Entrega de ejercicios Tema 4

Blanca Cano Camarero

25 de diciembre de 2022

Indice de contenidos

Ejercicio 5

El siguiente código carga en R los datos del fichero `lirios`:

```
 load(url('https://matematicas.uam.es/~joser.berrendero/datos/lirios.RData'))

# Cargamos los datos en un dataframe
data <- as.data.frame(lirios)
data$clases <- clases
colnames(data)
```

```
[1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"  "clases"
```

```
head(data)
```

```
   Sepal.Length Sepal.Width Petal.Length Petal.Width clases
51          7.0         3.2          4.7         1.4      0
52          6.4         3.2          4.5         1.5      0
53          6.9         3.1          4.9         1.5      0
54          5.5         2.3          4.0         1.3      0
55          6.5         2.8          4.6         1.5      0
56          5.7         2.8          4.5         1.3      0
```

Los datos corresponden a la longitud y anchura del pétalo y del sépalo de 100 lirios, 50 de ellos pertenecen a la especie versicolor y otros 50 a la especie virginica.

Apartado 1

Considera primero únicamente las dos variables correspondientes al sépalo.

Cálculo de los coeficientes de la función discriminante lineal de Fisher

```
library(MASS)
fisher_lineal <-lda(clases ~ Sepal.Length + Sepal.Width,
                    prior = c(0.5, 0.5), data = data)
fisher_lineal
```

```
Call:
lda(clases ~ Sepal.Length + Sepal.Width, data = data, prior = c(0.5,
    0.5))

Prior probabilities of groups:
  0   1
0.5 0.5

Group means:
  Sepal.Length Sepal.Width
0        5.936       2.770
1        6.588       2.974

Coefficients of linear discriminants:
                   LD1
Sepal.Length 1.6271842
Sepal.Width  0.3435524
```

Podemos observar que los coeficientes obtenidos han sido

```
Coefficients of linear discriminants:
                   LD1
Sepal.Length 1.6271842
Sepal.Width  0.3435524
```

Esto nos da una indicación de la relevancia de la etiqueta en cuanto a clasificar, a mayor en valor absoluto mayor importancia.

Estimación la probabilidad de error de esta regla mediante el riesgo empírico

La función de riesgo empírico se define como

$$\hat{L}_n = \frac{1}{n} \sum_{i=1}^{n} I_{g(x_i) \neq y_i}$$

Procedamos a calcularla

```r
predictions <- predict(fisher_lineal)$class
table(data$clases, predictions)
```

```
   predictions
     0  1
  0 38 12
  1 13 37
```

```r
riesgo_empirico <-mean(data$clases !=  predictions)
cat('El riesgo empírico es de ', riesgo_empirico)
```

```
El riesgo empírico es de  0.25
```

Tasa de error por validación cruzada

```r
predicciones.lda.cv <- lda(clases ~ Sepal.Length + Sepal.Width,
                           prior = c(0.5, 0.5),
                           data = data, CV=TRUE
                           )$class
tasa_error_cv <- mean(data$clases != predicciones.lda.cv)
cat("La tasa de error en validación cruzada es de ", tasa_error_cv)
```

```
La tasa de error en validación cruzada es de  0.27
```

Notemos que la tasa de error de validación cruzada es mayor, esto es natural ya que no estamos entrenando con todos los datos.

Compara los valores de estos estimadores con el estimador paramétrico basado en el resultado del ejercicio 4

$$1 - \phi(\hat{\Delta}/2)$$

donde

$$\hat{\Delta}^2 = (\hat{\mu}_0 - \hat{\mu}_1)'\hat{\Sigma}^{-1}(\hat{\mu}_0 - \hat{\mu}_1).$$

Solución

Los vectores de medias vendrán dados por:

```
Group means:
  Sepal.Length Sepal.Width
0        5.936       2.770
1        6.588       2.974
```

```r
mu_0 <- c( 5.936, 2.770)
mu_1 <- c( 6.588,  2.974)
diference <- mu_0 - mu_1
```

Estimaremos la matriz de covarianzas suponiendo caso homocedástico:

$$S = \frac{n_0 - 1}{n_0 + n_1 - 2} S_0 + \frac{n_1 - 1}{n_0 + n_1 - 2} S_1.$$

```r
library(matlib)

data_0 <- data[data$clases == 0,c("Sepal.Length", "Sepal.Width")]
n_0 <- nrow(data_0)
S_0 <- cov(data_0)

data_1 <- data[data$clases == 1,c("Sepal.Length", "Sepal.Width")]
n_1 <- nrow(data_1)
S_1 <- cov(data_1)

S <- (
  (n_0 - 1)*S_0
  +
  (n_1 - 1)*S_1
)/(n_0 + n_1 - 2)


S_inv <- inv(S)
delta_2 <- diference %*%  S_inv %*% matrix(diference)
delta <- sqrt(delta_2[1,1])

error <- 1 - pnorm(delta/2)

cat("EL error Bayes es ",error)
```

```
EL error Bayes es  0.2858654
```

A la vista de los resultados puede un pensar que se está violando la propocisión vista en clase de que **la regla de Bayes es la regla óptima**, ya que los errores obtenidos son inferiores al error de Bayes.

Sin embargo esta contradicción proviene de la imprecisión introducida en la estimación de los parámetros, ya sea en la matriz de covarianza como en las medias.

Además siempre hay que tener presente que estamos tratando con un conjunto de muestra (y en este caso relativamente pequeño) no con la población en su totalidad.

Repetición del apartado anterior pero considerando las cuatro variables

Coeficiente de la función discriminante lineal de Fisher

```
fisher_lineal <-lda(clases ~ Sepal.Length + Sepal.Width +  Petal.Length
+ Petal.Width
              ,
  prior = c(0.5, 0.5), data = data)
fisher_lineal
```

```
Call:
lda(clases ~ Sepal.Length + Sepal.Width + Petal.Length + Petal.Width,
    data = data, prior = c(0.5, 0.5))

Prior probabilities of groups:
  0   1
0.5 0.5

Group means:
  Sepal.Length Sepal.Width Petal.Length Petal.Width
0        5.936       2.770        4.260       1.326
1        6.588       2.974        5.552       2.026

Coefficients of linear discriminants:
                    LD1
Sepal.Length -0.9431178
Sepal.Width  -1.4794287
Petal.Length  1.8484510
Petal.Width   3.2847304
```

Estimación la probabilidad de error de esta regla mediante el riesgo empírico

```
predictions <- predict(fisher_lineal)$class
table(data$clases, predictions)
```

```
   predictions
     0  1
  0 48  2
  1  1 49
```

```
riesgo_empirico <-mean(data$clases !=  predictions)
cat('El riesgo empírico es de ', riesgo_empirico)
```

```
El riesgo empírico es de  0.03
```

Tasa de error por validación cruzada

```
predicciones.lda.cv <- lda(clases ~ Sepal.Length + Sepal.Width +  Petal.Length
+ Petal.Width
, prior = c(0.5, 0.5), data = data, CV=TRUE)$class
tasa_error_cv <- mean(data$clases != predicciones.lda.cv)
cat("La tasa de error en validación cruzada es de ", tasa_error_cv)
```

```
La tasa de error en validación cruzada es de  0.03
```

Estimador paramétrico

```
fisher_lineal
```

```
Call:
lda(clases ~ Sepal.Length + Sepal.Width + Petal.Length + Petal.Width,
    data = data, prior = c(0.5, 0.5))

Prior probabilities of groups:
  0   1
0.5 0.5

Group means:
  Sepal.Length Sepal.Width Petal.Length Petal.Width
0        5.936       2.770        4.260       1.326
1        6.588       2.974        5.552       2.026
```

```
Coefficients of linear discriminants:
                    LD1
Sepal.Length -0.9431178
Sepal.Width  -1.4794287
Petal.Length  1.8484510
Petal.Width   3.2847304
```

```r
mu_0 <- c( 5.936, 2.770, 4.260, 1.326)
mu_1 <- c(6.588, 2.974, 5.552, 2.026)
diference <- mu_0 - mu_1

columns <- c("Sepal.Length", "Sepal.Width", "Petal.Length","Petal.Width")
data_0 <- data[data$clases == 0, columns]
n_0 <- nrow(data_0)
S_0 <- cov(data_0)

data_1 <- data[data$clases == 1,columns]
n_1 <- nrow(data_1)
S_1 <- cov(data_1)

S <- (
  (n_0 - 1)*S_0
  +
  (n_1 - 1)*S_1
)/(n_0 + n_1 - 2)

#equivale a: delta <- mahalanobis(mu_0, mu_1, S)
S_inv <- inv(S)
delta_2 <- diference %*%  S_inv %*% matrix(diference)
delta <- sqrt(delta_2[1,1])

error <- 1 - pnorm(delta/2)

cat("EL error Bayes es ",error)
```

```
EL error Bayes es  0.02968814
```