MAT-269: Sesión 21, Análisis Discriminante II

Felipe Osorio

fosorios.mat.utfsm.cl

Departamento de Matemática, UTFSM

EX UMBRA IN SOLEM

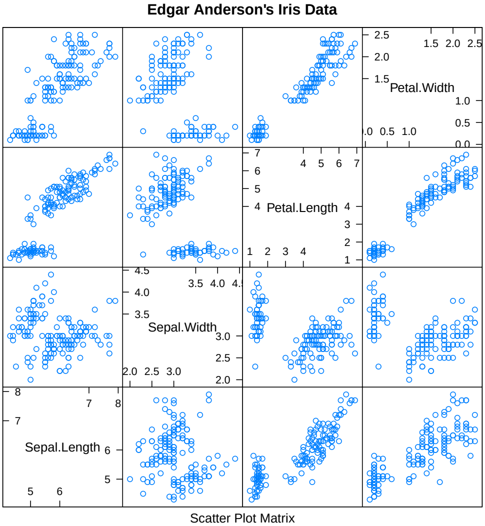
Datos de Flores Iris (Anderson, 1935; Fisher, 1936)

Iris setosa

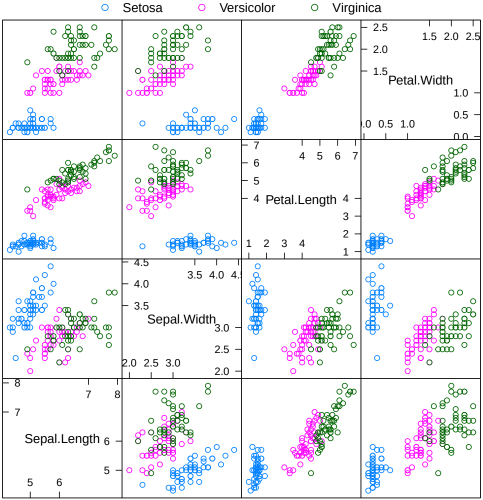
Iris versicolor

Iris virginica

Datos de Flores Iris (Anderson, 1935; Fisher, 1936)

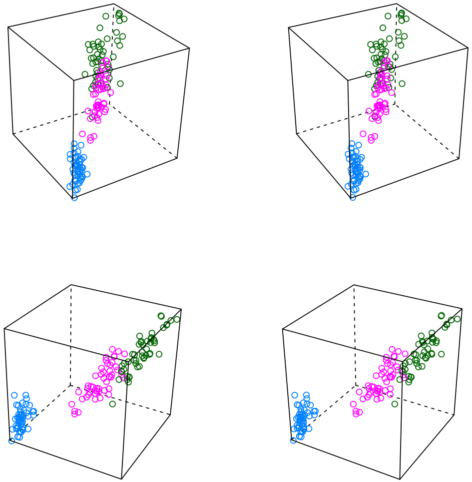


Datos de Flores Iris (Anderson, 1935; Fisher, 1936)

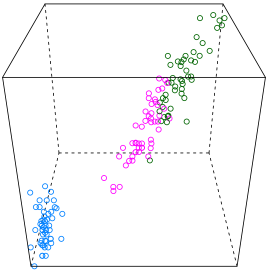


Scatter Plot Matrix

Datos de Flores Iris (Anderson, 1935; Fisher, 1936)



Datos de Flores Iris



Datos de Flores Iris (Anderson, 1935; Fisher, 1936)

Datos observados:

Mediciones (cm) del largo y ancho de los sépalos y el largo y ancho de pétalos para 50 flores desde 3 especies de Iris (setosa, virginica y versicolor).

Objetivo:

- Obtener una función que permita discriminar entre especies.
- Usando las medidas de una flor, clasificarla apropiadamente.

Características del problema:

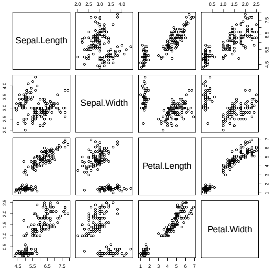
- El análisis exploratorio revela una separación evidente en 2 grupos.
- Técnicas más refinadas permiten identificar las 3 especies, p.ej.:
  - Análisis discriminante,
  - Técnicas de clasificación (Reconocimiento de patrones),
  - Aprendizaje de máquina (Máquinas de soporte vectorial, Data mining).

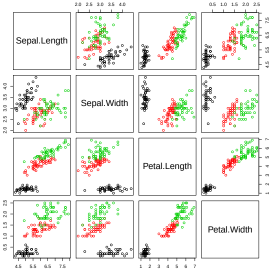
Conjunto de datos

```
> iris
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1            5.1         3.5          1.4         0.2     setosa
2            4.9         3.0          1.4         0.2     setosa
3            4.7         3.2          1.3         0.2     setosa
4            4.6         3.1          1.5         0.2     setosa
5            5.0         3.6          1.4         0.2     setosa
6            5.4         3.9          1.7         0.4     setosa
7            4.6         3.4          1.4         0.3     setosa
8            5.0         3.4          1.5         0.2     setosa
9            4.4         2.9          1.4         0.2     setosa
10           4.9         3.1          1.5         0.1     setosa
11           5.4         3.7          1.5         0.2     setosa
12           4.8         3.4          1.6         0.2     setosa
13           4.8         3.0          1.4         0.1     setosa

...

148          6.5         3.0          5.2         2.0  virginica
149          6.2         3.4          5.4         2.3  virginica
150          5.9         3.0          5.1         1.8  virginica
```

Gráfico del conjunto de datos

```
x <- iris[,1:4]
pairs(x)                         # 1er panel
pairs(x, col = iris$Species) # colores representando 'especies'
```





```
> library(MASS)
> zLDA <- lda(Species ~ ., data = iris)

> zLDA
Call:
lda(Species ~ ., data = iris)

Prior probabilities of groups:
    setosa versicolor  virginica
 0.3333333  0.3333333  0.3333333

Group means:
           Sepal.Length Sepal.Width Petal.Length Petal.Width
setosa            5.006       3.428        1.462       0.246
versicolor        5.936       2.770        4.260       1.326
virginica         6.588       2.974        5.552       2.026

Coefficients of linear discriminants:
                    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

Proportion of trace:
   LD1    LD2
0.9912 0.0088
```

```
> attributes(zLDA)
$names
 [1] "prior"   "counts"  "means"   "scaling" "lev"     "svd"     "N"
 [8] "call"    "terms"   "xlevels"

$class
[1] "lda"

# proporcion de varianza explicada
> prop <- zLDA$svd^2 / sum(zLDA$svd^2)
> prop
[1] 0.991212605 0.008787395
```

Es decir, para los conjuntos de datos de Iris, el 99.12% de la varianza entre-grupos es explicada por la primera función discriminante (lineal).

Note que, podemos escribir:

$$L_1 = 0.8294\,\text{Sepal.Length} + 1.5345\,\text{Sepal.Width} - 2.2012\,\text{Petal.Length} - 2.8105\,\text{Petal.Width}$$

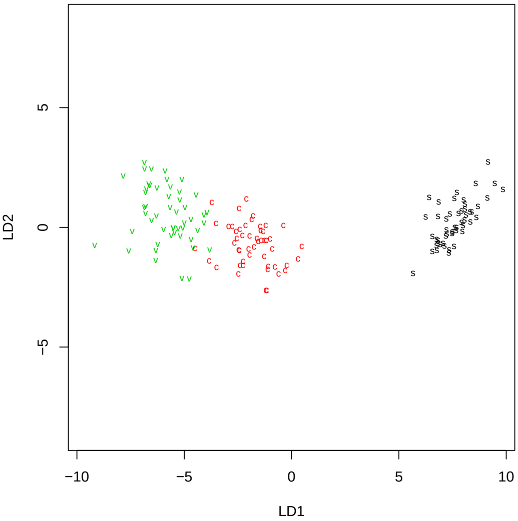
$$L_2 = 0.0241\,\text{Sepal.Length} + 2.1645\,\text{Sepal.Width} - 0.9319\,\text{Petal.Length} + 2.8392\,\text{Petal.Width}$$

que corresponden a las funciones discriminantes.

Análisis discriminante en R

```
# matriz de confusion:
> table(predict(zLDA, type="class")$class, iris$Species)

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


# grafico de los 1ros ejes discriminantes
> Iris <- iris
> levels(Iris$Species) <- c("s", "c", "v")
> zLDA <- lda(Species ~ ., data = Iris)
> plot(zLDA, col = as.integer(Iris$Species))
```

Análisis discriminante en R

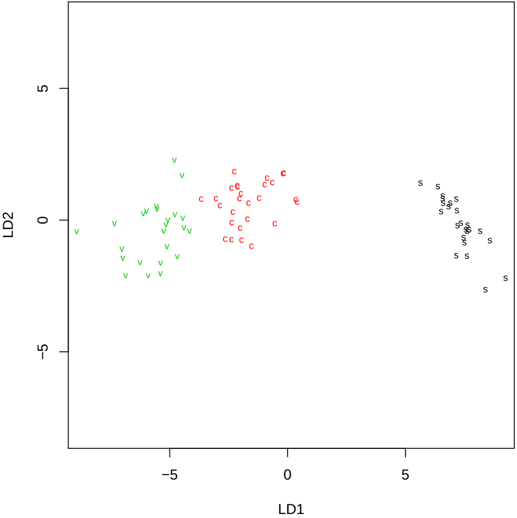


```
# muestra de entrenamiento
> train <- sample(1:150, 75)
> train
 [1] 115  57  93 108  97  47 146  19  23  33  60 114  94  85
[15]  56  28  41   6  91  26  59  63 124  10  80 119   5 148
[29]  20  92 111 144  68  36 147 135  35  86  62  77  21 126
[43]  99  58  52  30 143 149  70  95  65 130 150  46   3  40
[57]  42  48 136  54  74 123  67 140   4 132  15 120 104  73
[71]  51  16 127   9 117

# ajuste con datos de entrenamiento
> z <- lda(Species ~ ., data = Iris, subset = train)
> plot(z, col = as.integer(Iris$Species[train]))

# prediccion
> pLDA <- predict(object = z, newdata = Iris[-train,])
> attributes(pLDA)
$names
[1] "class"     "posterior" "x"
```

Análisis discriminante en R

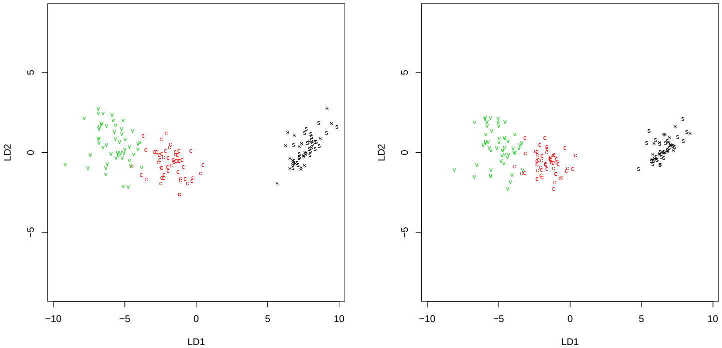
```
 > pLDA$class
 [1] s s s s s s s s s s s s s s s s s s s s s s s s s s
[27] c c c c c c v c c c v c c c c v c c c c c c c v v v
[54] v v v v v v v v v v v v v v v v c v v v v v v
Levels: s c v

> pLDA$posterior
   s             c             v
1   1.000000e+00 8.740338e-20 3.120120e-39
2   1.000000e+00 1.561053e-16 4.569270e-35
7   1.000000e+00 2.981436e-16 1.902425e-34
8   1.000000e+00 4.469592e-18 5.156442e-37
...
142 3.245900e-36 5.181672e-05 9.999482e-01
145 3.606253e-46 6.532700e-08 9.999999e-01

> pLDA$x
           LD1          LD2
1     7.7382492 -0.406234684
2     6.9812911  0.446793018
7     6.8781938  0.103031199
8     7.3362451 -0.005109942
...
142  -5.3569476 -2.681465191
145  -7.1323220 -2.306079275
```

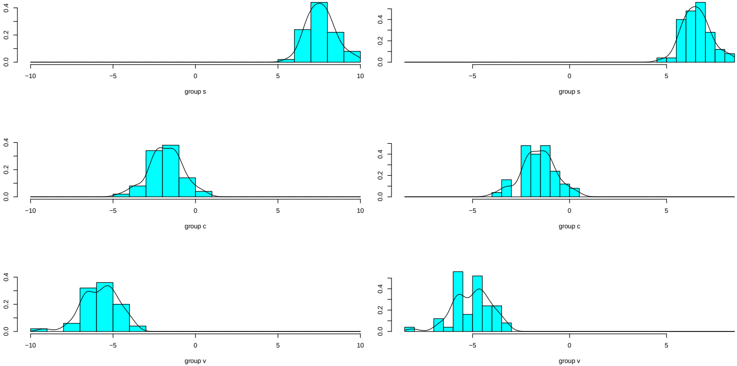
Análisis discriminante en R



```
> tLDA <- lda(Species ~ ., data = Iris, method = "t", nu = 4.)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.3133  0.5975  0.7381  0.7104  0.8303  0.9852
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1947  0.4604  0.6248  0.6066  0.7425  0.9776
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1538  0.3997  0.5724  0.5582  0.6895  0.9724
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1355  0.3678  0.5424  0.5330  0.6645  0.9690
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1262  0.3513  0.5244  0.5191  0.6518  0.9669
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1211  0.3423  0.5142  0.5112  0.6445  0.9656
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1183  0.3372  0.5079  0.5066  0.6402  0.9648
```

```
> tLDA
Call:
lda(Species ~ ., data = Iris, method = "t", nu = 4)

Prior probabilities of groups:
        s         c         v
0.3333333 0.3333333 0.3333333

Group means:
  Sepal.Length Sepal.Width Petal.Length Petal.Width
s     4.982772    3.400171     1.461310   0.2387506
c     5.949955    2.791431     4.259857   1.3221475
v     6.504207    2.966737     5.461265   2.0334488

Coefficients of linear discriminants:
                    LD1        LD2
Sepal.Length  0.5232151 -0.1348507
Sepal.Width   1.4649627  1.5219051
Petal.Length -1.7953005 -1.1831394
Petal.Width  -2.4778734  3.2659039

Proportion of trace:
   LD1    LD2
0.9907 0.0093
```


LD2
5
0
−5
−10
−5
0
5
10
LD1
LD2
5
0
−5
−10
−5
0
5
10
LD1


group s
group s
group c
group c
group v
group v