Cangrejos de roca Leptograpsus

Jaqueline Girabel

El set de datos **cangrejos** corresponde a un estudio de los cangrejos de roca Leptograpsus, de Australia. Consiste de cinco variables que describen distintas medidas del caparazón, y hay cuatro grupos: sexo (macho/hembra) - color (naranja/azul), cada uno con 50 ejemplares. Dado que, cuando se encuentran restos de cangrejos queda sólo la caparazón, que además se decolora con el tiempo, lo que interesa entonces es poder deducir el sexo y color mediante los datos de caparazón.

```
##   labio_frontal ancho_posterior longitud_caparazon ancho_caparazon
## 1           9.1             6.9               16.7            18.6
## 2          10.2             8.2               20.2            22.2
## 3          10.7             8.6               20.7            22.7
## 4          11.4             9.0               22.7            24.8
## 5          12.5             9.4               23.2            26.0
## 6          12.5             9.4               24.2            27.0
##   profundidad_cuerpo genero   color
## 1                7.4  macho naranja
## 2                9.0  macho naranja
## 3                9.2  macho naranja
## 4               10.1  macho naranja
## 5               10.8  macho naranja
## 6               11.2  macho naranja
```

```
dim(cangrejos)
```

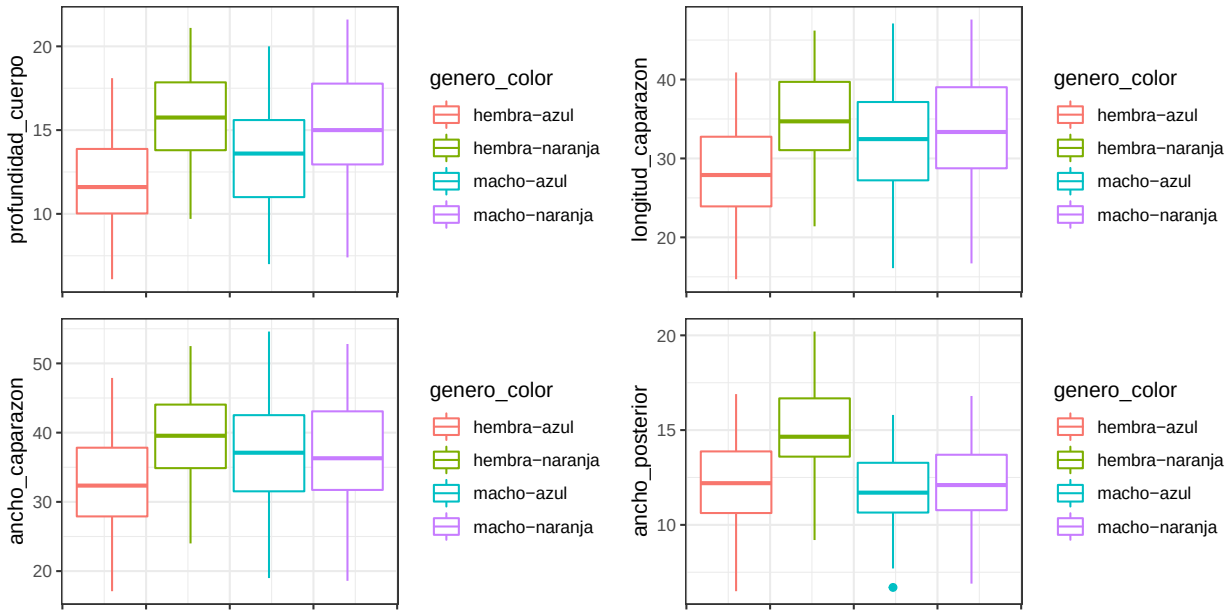
```
## [1] 200   7
```

Lo primero que hacemos es modificar el dataset pegando las últimas dos columnas, que corresponden a las variables **género** y **color** . De esta manera, obtenemos una nueva variable **género-color** en la que se identifican cuatro grupos distintos.

```
##   labio_frontal ancho_posterior longitud_caparazon ancho_caparazon
## 1           9.1             6.9               16.7            18.6
## 2          10.2             8.2               20.2            22.2
## 3          10.7             8.6               20.7            22.7
## 4          11.4             9.0               22.7            24.8
## 5          12.5             9.4               23.2            26.0
## 6          12.5             9.4               24.2            27.0
##   profundidad_cuerpo  genero_color
## 1                7.4 macho-naranja
## 2                9.0 macho-naranja
## 3                9.2 macho-naranja
## 4               10.1 macho-naranja
## 5               10.8 macho-naranja
## 6               11.2 macho-naranja
```

Por cada variable explicativa, observamos algunas características básicas de las ditribuciones por grupos. Cada uno de los gráficos que siguen muestra que los cuatro grupos parecen tener una distribución simétrica, y como los tamaños de las cajas son similares entre grupos, la varianza parece ser constante en ellos. Sí puede notarse una diferencia en los centros de cada grupo, pero no demasiado significativa. Por ejemplo, en la variable **profundidad_cuerpo** hay una diferencia entre las medias de dos poblaciones distintas, que son las que se diferencian por color.

```
library(ggplot2)
library(cowplot)

gg1<-ggplot(data = cangrejos, aes(y = profundidad_cuerpo, color = genero_color)) +
    geom_boxplot() +
    theme_bw()+theme(axis.text.x=element_blank())

gg2<-ggplot(data = cangrejos, aes(y = longitud_caparazon, color = genero_color)) +
    geom_boxplot() +
    theme_bw()+theme(axis.text.x=element_blank())

gg3<-ggplot(data = cangrejos, aes(y = ancho_caparazon, color = genero_color)) +
    geom_boxplot() +
    theme_bw()+theme(axis.text.x=element_blank())

gg4<-ggplot(data = cangrejos, aes(y = ancho_posterior, color = genero_color)) +
    geom_boxplot() +
    theme_bw()+theme(axis.text.x=element_blank())

plot_grid(gg1, gg2, gg3, gg4)
```



Separamos un 70% de los datos para el entrenamiento de los modelos, reservando el 30% restante para testeo. Se obtienen los respectivos datasets resultantes **cangrejos.T** y **cangrejos.V**.

```
set.seed(123)
indices<-sort(sample(1:dim(cangrejos)[1], round(dim(cangrejos)[1]*0.7, 0 ), replace = FALSE))
cangrejos.V<-cangrejos[-indices,]
cangrejos.T<-cangrejos[indices,]
```

Ajustamos un modelo usando análisis discriminante lineal. Como hay cuatro grupos para clasificar y cinco variables explicativas, son exactamente tres funciones discriminantes que separan a los cangrejos por su condición genero-color. Reciben los nombres LD1, LD2 y LD3.

```
library(MASS)

modelo.lda<- lda(genero_color ~ ., data = cangrejos.T)
modelo.lda$scaling
```

```
##                          LD1        LD2        LD3
## labio_frontal      -1.9034702 -0.3464898 -1.8034135
## ancho_posterior    -0.6344594 -1.8321687  0.4610657
## longitud_caparazon -0.2188736  0.9954186  0.5885658
## ancho_caparazon     1.6279120 -0.4691861 -0.6236399
## profundidad_cuerpo -1.2336663  0.6450779  1.5582165
```

Los coeficientes que se imprimieron son los que acompañan a cada variable por cada una de las funciones discriminantes.

Aplicamos este modelo al set de testing y calculamos el error de predicción:

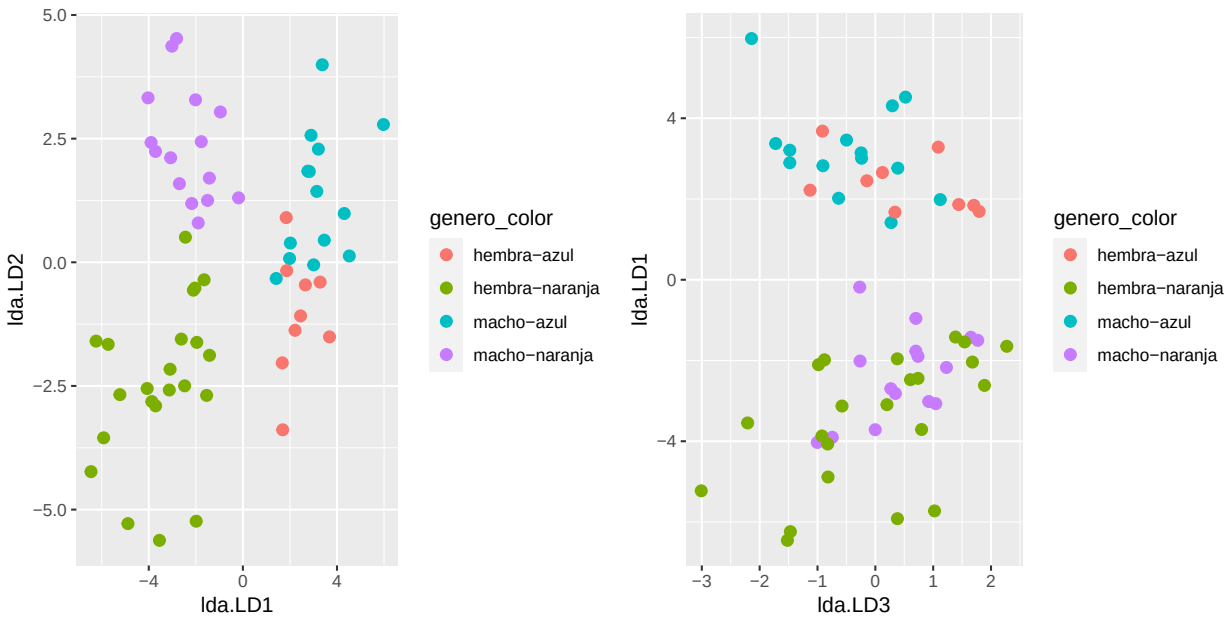
```
modelo.lda.predicciones<-predict(modelo.lda, cangrejos.V)
clasificador<-modelo.lda.predicciones$class
```

```
error<-round(mean(clasificador != cangrejos.V$genero_color),4)
error
```

```
## [1] 0.0833
```

Lo que podemos observar en los gráficos a continuación es la separación de las cuatro clases a partir de los resultados de este modelo lineal. La función LD1 separa muy bien por color pero no distingue por género, mientras que LD2 logra diferenciar grupos por género y no por color. Por otro lado, LD3 pretende distinguir también grupos por color, pero la separación no es clara.

```
library(ggplot2)
library(cowplot)


newdata <- data.frame(genero_color = cangrejos.V[,6], lda = modelo.lda.predicciones$x)
attach(newdata)

g1<-ggplot(newdata) + geom_point(aes(lda.LD1, lda.LD2, colour = genero_color), size = 2.5)
g2<-ggplot(newdata) + geom_point(aes(lda.LD3, lda.LD1, colour = genero_color), size = 2.5)

plot_grid(g1, g2)
```



Comparamos ahora los resultados del modelo lineal con la performance de un modelo cuadrático.

```
modelo.qda<- qda(genero_color ~ ., data = cangrejos.T)

modelo.qda.predicciones<-predict(modelo.qda, cangrejos.V)
clasificador.qda<-modelo.qda.predicciones$class

#error de predicción del modelo cuadrático
round(mean(clasificador.qda != cangrejos.V$genero_color),4)
```

```
## [1] 0.1
```

El modelo cuadrático es ligeramente menos preciso que el modelo lineal en el set de testing, pero ambos se ajustan bien a los datos.