Análisis Discriminante Lineal

Karina Itzel Rodríguez Conde

2022-05-12

Matriz de datos

La base de datos utilizada es el de **anorexia**, la cual se encuentra precargada en r, en la librería *MASS* y contiene los datos del cambio de peso de pacientes mujeres jóvenes con anorexia, así como las siguientes columnas:

1.- Factor de tratamiento de tres niveles: "Cont" (control), "CBT" (tratamiento cognitivo conductual) y "FT" (tratamiento familiar).

2.- **Prewt** Peso del paciente antes del período de estudio, en lbs.

3.- **Postwt** peso del paciente después del período de estudio, en lbs.

1.- Paquetería y librería a utilizar

```
install.packages("MASS")
library(MASS)
```

2.- Se cargan los datos de anorexia

```
Z <- as.data.frame(anorexia)
```

Exploración de la matriz

1.- Dimensión

```
dim(Z)
```

```
## [1] 72  3
```

La base de datos cuenta con 72 observaciones y 3 variables.

2.- Nombre de las variables

```
colnames(Z)
```

```
## [1] "Treat"  "Prewt"  "Postwt"
```

3.- Tipo de variables

```
str(Z)
```

```
## 'data.frame':    72 obs. of  3 variables:
##  $ Treat : Factor w/ 3 levels "CBT","Cont","FT": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Prewt : num  80.7 89.4 91.8 74 78.1 88.3 87.3 75.1 80.6 78.4 ...
##  $ Postwt: num  80.2 80.1 86.4 86.3 76.1 78.1 75.1 86.7 73.5 84.6 ...
```

4.- Saber si existen datos nulos

```
anyNA(Z)
```

```
## [1] FALSE
```

Esta base de datos no contiene datos nulos.

Tratamiento de la matriz

1.- Se define la matriz de datos y la variable respuesta con las clasificaciones.

```
x<-Z[,2:3]
y<-Z[,1]
```

2.- Definir como n y p el número de pacientes y variables

```
n<-nrow(x)
p<-ncol(x)
```

Se aplica el Análisis discriminante lineal (LDA)

1.- Cross validation (cv): clasificación óptima

```
lda.anorexia<-lda(y~.,data=x, CV=TRUE)
```

2.- lda.anorexia$class contiene las clasificaciones hechas por CV usando LDA.

```
lda.anorexia$class
```

```
##  [1] Cont Cont CBT  CBT  Cont Cont Cont CBT  Cont CBT  Cont Cont CBT  Cont CBT
## [16] Cont CBT  Cont Cont Cont CBT  CBT  Cont Cont CBT  Cont Cont CBT  Cont Cont
## [31] Cont FT   FT   FT   Cont Cont FT   FT   CBT  Cont FT   Cont CBT  Cont CBT
## [46] CBT  CBT  Cont CBT  FT   CBT  CBT  CBT  FT   CBT  FT   FT   CBT  CBT  FT
## [61] Cont Cont FT   FT   Cont Cont FT   CBT  CBT  CBT  CBT  FT
## Levels: CBT Cont FT
```

3.- Creación de la tabla de clasificaciones buenas y malas

```
table.lda<-table(y,lda.anorexia$class)
table.lda
```

```
##
## y      CBT Cont FT
##   CBT   11   10  8
##   Cont  10   16  0
##   FT     6    4  7
```

4.- Proporción de errores

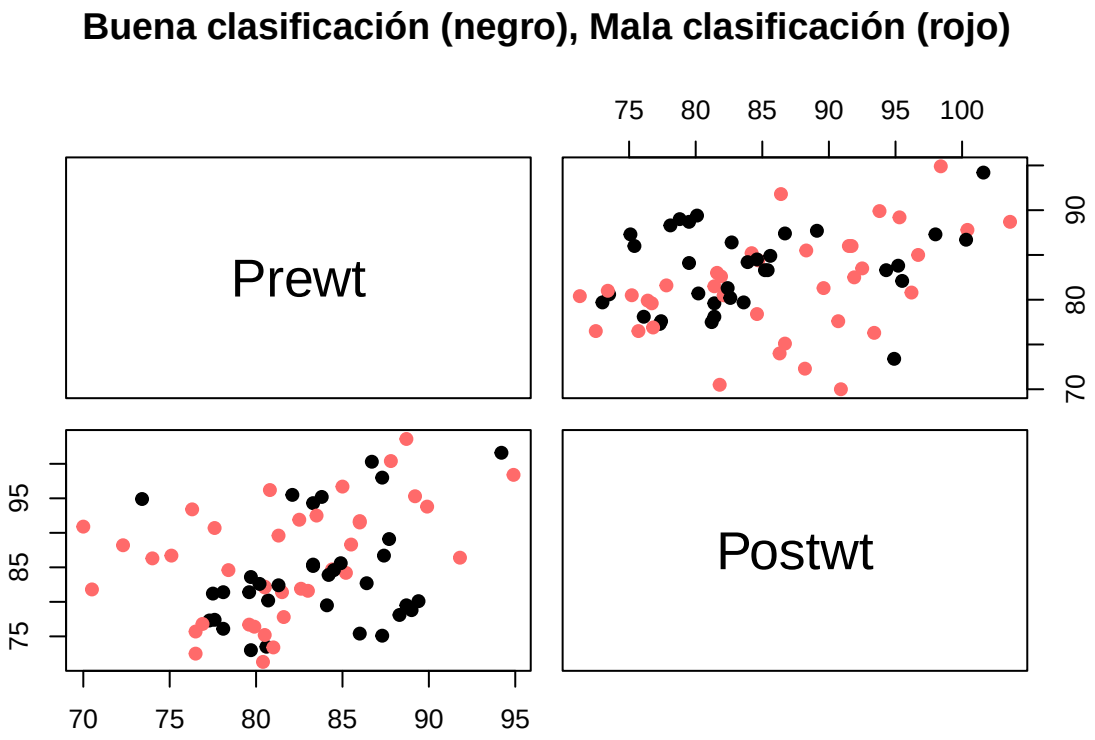
```
mis.lda<- n-sum(y==lda.anorexia$class)
mis.lda/n
```

```
## [1] 0.5277778
```

Por cada 100 pacientes que se clasifique se equivocará 52 veces.

5.- Scater plot de Buenas clasificaciones en negro y Malas en rojo

```
col.lda.anorexia<-c("indianred1","black")[1*(y==lda.anorexia$class)+1]
pairs(x,main="Buena clasificación (negro), Mala clasificación (rojo)",
      pch=19,col=col.lda.anorexia)
```



6.- Probabilidad de pertenencia a uno de los tres grupos

```
lda.anorexia$posterior
```

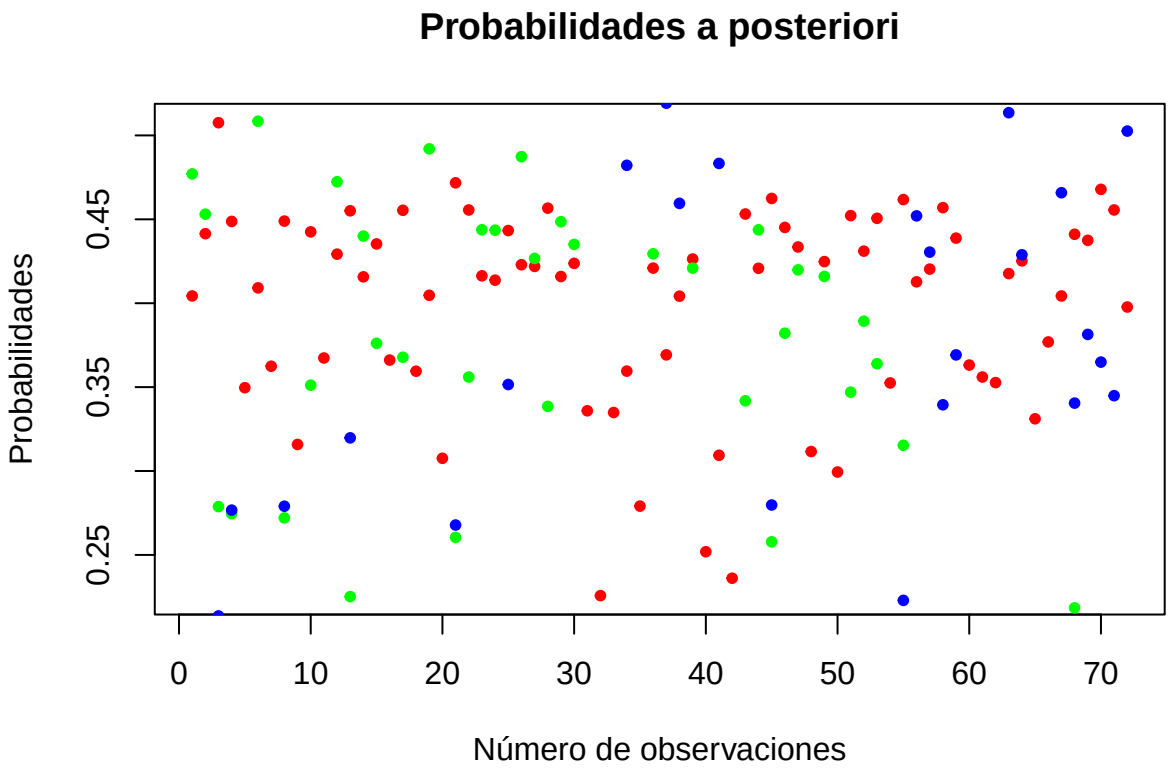
```
##          CBT       Cont         FT
## 1  0.4043290 0.47705482 0.11861619
## 2  0.4414188 0.45305554 0.10552561
## 3  0.5075629 0.27878409 0.21365306
## 4  0.4487312 0.27457551 0.27669333
## 5  0.3496473 0.57713852 0.07321413
## 6  0.4091361 0.50846418 0.08239970
## 7  0.3624352 0.58235879 0.05520601
## 8  0.4489345 0.27204601 0.27901949
## 9  0.3158225 0.63571480 0.04846271
## 10 0.4425225 0.35114693 0.20633062
## 11 0.3673504 0.54452728 0.08812229
```

```
## 12 0.4292390 0.47236962 0.09839133
## 13 0.4550708 0.22517166 0.31975750
## 14 0.4157067 0.43999786 0.14429539
## 15 0.4353164 0.37609179 0.18859180
## 16 0.3661119 0.54636933 0.08751875
## 17 0.4553781 0.36779684 0.17682504
## 18 0.3594913 0.58194637 0.05856229
## 19 0.4047231 0.49197489 0.10330205
## 20 0.3075810 0.64665500 0.04576398
## 21 0.4717416 0.26045337 0.26780506
## 22 0.4555722 0.35600817 0.18841967
## 23 0.4163213 0.44375673 0.13992193
## 24 0.4137215 0.44350623 0.14277226
## 25 0.4433088 0.20514332 0.35154787
## 26 0.4229465 0.48733222 0.08972129
## 27 0.4219465 0.42670032 0.15135323
## 28 0.4566604 0.33851988 0.20481975
## 29 0.4159020 0.44855050 0.13554750
## 30 0.4237582 0.43503849 0.14120328
## 31 0.3358862 0.59142408 0.07268974
## 32 0.2257488 0.03843641 0.73581478
## 33 0.3348680 0.09127334 0.57385864
## 34 0.3595406 0.15826462 0.48219478
## 35 0.2790934 0.67577851 0.04512810
## 36 0.4209267 0.42945547 0.14961784
## 37 0.3692137 0.11159118 0.51919511
## 38 0.4041943 0.13631759 0.45948809
## 39 0.4263332 0.42089174 0.15277504
## 40 0.2518826 0.70620103 0.04191634
## 41 0.3093705 0.20732203 0.48330752
## 42 0.2361239 0.73189924 0.03197681
## 43 0.4531472 0.34184031 0.20501250
## 44 0.4207942 0.44373833 0.13546742
## 45 0.4624380 0.25781486 0.27974713
## 46 0.4451559 0.38213913 0.17270494
## 47 0.4335038 0.41987474 0.14662145
## 48 0.3116036 0.61900998 0.06938645
## 49 0.4247919 0.41594127 0.15926680
## 50 0.2994030 0.06667817 0.63391884
## 51 0.4521984 0.34698216 0.20081947
## 52 0.4310308 0.38928000 0.17968920
## 53 0.4505897 0.36395519 0.18545506
## 54 0.3524833 0.11504419 0.53247249
## 55 0.4617649 0.31532231 0.22291277
## 56 0.4127439 0.13521328 0.45204281
## 57 0.4203198 0.14923533 0.43044485
## 58 0.4569578 0.20357053 0.33947172
## 59 0.4387823 0.19202112 0.36919663
## 60 0.3631442 0.07377692 0.56307893
## 61 0.3560293 0.58695575 0.05701495
## 62 0.3526889 0.58883679 0.05847431
## 63 0.4176513 0.06880023 0.51354852
## 64 0.4252170 0.14592082 0.42886218
## 65 0.3311223 0.62736690 0.04151080
```

```
## 66 0.3769130 0.55580200 0.06728503
## 67 0.4043237 0.12983390 0.46584237
## 68 0.4410736 0.21845638 0.34047002
## 69 0.4374726 0.18113118 0.38139619
## 70 0.4678569 0.16724076 0.36490236
## 71 0.4555694 0.19953689 0.34489368
## 72 0.3977378 0.09968493 0.50257723
```

7.- Gráfico de probabilidades

```
plot(1:n, lda.anorexia$posterior[,1],
     main="Probabilidades a posteriori",
     pch=20, col="red",
     xlab="Número de observaciones", ylab="Probabilidades")
points(1:n,lda.anorexia$posterior[,2],
       pch=20, col="green")
points(1:n,lda.anorexia$posterior[,3],
       pch=20, col="blue")
```



El gráfico anterior muestra aquellas observaciones que están bien y mal clasificadas con base a su probabilidad. Se observa que hay muchas variables cuya probabilidad va de 0.25 a 0.45, recordando que el error de clasificación es de 0.52, siendo una probabilidad muy alta y habiendo un total de 38 observaciones con mala clasificación, era de esperar un comportamiento de esa manera.