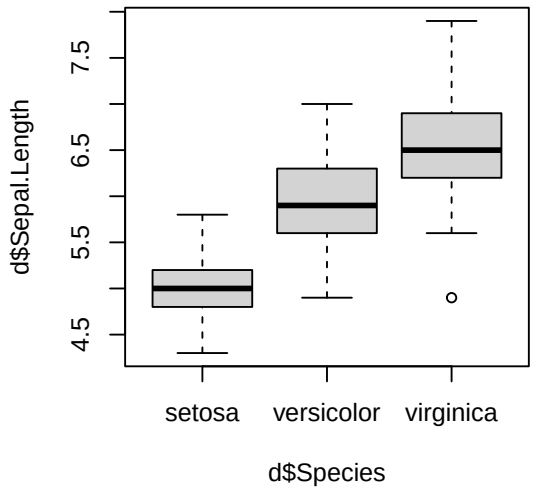
1) Cargar el conjunto de datos y realizar un estudio descriptivo previo atendiendo a nuestro objetivo. En particular, debes dar respuesta a las cuestiones:

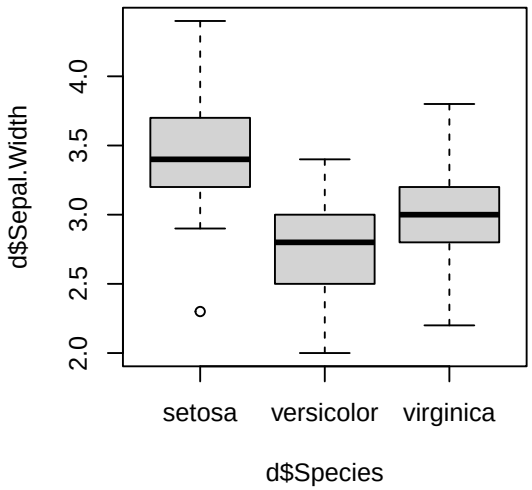
```
d <- iris
summary(d)
```

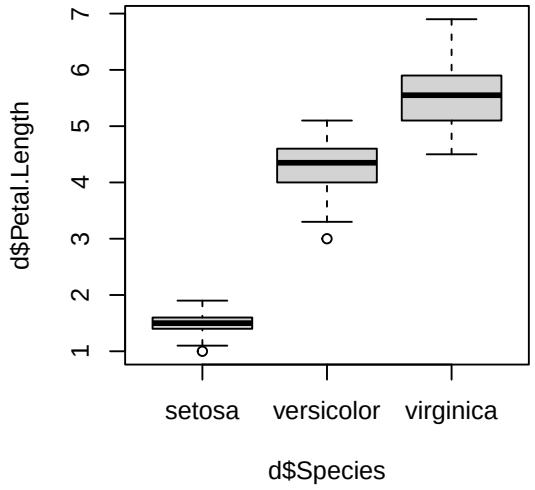
```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

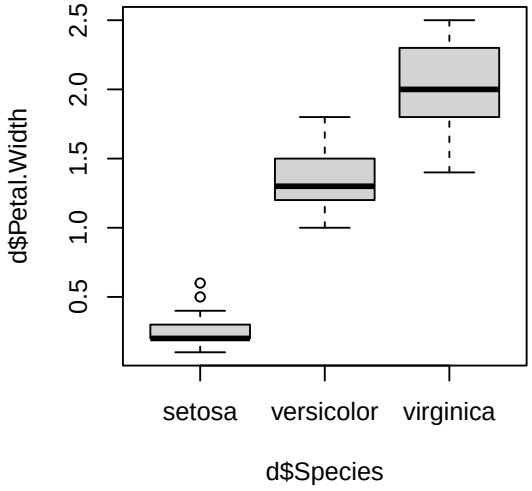
   a. ¿Existe alguna variable que permita discriminar entre las especies? Hacer gráficos caja-bigotes de cada variable distinguiendo por especie.

```
par(mfrow = c(2,2))
boxplot(d$Sepal.Length ~ d$Species)
boxplot(d$Sepal.Width ~ d$Species)
boxplot(d$Petal.Length ~ d$Species)
boxplot(d$Petal.Width ~ d$Species)
```







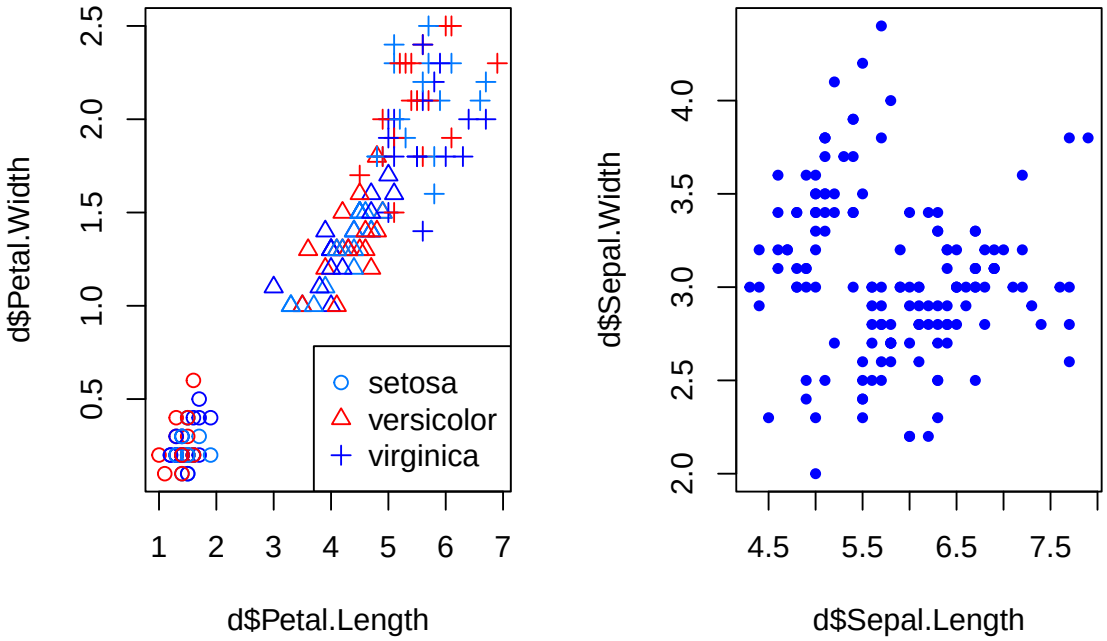


Para cada una de las variables `Petal.Length` y `Petal.width` se pude separar la especie Setosa de las otras dos poque no se solapan los gráficos de caja.

b. ¿Existe alguna variable que permita discriminar entre las especies? Hacer gráficos caja-bigotes de cada variable distinguiendo por especie.

```
par(mfrow = c(1,2))
plot(d$Petal.Length, d$Petal.Width, pch = as.integer(d$Species), col = c("#007AFF",
 ↪  "#FF0000", "#0000FF"))
legend("bottomright", legend = unique(d$Species), pch = 1:3, col = c("#007AFF",
 ↪  "#FF0000", "#0000FF"))

plot(d$Sepal.Length, d$Sepal.Width, pch=20, col = "blue")
```



2) Realizar un análisis discriminante lineal (LDA) con todas las variables y probabilidades a priori iguales.

```
library("MASS")
LDA <- lda(x = d[1:150, 1:4], grouping = d[1:150, 5],
           prior = c(1/3, 1/3, 1/3))
```

a. ¿Cuál es la expresión de las funciones discriminantes en función de las variables observadas?

```
LDA$scaling
```

```
##                    LD1         LD2
## Sepal.Length  0.8293776 -0.02410215
## Sepal.Width   1.5344731 -2.16452123
## Petal.Length -2.2012117  0.93192121
## Petal.Width  -2.8104603 -2.83918785
```

b. ¿Cómo se clasificaría una flor con medidas z = (6, 3, 5, 2)? ¿Cuáles serían las probabilidades a posteriori de pertenencia a cada grupo? ¿Sería fiable la clasificación?

```
z = c(6,3,5,2)
predict(LDA, z)
```

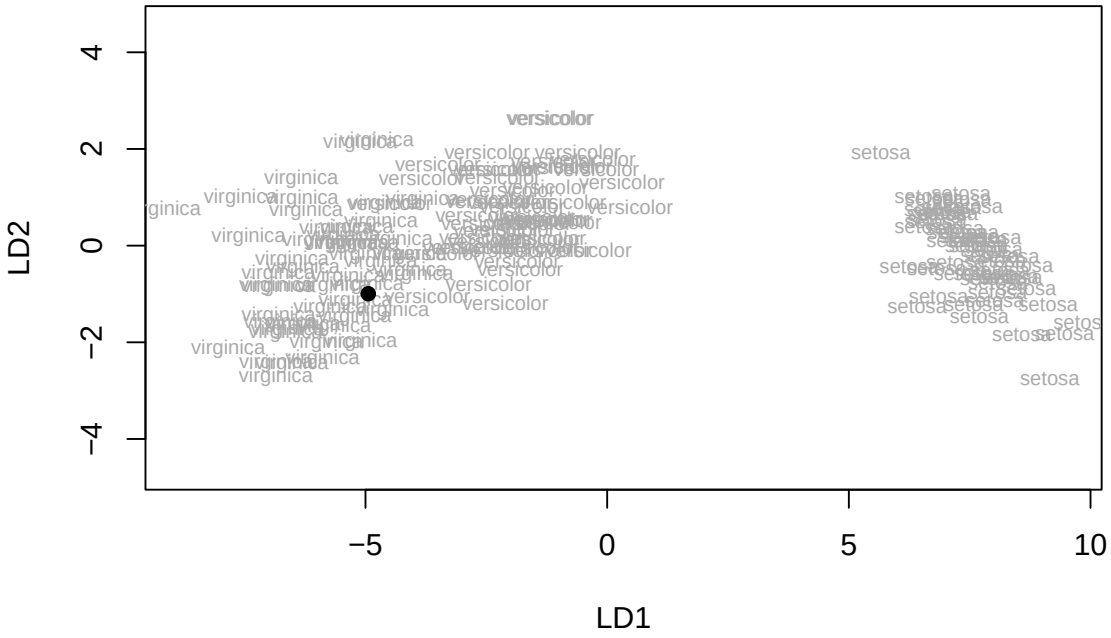
```
## $class
## [1] virginica
## Levels: setosa versicolor virginica
##
## $posterior
##            setosa  versicolor virginica
## [1,] 7.420387e-35 0.002804633 0.9971954
##
```

```
## $x
##             LD1        LD2
## [1,] -4.942187 -0.9954737
```

Se clasifica en virgínica, y la clasificación sería fiable.

c. Representar gráficamente las funciones discriminantes y la flor anterior a clasificar. ¿Dónde se sitúa esta flor? Según el gráfico, ¿en qué grupo la clasificaría?

```
plot(LDA, col = "darkgrey")
points(predict(LDA, z)$x[1, 1], predict(LDA, z)$x[1,2], pch = 19)
```



3) Realizar un análisis discriminante cuadrático (QDA) incluyendo todas las variables y considerando probabilidades a priori iguales.

```
QDA <- qda(x = d[1:150, 1:4], grouping = d[1:150, 5],
           prior = c(1/3, 1/3, 1/3))
```

a. ¿Cómo se clasificaría a una flor con medidas z = (6, 3, 5, 2)? ¿Cuáles serían las probabilidades a posteriori de pertenencia a cada grupo? ¿Sería fiable la clasificación?

```
predict(QDA, z)
```

```
## $class
## [1] virginica
## Levels: setosa versicolor virginica
##
## $posterior
##              setosa   versicolor virginica
## [1,] 6.952495e-124 0.0003875667 0.9996124
```

Se clasifica en virgínica, y la clasificación sería fiable.

b. ¿Dónde se clasificaría z si las probabilidades a priori fueran 0.5, 0.25 y 0.25 para las especies *setosa, versicolor* y *virginica*, respectivamente? ¿Cuánto valdrán las probabilidades a posteriori? ¿Sería fiable la clasificación en este caso?

```
QDA <- qda(x = d[1:150, 1:4], grouping = d[1:150, 5],
           prior = c(0.5, 0.25, 0.25))
predict(QDA, z)
```

```
## $class
## [1] virginica
## Levels: setosa versicolor virginica
##
## $posterior
##              setosa   versicolor virginica
## [1,] 1.390499e-123 0.0003875667 0.9996124
```

Se clasifica en virgínica, y la clasificación sería fiable.

4) Estimar las probabilidades de clasificar correctamente a una flor desconocida usando LDA y QDA con validación cruzada y suponiendo las proporciones a priori iguales.

```
LDACV <- lda(x = d[1:150, 1:4], grouping = d[1:150, 5],
             prior = c(1/3, 1/3, 1/3), CV = TRUE)
QDACV <- qda(x = d[1:150, 1:4], grouping = d[1:150, 5],
             prior = c(1/3, 1/3, 1/3), CV = TRUE)
```

a. ¿Cuál es el porcentaje de clasificaciones correctas usando LDA y QDA?

```
table(LDACV$class, d$Species)
```

```
##
##              setosa versicolor virginica
##   setosa         50          0         0
##   versicolor      0         48         1
##   virginica       0          2        49
```

```
table(QDACV$class, d$Species)
```

```
##
##              setosa versicolor virginica
##   setosa         50          0         0
##   versicolor      0         47         1
##   virginica       0          3        49
```

Clasificaciones correctas para `LDA = 147/150` y para `QDACV = 146/150`.

b. ¿Cuál es el porcentaje de clasificaciones correctas usando LDA y QDA dentro de la especie *versicolor*?

```
48/50
```

```
## [1] 0.96
```

```
2/50
```

```
## [1] 0.04
```

c. ¿Cuál es el porcentaje de clasificaciones correctas usando LDA y QDA entre las flores clasificadas como *virginica*?

```
49/51
```

```
## [1] 0.9607843
```

```
49/52
```

```
## [1] 0.9423077
```

d. ¿Cuáles son las flores mal clasificadas? ¿En qué grupo se encuentran y dónde se clasifican?

```
index = which(LDACV$class != d[, 5])
index
```

```
## [1]  71  84 134
```

```
d[index, 5]
```

```
## [1] versicolor versicolor virginica
## Levels: setosa versicolor virginica
```

```
LDACV$class[index]
```

```
## [1] virginica  virginica  versicolor
## Levels: setosa versicolor virginica
```

5) Obtener las matrices de covarianzas y realizar los test de normalidad en cada grupo. Con los resultados obtenidos, ¿qué procedimiento sería más adecuado, LDA o QDA? Razonar las respuestas.

```
library("dplyr")
d1 = d %>%
  filter(Species == "setosa")
d2 = d %>%
  filter(Species == "versicolor")
d3 = d %>%
  filter(Species == "virginica")
cov(d1[, 1:4])
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length   0.12424898 0.099216327  0.016355102 0.010330612
## Sepal.Width    0.09921633 0.143689796  0.011697959 0.009297959
## Petal.Length   0.01635510 0.011697959  0.030159184 0.006069388
## Petal.Width    0.01033061 0.009297959  0.006069388 0.011106122
```

```
cov(d2[, 1:4])
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length   0.26643265  0.08518367   0.18289796  0.05577959
## Sepal.Width    0.08518367  0.09846939   0.08265306  0.04120408
## Petal.Length   0.18289796  0.08265306   0.22081633  0.07310204
## Petal.Width    0.05577959  0.04120408   0.07310204  0.03910612
```

```
cov(d3[, 1:4])
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length   0.40434286  0.09376327   0.30328980  0.04909388
## Sepal.Width    0.09376327  0.10400408   0.07137959  0.04762857
```

```
## Petal.Length   0.30328980  0.07137959   0.30458776  0.04882449
## Petal.Width    0.04909388  0.04762857   0.04882449  0.07543265
```

```
library("mvnormtest")
mshapiro.test(t(d1[, 1:4]))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Z
## W = 0.95878, p-value = 0.07906
```

```
mshapiro.test(t(d2[, 1:4]))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Z
## W = 0.93043, p-value = 0.005739
```

```
mshapiro.test(t(d3[, 1:4]))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Z
## W = 0.93414, p-value = 0.007955
```