Martha Rocío Estrada
Luis Alberto Gamba Mondragón
José Francisco Méndez de los Ríos

ANÁLISIS DISCRIMINANTE

El análisis discriminante, se utiliza para identificar las características que permiten diferenciar a dos o más grupos de sujetos; además para clasificar nuevos casos como pertenecientes a uno u otro grupo. El análisis discriminante ayuda a identificar las características que diferencian a dos o más grupos; ayuda también a crear una función capaz de distinguir con la mayor precisión posible a los miembros de grupos.

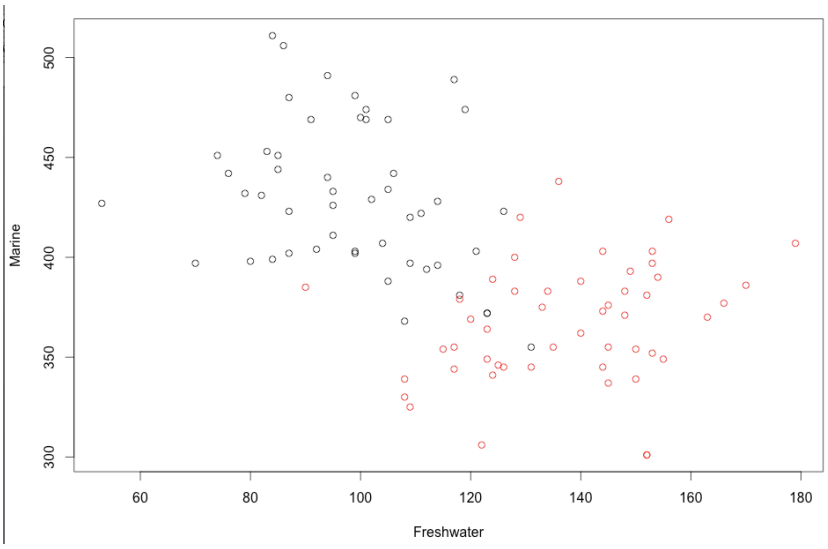
Mediante esta técnica estadística es posible conocer qué variables permiten diferenciar a los grupos y cuántas de estas son necesarias para alcanzar la mejor clasificación posible. Las variables discriminantes, también conocidas como variables de clasificación o dependientes, deben ser variables cuantitativas o continuas, o al menos admitir un tratamiento numérico con significado. Se desea reducir la dimensionalidad de las "p" variables independientes a una sola dimensión (una combinación lineal D, con las ponderaciones de las variables independientes que consiguen hacer que los sujetos de uno de los dos grupos obtenga puntuaciones máximas en D, y los sujetos del otro grupo puntuaciones mínimas); en esta dimensión se espera que los grupos difieran los más posible.

Son las puntuaciones de los sujetos en la nueva dimensión (puntuaciones discriminantes), las que permiten llevar a cabo la clasificación de los sujetos.

COMENTARIOS DEL EJEMPLO: SALMON (Alaska/Canadá)

> plot(salmon[,-1],col=as.factor(salmon[,1]))

Se obtiene un diagrama de dispersión entre las variables "Freshwater" y "Marine", se toma como variable de agrupación por color el origen del salmón Alaska/Canadá:



Se forman dos grupos, entre las características que manifiesta el salmón, de acuerdo al país de origen: Alaska o Canadá.

Martha Rocío Estrada
Luis Alberto Gamba Mondragón
José Francisco Méndez de los Ríos

> strain=salmon[c(1:40,51:90),]

> stest=salmon[c(41:50,91:100),]

Estos comandos extraen algunos elementos de la base de datos, el primero deja fuera las observaciones en los rangos (41-50) y (91-100). El segundo comando, conforma una base de datos que es el complemento de la primer base, dejando fuera las observaciones de los rangos (1-40) y (51-90)

> lsol=lda(strain[,c(2,3)],grouping=strain[,1])

La función “lda”, obtiene los grupos en los cuales se reúnen las observaciones. En el caso del comando anterior, se genera una variable “lsol” que contiene los resultados del análisis discriminante de la primera selección de observaciones de la base.

A continuación se presentan los resultados obtenidos de la función “lda”, lsol$prior arroja los valores de la probabilidad de que los elementos pertenezcan a uno u otro grupo, en este caso es 0.5, ya que las observaciones se agrupan en un 50% en cada grupo. Por otro lado, lsol$means, muestra las medias obtenidas de las variables en cada grupo, de a cuerdo con la división de grupos que se conforma. La variable que se utilizó para agrupar a los sujetos fue la primera variable de la base “strain”.

```
> lsol=lda(strain[,c(2,3)], grouping=strain[,1])
> lsol$prior
Alaska Canada
   0.5    0.5
> lsol$means
       Freshwater  Marine
Alaska    100.550 422.275
Canada    138.625 368.650
```

Martha Rocío Estrada
Luis Alberto Gamba Mondragón
José Francisco Méndez de los Ríos

> alaskasalmon=salmon[c(1:40),c(2,3)]
> canadasalmon=salmon[c(51:90),c(2,3)]

Los comandos anteriores conforman dos bases, el primero selecciona las observaciones relativas a las características de los salmones de Alaska y el segundo selecciona las observaciones referentes al salmón de Canadá, es importante mencionar que dejan fuera diez observaciones en cada grupo . Posteriormente con el comando siguiente se obtienen los promedios de las covarianzas de los grupos:

> singlecov=(39/78)*(cov(alaskasalmon)+cov(canadasalmon))

```
> singlecov=(39/78)*(cov(alaskasalmon)+cov(canadasalmon))
> singlecov
           Freshwater     Marine
Freshwater  322.22147  -15.24744
Marine      -15.24744 1087.44968
```

También se obtuvieron los coeficientes lineales discriminantes, los cuales constituyen las ponderaciones las variables independientes para construir la proyección D unidimensional de las nubes de datos:

```
> lsol$scaling
                    LD1
Freshwater  0.04390178
Marine     -0.01806237
```

A continuación se realiza la predicción de una observación, es decir se busca a que grupo pertenece una observación cuyas características son Freshwater=120 y Marine=380:

```
> predict(lsol,c(120,380))
$class
[1] Canada
Levels: Alaska Canada

$posterior
       Alaska    Canada
[1,] 0.3132047 0.6867953

$x
          LD1
[1,] 0.2973989
```

Un sujeto con las características (120,380), pertenece al grupo de Canadá, de acuerdo con estas características se trata de un salmón de Canadá. Por otro lado $posterior, indica la probabilidad de que el elemento pertenezca a uno u otro grupo, en este caso la probabilidad mayor está en el grupo de Canadá.

Martha Rocío Estrada
Luis Alberto Gamba Mondragón
José Francisco Méndez de los Ríos

El ejercicio siguiente, consistió en hacer la predicción del lugar que ocuparían los elementos seleccionados de la base "stest" donde se agrupaban las observaciones (41-50) y (91-100), con los grupos formados por "lsol":

```
> predict(lsol,stest[,c(2,3)])
$class
 [1] Alaska Alaska Alaska Alaska Alaska Alaska Alaska Alaska Alaska
[10] Alaska Canada Canada Canada Canada Canada Canada Canada Canada
[19] Canada Canada
Levels: Alaska Canada
```

Se observa que son asignados de acuerdo a la base principal "SALMON", la asignación concuerda con los datos originales.

```
$posterior
         Alaska       Canada
41  0.999934575 6.542453e-05
42  0.998909821 1.090179e-03
43  0.999641196 3.588039e-04
44  0.997267179 2.732821e-03
45  0.991071121 8.928879e-03
46  0.990434148 9.565852e-03
47  0.973525192 2.647481e-02
48  0.998445913 1.554087e-03
49  0.999459094 5.409062e-04
50  0.999593904 4.060962e-04
91  0.073753358 9.262466e-01
92  0.172305247 8.276948e-01
93  0.068420264 9.315797e-01
94  0.019825308 9.801747e-01
95  0.061697460 9.383025e-01
96  0.001990077 9.980099e-01
97  0.042753089 9.572469e-01
98  0.048058245 9.519418e-01
99  0.002611083 9.973889e-01
100 0.205956271 7.940437e-01
```

Probabilidades de pertenecer a uno u otro grupo, para la asignación de la base "stest"

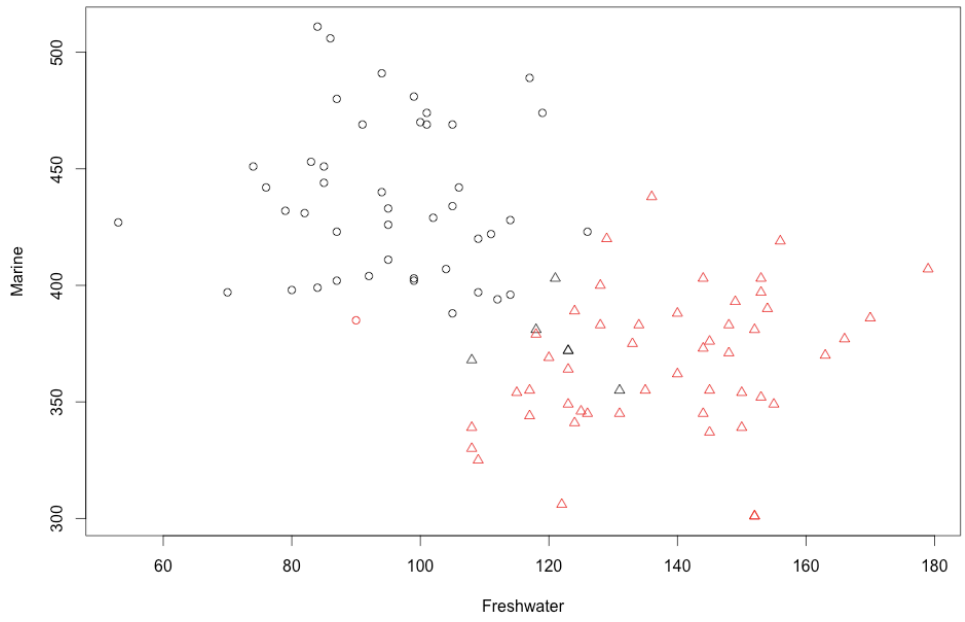
Martha Rocío Estrada
Luis Alberto Gamba Mondragón
José Francisco Méndez de los Ríos

Se realiza un Cross-Validation, de tal manera que se observa el ajuste de la asignación en grupos de los elementos. Al agregar CV a la función "lda", regresa otro ajuste de los datos en los grupos:

```
> lsolcv=lda(SALMON[,c(2,3)], grouping=SALMON[,1], CV=TRUE)
> lsolcv
$class
  [1] Canada Canada Alaska Alaska Alaska Alaska Alaska Alaska Alaska Alaska
 [11] Alaska Canada Canada Alaska Alaska Alaska Alaska Alaska Alaska Alaska
 [21] Alaska Alaska Alaska Alaska Alaska Alaska Alaska Alaska Alaska Canada
 [31] Alaska Canada Alaska Alaska Alaska Alaska Alaska Alaska Alaska Alaska
 [41] Alaska Alaska Alaska Alaska Alaska Alaska Alaska Alaska Alaska Alaska
 [51] Canada Canada Canada Canada Canada Canada Canada Canada Canada Canada
 [61] Canada Canada Canada Canada Canada Canada Canada Canada Canada Canada
 [71] Alaska Canada Canada Canada Canada Canada Canada Canada Canada Canada
 [81] Canada Canada Canada Canada Canada Canada Canada Canada Canada Canada
 [91] Canada Canada Canada Canada Canada Canada Canada Canada Canada Canada
Levels: Alaska Canada
```

Del análisis anterior se observa que hay observaciones que son reasignadas al grupo opuesto. El comando anterior regresa también las probabilidades de los sujetos de pertenecer a uno u otro grupo. Para visualizar la dispersión en los grupos se presenta el siguiente comando con la gráfica correspondiente, con la figura se observa su asignación real, situación que por color no se manifiesta:

```
> plot(SALMON[,c(2,3)], col=as.factor(SALMON[,1]), pch=as.numeric(lsolcv$class))
```



Martha Rocío Estrada
Luis Alberto Gamba Mondragón
José Francisco Méndez de los Ríos

Análisis discriminante cuadrático

Otra forma de asignar los elementos a los grupos se efectúa mediante el comando “qda”, como se muestra a continuación:

```
> qsol=qda(strain[,c(2,3)], grouping=strain[,1])
> qsol
Call:
qda(strain[, c(2, 3)], grouping = strain[, 1])

Prior probabilities of groups:
Alaska Canada
   0.5    0.5

Group means:
       Freshwater  Marine
Alaska    100.550 422.275
Canada    138.625 368.650
```

Se obtienen todos los valores que se tienen con la función “lda”.

Para predecir la asignación de observaciones en los grupos, se emplea la siguiente línea de comandos:

```
> predict(qsol, stest[,c(2,3)])
$class
 [1] Alaska Alaska Alaska Alaska Alaska Alaska Alaska Alaska Alaska Alaska
[11] Canada Canada Canada Canada Canada Canada Canada Canada Canada Canada
Levels: Alaska Canada

$posterior
         Alaska       Canada
41  0.999999603 3.973313e-07
42  0.999934364 6.563641e-05
43  0.999975258 2.474169e-05
44  0.999739798 2.602019e-04
45  0.992748801 7.251199e-03
46  0.997730481 2.269519e-03
47  0.985325240 1.467476e-02
48  0.999895544 1.044557e-04
49  0.999982257 1.774293e-05
50  0.999987507 1.249314e-05
91  0.112831650 8.871684e-01
92  0.237765000 7.622350e-01
93  0.111230243 8.887698e-01
94  0.030850466 9.691495e-01
95  0.074449894 9.255501e-01
96  0.008231741 9.917683e-01
97  0.076453258 9.235467e-01
98  0.085706823 9.142932e-01
99  0.007297430 9.927026e-01
100 0.182480379 8.175196e-01
```