Análisis discriminante lineal (LDA)

Omar Sanchez Hernandez

4/6/2022

Introducción

El análisis discriminante lineal es un tipo de analisis que construye una función a partir de datos numericos continuos para poder calcular la probabilidad de que un individuo pertenezca a un grupo (que vendria a ser la variable respuesta) para poder clasificarlo.

Para desarrollar el siguiente ejemplo se hara uso de la base de datos de *iris*

Cargamos la matriz de datos de *iris*

```
library(MASS)
library(knitr)
Z<-as.data.frame(iris)
kable(head(Z))
```

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa

Acerca de la matriz de datos

```
dim(Z)
```

```
## [1] 150   5
```

```
str(Z)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

La matriz de datos contiene 150 observaciones y 5 variables de las cuales 4 son numericas continuas y una es categorica.

Se define la matriz de datos y la variable respuesta con las categorías.

```
x<-Z[,1:4]
y<-Z[,5]
```

Definir como n como el numero de sujetos (flores) que participan y p como el numero de variables presentes.

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

Se aplica el Análisis discriminante lineal (LDA) con Validación cruzada (cv): clasificación optima.

```
lda.iris<-lda(Z$Species~.,data=Z,CV=TRUE)
```

El objeto lda.iris contiene las clasificaciones hechas por *Validación Cruzada* usando el discriminante lineal.

```
lda.iris$class
```

```
##   [1] setosa     setosa     setosa     setosa     setosa     setosa
##   [7] setosa     setosa     setosa     setosa     setosa     setosa
##  [13] setosa     setosa     setosa     setosa     setosa     setosa
##  [19] setosa     setosa     setosa     setosa     setosa     setosa
##  [25] setosa     setosa     setosa     setosa     setosa     setosa
##  [31] setosa     setosa     setosa     setosa     setosa     setosa
##  [37] setosa     setosa     setosa     setosa     setosa     setosa
##  [43] setosa     setosa     setosa     setosa     setosa     setosa
```

```
##  [49] setosa     setosa     versicolor versicolor versicolor versicolor
##  [55] versicolor versicolor versicolor versicolor versicolor versicolor
##  [61] versicolor versicolor versicolor versicolor versicolor versicolor
##  [67] versicolor versicolor versicolor versicolor virginica  versicolor
##  [73] versicolor versicolor versicolor versicolor versicolor versicolor
##  [79] versicolor versicolor versicolor versicolor versicolor virginica
##  [85] versicolor versicolor versicolor versicolor versicolor versicolor
##  [91] versicolor versicolor versicolor versicolor versicolor versicolor
##  [97] versicolor versicolor versicolor versicolor virginica  virginica
## [103] virginica  virginica  virginica  virginica  virginica  virginica
## [109] virginica  virginica  virginica  virginica  virginica  virginica
## [115] virginica  virginica  virginica  virginica  virginica  virginica
## [121] virginica  virginica  virginica  virginica  virginica  virginica
## [127] virginica  virginica  virginica  virginica  virginica  virginica
## [133] virginica  versicolor virginica  virginica  virginica  virginica
## [139] virginica  virginica  virginica  virginica  virginica  virginica
## [145] virginica  virginica  virginica  virginica  virginica  virginica
## Levels: setosa versicolor virginica
```

Se crea la matriz de confusion para evaluar que tan bien logra clasificar a los sujetos.

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

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

Se observa en la matriz de confusión que el modelo se equivoca clasificando a 3 sujetos.

Proporción de errores

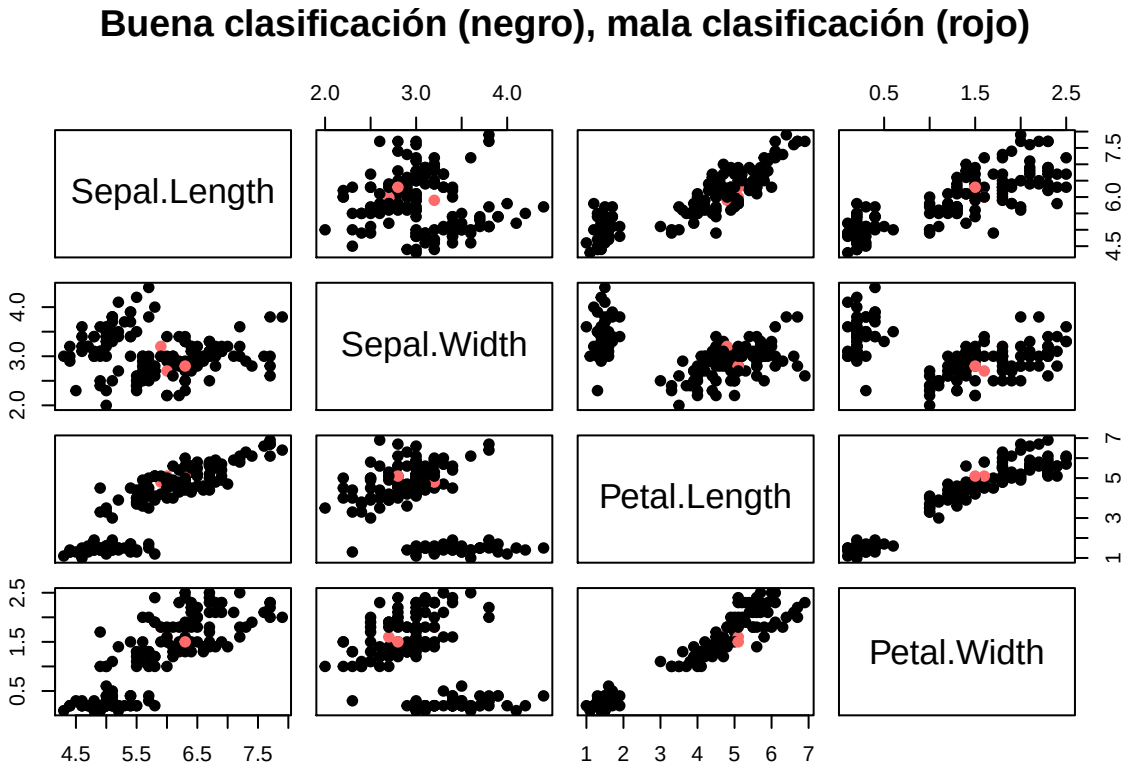
```
mis.lda<- n-sum(y==lda.iris$class)
(mis.lda/n)*100
```

```
## [1] 2
```

Tras calcular la proporción de errores se encuentra que el modelo se equivoca discriminando un 2% de los casos totales.

Scatterplot para localizar a los sujetos mal clasificados

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



En el gráfico se puede observar como hay pocos valores de color rojo que son los que se clasifican mal, sin embargo hay relaciones entre variables que se pueden observar muy bajas e incluso nulas como lo es en el caso de el largo del sepalo con el ancho del mismo, o el acho del sepalo con el largo del pétalo.

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

```
round(lda.iris$posterior,3)
```

```
##     setosa versicolor virginica
## 1        1      0.000     0.000
## 2        1      0.000     0.000
## 3        1      0.000     0.000
## 4        1      0.000     0.000
## 5        1      0.000     0.000
## 6        1      0.000     0.000
## 7        1      0.000     0.000
## 8        1      0.000     0.000
## 9        1      0.000     0.000
## 10       1      0.000     0.000
## 11       1      0.000     0.000
## 12       1      0.000     0.000
## 13       1      0.000     0.000
```

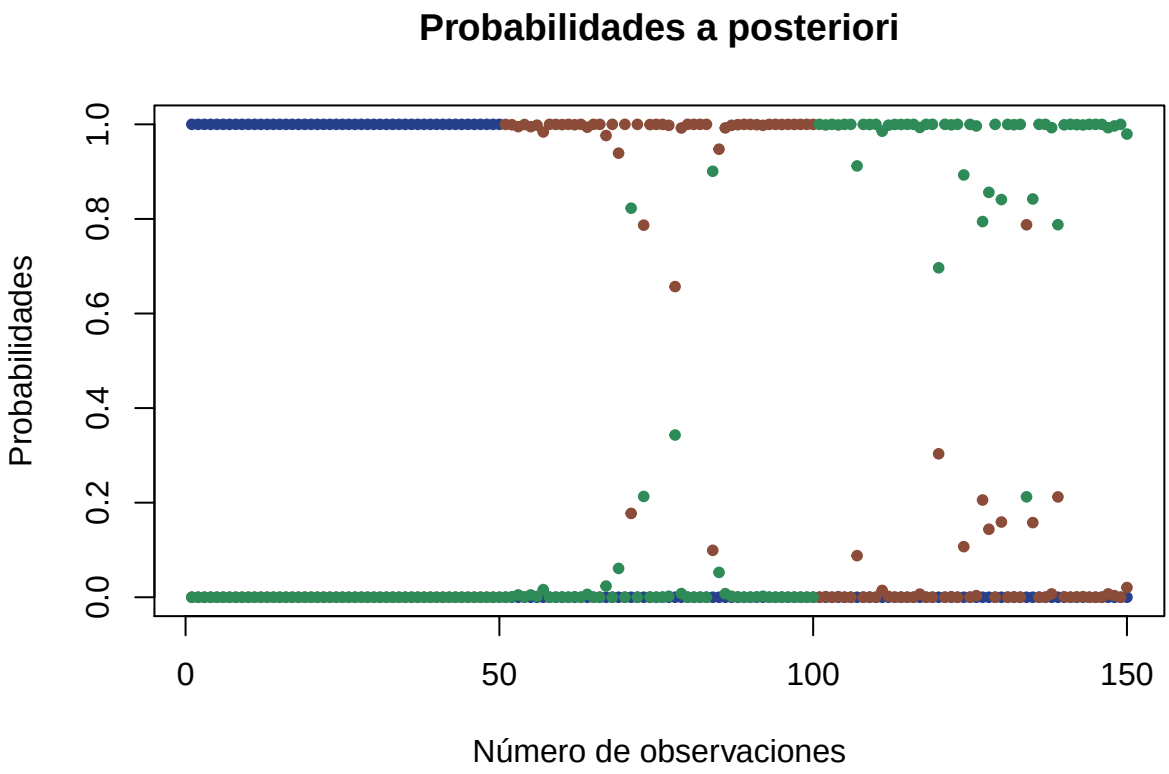
```
## 14       1      0.000     0.000
## 15       1      0.000     0.000
## 16       1      0.000     0.000
## 17       1      0.000     0.000
## 18       1      0.000     0.000
## 19       1      0.000     0.000
## 20       1      0.000     0.000
## 21       1      0.000     0.000
## 22       1      0.000     0.000
## 23       1      0.000     0.000
## 24       1      0.000     0.000
## 25       1      0.000     0.000
## 26       1      0.000     0.000
## 27       1      0.000     0.000
## 28       1      0.000     0.000
## 29       1      0.000     0.000
## 30       1      0.000     0.000
## 31       1      0.000     0.000
## 32       1      0.000     0.000
## 33       1      0.000     0.000
## 34       1      0.000     0.000
## 35       1      0.000     0.000
## 36       1      0.000     0.000
## 37       1      0.000     0.000
## 38       1      0.000     0.000
## 39       1      0.000     0.000
## 40       1      0.000     0.000
## 41       1      0.000     0.000
## 42       1      0.000     0.000
## 43       1      0.000     0.000
## 44       1      0.000     0.000
## 45       1      0.000     0.000
## 46       1      0.000     0.000
## 47       1      0.000     0.000
## 48       1      0.000     0.000
## 49       1      0.000     0.000
## 50       1      0.000     0.000
## 51       0      1.000     0.000
## 52       0      0.999     0.001
## 53       0      0.995     0.005
## 54       0      1.000     0.000
## 55       0      0.995     0.005
## 56       0      0.998     0.002
## 57       0      0.984     0.016
## 58       0      1.000     0.000
## 59       0      1.000     0.000
## 60       0      0.999     0.001
## 61       0      1.000     0.000
## 62       0      0.999     0.001
## 63       0      1.000     0.000
## 64       0      0.994     0.006
## 65       0      1.000     0.000
## 66       0      1.000     0.000
## 67       0      0.976     0.024
```

```
## 68       0      1.000     0.000
## 69       0      0.939     0.061
## 70       0      1.000     0.000
## 71       0      0.177     0.823
## 72       0      1.000     0.000
## 73       0      0.787     0.213
## 74       0      1.000     0.000
## 75       0      1.000     0.000
## 76       0      1.000     0.000
## 77       0      0.998     0.002
## 78       0      0.657     0.343
## 79       0      0.992     0.008
## 80       0      1.000     0.000
## 81       0      1.000     0.000
## 82       0      1.000     0.000
## 83       0      1.000     0.000
## 84       0      0.099     0.901
## 85       0      0.947     0.053
## 86       0      0.992     0.008
## 87       0      0.998     0.002
## 88       0      0.999     0.001
## 89       0      1.000     0.000
## 90       0      1.000     0.000
## 91       0      0.999     0.001
## 92       0      0.998     0.002
## 93       0      1.000     0.000
## 94       0      1.000     0.000
## 95       0      1.000     0.000
## 96       0      1.000     0.000
## 97       0      1.000     0.000
## 98       0      1.000     0.000
## 99       0      1.000     0.000
## 100      0      1.000     0.000
## 101      0      0.000     1.000
## 102      0      0.001     0.999
## 103      0      0.000     1.000
## 104      0      0.001     0.999
## 105      0      0.000     1.000
## 106      0      0.000     1.000
## 107      0      0.088     0.912
## 108      0      0.000     1.000
## 109      0      0.000     1.000
## 110      0      0.000     1.000
## 111      0      0.014     0.986
## 112      0      0.002     0.998
## 113      0      0.000     1.000
## 114      0      0.000     1.000
## 115      0      0.000     1.000
## 116      0      0.000     1.000
## 117      0      0.007     0.993
## 118      0      0.000     1.000
## 119      0      0.000     1.000
## 120      0      0.303     0.697
## 121      0      0.000     1.000
```

```
## 122      0      0.001     0.999
## 123      0      0.000     1.000
## 124      0      0.107     0.893
## 125      0      0.000     1.000
## 126      0      0.003     0.997
## 127      0      0.206     0.794
## 128      0      0.144     0.856
## 129      0      0.000     1.000
## 130      0      0.159     0.841
## 131      0      0.000     1.000
## 132      0      0.001     0.999
## 133      0      0.000     1.000
## 134      0      0.788     0.212
## 135      0      0.158     0.842
## 136      0      0.000     1.000
## 137      0      0.000     1.000
## 138      0      0.007     0.993
## 139      0      0.212     0.788
## 140      0      0.001     0.999
## 141      0      0.000     1.000
## 142      0      0.001     0.999
## 143      0      0.001     0.999
## 144      0      0.000     1.000
## 145      0      0.000     1.000
## 146      0      0.000     1.000
## 147      0      0.007     0.993
## 148      0      0.003     0.997
## 149      0      0.000     1.000
## 150      0      0.021     0.979
```

10.- Gráfico de probabilidades

```
plot(1:n, lda.iris$posterior[,1],
     main="Probabilidades a posteriori",
     pch=20, col="#27408B",
     xlab="Número de observaciones", ylab="Probabilidades")
points(1:n,lda.iris$posterior[,2],
       pch=20, col="#8B4C39")
points(1:n,lda.iris$posterior[,3],
       pch=20, col="#2E8B57")
```



En el gráfico de dispersion de las probabilidades a priori y posteriori se puede observar como existen comportamientos de traslape entre las observaciones de las especies de **Virginica** y **Versicolor**.