

## Análisis discriminante

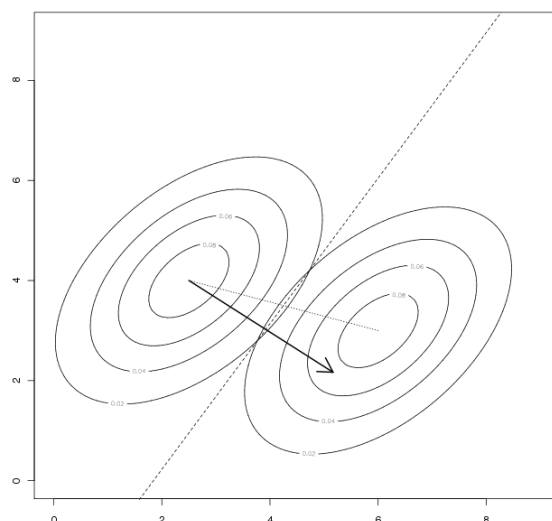
Francesc Carmona

9 de junio de 2022

1. Consideremos dos variables aleatorias  $(X_1, X_2)$  con distribución normal bivalente  $N_2(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$  para los elementos de una población y con distribución  $N_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$  para los elementos de otra población. Los vectores de medias son  $\boldsymbol{\mu}_1 = (2.5, 4)$  y  $\boldsymbol{\mu}_2 = (6, 3)$  respectivamente, pero las dos poblaciones tienen la misma matriz de covarianzas

$$\boldsymbol{\Sigma} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

La siguiente imagen<sup>1</sup> nos puede ayudar



- a) Hallar el discriminador lineal de Fisher

$$L(\mathbf{x}) = \mathbf{a}'\mathbf{x} - \mathbf{a}'\boldsymbol{\mu}_c = \mathbf{a}'(\mathbf{x} - \boldsymbol{\mu}_c)$$

donde  $\boldsymbol{\mu}_c = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$ , que asigna  $\mathbf{x}$  a una de las dos poblaciones según su signo.

- b) Comprobar matricialmente que  $L(\mathbf{x})$  es básicamente la diferencia entre los cuadrados de las distancias de Mahalanobis del punto  $\mathbf{x}$  a las medias:

$$L(\mathbf{x}) = \frac{1}{2} (d_M^2(\mathbf{x}, \boldsymbol{\mu}_2) - d_M^2(\mathbf{x}, \boldsymbol{\mu}_1))$$

Luego el criterio de Mahalanobis de asignar un punto a la población más cercana a su media coincide con el criterio de Fisher.

---

<sup>1</sup>El código para dibujarla, así como otros detalles, se hallan en el artículo  
<http://erre-que-erre-paco.blogspot.com/2010/05/discriminador-lineal-de-fisher.html>

2. En el archivo `wine.data` se recogen los resultados de un análisis químico del vino criado en la misma región italiana pero de tres viticultores distintos. El análisis determinó las cantidades de 13 constituyentes hallados en cada uno de los tres tipos de vino.

```
archivo <-  
"http://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data"  
vinos <- read.csv(archivo,header=F)  
names(vinos)<- c("tipo",paste("X",1:13,sep=""))  
vinos$tipo <- as.factor(vinos[,1])  
attach(vinos)
```

Considerando únicamente los tipos de vino 1 y 2 y las variables  $X_1$  (Alcohol) y  $X_{13}$  (Proline),

- Expresar la regla de clasificación lineal de Fisher para una nueva observación  $(x_1, x_{13})$ .
- Aplicar la regla anterior al caso concreto en que  $(x_1, x_{13}) = (13.05, 515)$ . ¿A qué tipo de vino corresponde?
- Representar gráficamente los puntos de las dos poblaciones y la recta equidistante de las medias<sup>2</sup>.
- Probar la función `lda()` del paquete MASS para este conjunto restringido de datos.
- Probar la función `plot()` sobre el objeto resultado de la función `lda()`. Obtener el gráfico con histogramas y con densidades (mejor con `ggplot2`).
- Probar la función `predict()` para la observación  $(x_1, x_{13}) = (13.05, 515)$ .
- Obtener la tabla de clasificación, también llamada tabla de *confusión*, donde se cruzan los datos reales con las clases obtenidas.  
Calcular el error o *misclassification rate*

$$\text{Error} = 1 - \frac{\text{Total bien clasificados}}{\text{Total}} = 1 - \frac{\sum n_{ii}}{n}$$

- Obtener la tabla de validación cruzada mediante el parámetro `CV = TRUE` que genera las predicciones por el sistema *leave one out*.  
¡Atención! Cuando se utiliza este parámetro el objeto que resulta no es un objeto `lda`, lo que provoca que no se pueda utilizar con las funciones `plot()` y `predict()` asociadas a dichos objetos.

3. Consideremos los datos sobre cráneos de varones egipcios de cinco épocas históricas que se pueden obtener en el siguiente enlace:

<http://lib.stat.cmu.edu/DASL/Datafiles/EgyptianSkulls.html>

También podemos bajarlos desde la página del libro de Everitt(2005) y así los podremos cargar directamente en **R** con las siguientes instrucciones:

```
skulls <- source("/(path)/chap5skulls.dat")$value  
str(skulls)  
attach(skulls)
```

donde el *path* debe ser la dirección a la carpeta donde hemos dejado el archivo una vez descomprimido. Así tendremos la base de datos `skulls` con cinco variables. La primera variable es el factor `EPOCH` y las otras cuatro son las medidas biométricas estudiadas del cráneo.

---

<sup>2</sup>Es la recta perpendicular al discriminador lineal de Fisher y que pasa por el punto medio entre los dos puntos medios de las poblaciones.

- a) En primer lugar podemos realizar un MANOVA para contrastar la diferencia de medias entre los niveles del factor o poblaciones. No entraremos aquí en la comprobación de las hipótesis de normalidad y de igualdad de las matrices de covarianzas.

Realizar un test de Wilks.

El test rechaza la igualdad de medias y, por lo tanto, justifica el análisis discriminante.

- b) Realizar un análisis lineal discriminante con la función `lda()` del paquete MASS.

Obtener los vectores de medias para las distintas poblaciones y los coeficientes de las variables canónicas.

Observar que con el nombre de `Coefficients of linear discriminants`<sup>3</sup> justamente se obtienen los coeficientes de las variables canónicas. Éstos están normalizados respecto a la matriz **S** de covarianzas dentro de los grupos.

- c) Calcular la matriz **S** de covarianzas dentro de los grupos o *pooled within-groups covariance matrix* sobre los cinco grupos.

Para ello podemos utilizar el siguiente código<sup>4</sup>, donde **g** es el factor y **X** la matriz de datos numéricos,

```
S <- (n-1)*var(X - medias[g,])/(n-length(levels(g)))
```

Comprobar que los coeficientes **b<sub>i</sub>** de las variables canónicas verifican **b<sub>i</sub>'Sb<sub>i</sub> = 1**.

- d) Calcular los coeficientes<sup>5</sup> lineales  $\bar{\mathbf{x}}_i' \mathbf{S}^{-1}$  y la “constante”  $-\frac{1}{2} \bar{\mathbf{x}}_i' \mathbf{S}^{-1} \bar{\mathbf{x}}_i + \log(\pi_i)$  para cada población, donde  $\pi_i$  es la probabilidad a priori de la *i*-ésima población. En nuestro caso suponemos  $\pi_i = 1/5$ .

- e) Hallar los coeficientes de las funciones<sup>6</sup> discriminantes  $h_{ij}(\mathbf{x})$

$$\mathbf{a}'_{ij} = \bar{\mathbf{x}}_i' \mathbf{S}^{-1} - \bar{\mathbf{x}}_j' \mathbf{S}^{-1} = (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)' \mathbf{S}^{-1}$$

y su constante

$$\begin{aligned} c_{ij} &= -\frac{1}{2} \bar{\mathbf{x}}_i' \mathbf{S}^{-1} \bar{\mathbf{x}}_i + \log(1/5) - \left[ -\frac{1}{2} \bar{\mathbf{x}}_j' \mathbf{S}^{-1} \bar{\mathbf{x}}_j + \log(1/5) \right] \\ &= -(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)' \mathbf{S}^{-1} \left[ \frac{1}{2} (\bar{\mathbf{x}}_i + \bar{\mathbf{x}}_j) \right] \\ &= -\mathbf{a}'_{ij} \mathbf{m}_{ij} \end{aligned}$$

de modo que la función discriminante entre dos poblaciones es

$$\begin{aligned} h_{ij}(\mathbf{x}) &= \mathbf{a}'_{ij} \mathbf{x} + c_{ij} = \mathbf{a}'_{ij} \mathbf{x} - \mathbf{a}'_{ij} \mathbf{m}_{ij} = \mathbf{a}'_{ij} (\mathbf{x} - \mathbf{m}_{ij}) \\ &= (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)' \mathbf{S}^{-1} \left[ \mathbf{x} - \frac{1}{2} (\bar{\mathbf{x}}_i + \bar{\mathbf{x}}_j) \right] \end{aligned}$$

- f) Obtener con la función `plot()` el gráfico que muestra la separación entre las poblaciones que proporciona la primera variable discriminante.
- g) Hallar la tabla de clasificación sencilla o *plug-in* y la tabla de validación cruzada.
- h) Representar los puntos medios de las poblaciones en el diagrama con las dos primeras variables canónicas.

Para ello podemos utilizar el siguiente código:

```
dsfs <- as.matrix(skulls[, -1]) %*% dis$scaling[, 1:2]
medias.cv <- aggregate(dsfs, skulls["EPOCH"], mean)
plot(medias.cv[, 2:3], type="n", xlim=c(1, 3.5))
text(medias.cv[, 2:3], labels=medias.cv[, 1])
```

<sup>3</sup>Se obtienen con el valor `scaling`.

<sup>4</sup>Algunos autores dividen la suma de cuadrados por *n* o por *n* - 1.

<sup>5</sup>En la página 402 del libro de Peña(2002) se llaman  $\mathbf{w}_g = \mathbf{V}^{-1} \boldsymbol{\mu}_g$ .

<sup>6</sup>En la página 402 del libro de Peña(2002) esencialmente son las funciones  $A_{ij}(\mathbf{x})$ .

o con los datos de las variables canónicas centrados

```
skulls.pr <- predict(dis, dimen=2)
skulls.t <- skulls.pr$x[,1:2] # =dsfs pero con columnas centradas
skulls.m <- lda(skulls.t, EPOCH)$means
# plot
eqscplot(skulls.t, type="n")
text(skulls.t, labels=1:5)
text(skulls.m, labels=rownames(skulls.m), col="blue")
```

4. Vamos a construir una regla para predecir el sexo de un futuro cangrejo *Leptograpsus* de especie desconocida tomando como muestra la base de datos **crabs** del paquete **MASS** de **R**. Los autores del paquete Venables & Ripley(2003, pág. 334-336) sugieren eliminar del análisis la variable **BD** ya que se midió de forma distinta para machos y hembras. También argumentan la utilización de los logaritmos de las variables.

```
library(MASS)
data(crabs)
lcrabs <- log(crabs[,4:8])
```

- a) Realizar un análisis discriminante lineal con la función **lda()** para el sexo con las cuatro variables numéricas transformadas.

Calcular la tabla de clasificación.

- b) Repetir el análisis teniendo en cuenta las cuatro clases en función del sexo y la especie. Los machos se codifican con una letra mayúscula y las hembras con una minúscula:

```
crabs.grp <- as.factor(c("B","b","O","o")[rep(1:4, each=50)])
```

- c) Como las dos primeras variables canónicas dominan la variación entre grupos, dibujar el gráfico de dispersión de los datos con ellas. El siguiente código nos puede servir:

```
(dcrabs.lda4 <- lda(crabs.grp ~ FL + RW + CL + CW, lcrabs))
dcrabs.pr4 <- predict(dcrabs.lda4, dimen=2)
```

```
# Male posterior probabilities
dcrabs.pr2 <- dcrabs.pr4$pos[, "B"] + dcrabs.pr4$pos[, "O"]
table(crabs$sex, dcrabs.pr2 > 0.5)
```

```
cr.t <- dcrabs.pr4$x[,1:2]
eqscplot(cr.t, type="n", xlab="First LD", ylab="Second LD")
text(cr.t, labels=as.character(crabs.grp))
```

- d) Añadir al gráfico anterior los puntos medios que representan a los dos sexos.

```
cr.m <- lda(cr.t, crabs$sex)$means
```

Añadir también la recta discriminante.

El siguiente código calcula la recta perpendicular al segmento que une dos puntos **x** e **y** y que pasa por su punto medio:

```
perp <- function(x,y) {
  m <- (x+y)/2
  s <- - (x[1] - y[1])/(x[2] - y[2])
  abline(c(m[2] - s*m[1], s))
  invisible()
}
```

5. Los científicos que han estudiado algunos de los insectos sociales, tales como las abejas y las hormigas, han observado que las reinas y las trabajadoras tienen diferentes tamaños y formas. En la página web

<https://pages.stat.wisc.edu/~yandell/pda/data/Wasp/>

se describen los datos y las variables de un experimento sobre las abejas.

Para cargar esos datos en **R** hacemos:

```
wasp <- read.table("https://pages.stat.wisc.edu/~yandell/pda/data/Wasp/wasp.dat",
                  header=T, stringsAsFactors = T)
```

- a) Analizar las características morfológicas de las abejas de forma descriptiva univariante y multivariante según el factor `caste`. Añadir algunos gráficos ilustrativos.
  - b) Realizar un análisis discriminante lineal.  
La función `lda()` del paquete **MASS** puede servir.
  - c) (\*\*) Comparar las matrices de covarianzas de las dos poblaciones con el test de la razón de verosimilitudes (ver el ejercicio 13e del segundo módulo).  
También se puede aplicar el test  $M$  de Box<sup>7</sup>. Ambos son muy sensibles a la no normalidad de los datos y tienden a rechazar la igualdad de covarianzas.
  - d) En el caso de poblaciones normales con diferentes matrices de covarianzas se clasificará cada observación en el grupo con máxima probabilidad a posteriori, pero entonces las funciones discriminantes no son lineales, ya que tienen un término de segundo grado.  
Realizar un análisis discriminante cuadrático.  
La función `qda()` del paquete **MASS** nos ayudará.
  - e) Calcular el número de parámetros que hay que estimar en la discriminación lineal y en la cuadrática.
  - f) Calcular los errores de clasificación con ambas reglas utilizando validación cruzada.  
Si son similares, nos quedaremos con el análisis lineal que además es más robusto y de mejor interpretación.
6. El banco de datos `wbca` contiene información sobre un estudio de cáncer de mama en el estado norteamericano de Wisconsin. La variable `Class` indica si el tumor es maligno (valor 0) o si es benigno (valor 1). El resto son variables que describen el estado de la paciente en una escala de 1 (normal) a 10 (muy poco normal) evaluadas por la observación de un médico. El objetivo es clasificar a la paciente como enferma o no de cáncer utilizando estas variables en lugar de algún otro método más invasivo.

```
library(faraway)
library(MASS)
attach(wbca)
```

- a) Realizar un análisis discriminante lineal.  
Añadir algún gráfico ilustrativo.
- b) Como tenemos una muestra muy grande, podemos elegir una muestra de entrenamiento para estimar mejor los errores de clasificación.

```
train <- sample(1:nrow(wbca), 400)
```

---

<sup>7</sup>[http://publib.boulder.ibm.com/infocenter/spssstat/v20r0m0/topic/com.ibm.spss.statistics.help/alg\\_boxs-m.htm](http://publib.boulder.ibm.com/infocenter/spssstat/v20r0m0/topic/com.ibm.spss.statistics.help/alg_boxs-m.htm)  
Un código para su cálculo en **R**: <http://finzi.psych.upenn.edu/R/Rhelp02a/archive/33330.html>

Utilizar el parámetro `subset = train` para limitar el análisis a este subconjunto. Predecir el resto de elementos con las variables obtenidas y calcular la tabla de clasificación.

Comparar el resultado con las tablas de clasificación *plug-in* y de valoración cruzada con todas las pacientes, normalmente más optimistas.

- c) (\*) Cuando el factor tiene dos clases, la primera combinación discriminante es equivalente a la que se obtiene en una regresión múltiple<sup>8</sup> tomando como variable respuesta una variable numérica con valores  $-1/n_1$  y  $1/n_2$ , según la clase, y donde  $n_1$  y  $n_2$  son los tamaños de los dos grupos.
- d) (\*\*) Como el factor considerado sólo tiene dos estados, también podemos acometer una regresión logística:

```
glm(Class ~ ., wbca, family = binomial, subset = train)
```

7. Consideremos un grupo de pacientes de los cuales se conocen algunas variables obtenidas de un análisis de orina. En concreto las variables nos dan la gravidez específica `grav`, la osmolaridad `osmo`, la conductibilidad `conduc`, la concentración de urea `urea` y la concentración de calcio `calcio`. También tenemos una variable `grupo` que nos indica la presencia o ausencia de cristales en la orina del individuo, donde 1 indica la ausencia y 2 la presencia. El siguiente código prepara los datos para su análisis:

```
archivo <- "https://www2.stat.duke.edu/courses/Spring01/sta114/data/Andrews/T44.1"

x <- read.table(archivo, header=F)
x <- x[, -(1:4)]
x[x == -9999] <- NA
x[, 1] <- factor(x[, 1], levels=1:2, labels=c("A", "P"))
cristal <- x; remove(x)
names(cristal) <- c("grupo", "calcio", "conduc", "grav", "osmo", "ph", "urea")
```

Observemos la substitución del valor original *missing*  $-9999$  por el NA de **R**. También es conveniente tratar a la variable que indica el grupo de los individuos como un factor.

- a) Realizar un análisis discriminante lineal con la función `lda()` del paquete MASS y probabilidades a priori iguales a 0.5.  
Habrá que utilizar el parámetro `prior` y también<sup>9</sup> `na.action = na.omit`.
- b) Realizar otro análisis discriminante lineal con la función `lda()` pero con las probabilidades a priori que coincidan con las proporciones observadas dentro de la muestra.
- c) Hallar las probabilidades a posteriori en ambos casos.  
Deberemos utilizar el parámetro `CV = T`.

8. (\*) En este ejercicio vamos a utilizar los famosos datos `iris` que Fisher usó para introducir la metodología. Tenemos tres especies de iris y cuatro características numéricas medidas sobre cada iris. Los datos están en el *data.frame* `iris` del paquete MASS. El objetivo es capacitarnos para clasificar un futuro individuo en la especie correcta basándonos únicamente en las medidas físicas. En primer lugar recodificaremos el factor para utilizarlo mejor en los gráficos.

```
library(MASS)
data(iris)
levels(iris$Species)
Sp <- c("s", "v", "c")[as.numeric(iris$Species)]
```

---

<sup>8</sup>Ver el apéndice 13.2 de Peña(2002).

<sup>9</sup>Otra posibilidad es excluir de la base de datos todos los individuos con algún valor *missing*: `cc <- complete.cases(x); cristal <- x[cc, ]`.

- a) Realizar un gráfico con la función `pairs()` de las variables numéricas con distintos símbolos y colores para cada especie.
  - b) Antes de proceder con un análisis discriminante lineal, realizar un análisis de clasificación jerárquico y otro por el método  $k$ -medias.  
Obtener las tablas de clasificación por ambos métodos.
  - c) Realizar un análisis discriminante lineal y obtener la tabla de clasificación con la primera variable discriminante.
  - d) Dibujar el gráfico de dispersión de los datos transformados por las dos primeras variables canónicas.
  - e) Dibujar las distribuciones de los grupos en la primera variable discriminante.
9. (\*) El método  $k$ -nn ( $k$ -nearest neighbors, Fix y Hodges, 1951) es un método de clasificación supervisada (o aprendizaje con estimación basada en un conjunto de entrenamiento y prototipos) que sirve para estimar la función de densidad  $f(\mathbf{x}|C_j)$  de las variables predictoras  $\mathbf{x}$  en cada clase  $C_j$ ,  $j = 1, \dots, g$ .

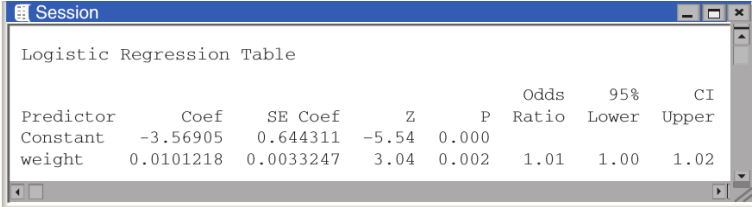
Este es un método de clasificación no paramétrico, que estima el valor de la función de densidad de probabilidad o directamente la probabilidad a posteriori de que un elemento  $\mathbf{x}$  pertenezca a la clase  $C_j$  a partir de la información proporcionada por el conjunto de prototipos. En el proceso de aprendizaje no se hace ninguna suposición acerca de la distribución de las variables predictoras.

Otros métodos no paramétricos difieren en la elección del conjunto de prototipos de referencia. El  $k$ -nn usa el conjunto de entrenamiento o un subconjunto de él. La “cercanía” se mide con la distancia euclídea.

Aplicar la función `knn()` del paquete `class` al conjunto de entrenamiento de 400 datos de `wbca` para predecir el tipo de tumor del resto.

10. (\*) Un estudio analizó una muestra aleatoria de personas adultas con diabetes de tipo 2 con el objetivo de identificar algunos de los factores de riesgo asociados con la enfermedad. El archivo `ex28_49.dat` contiene datos <sup>10</sup> sobre el estado de la diabetes, peso (en libras), circunferencia de la cintura (en pulgadas) y relación de colesterol (relación entre el colesterol HDL y el colesterol total en sangre) para todos los 386 sujetos. Cincuenta y nueve de los 386 se diagnosticaron con diabetes tipo 2 (diabetes = 1).

La diabetes de aparición en adultos o tipo 2 se ha asociado con la obesidad. La salida de la siguiente figura muestra los resultados de una regresión logística con la variable respuesta `diabetes` y `weight` como variable explicativa. Dar la ecuación del modelo de regresión logística. ¿Es significativo el coeficiente de la pendiente? ¿El peso parece aumentar las probabilidades de tener diabetes tipo 2?



Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% Lower	95% Upper
Constant	-3.56905	0.644311	-5.54	0.000			
weight	0.0101218	0.0033247	3.04	0.002	1.01	1.00	1.02

- a) Utilizar **R** para obtener el análisis de la regresión logística con la `diabetes` como variable respuesta y `weight` como variable explicativa. Se pueden obtener los mismos resultados que los mostrados en la figura anterior, salvo errores de redondeo.

Obtener la tabla de clasificación, también llamada tabla de *confusión*, y calcular el error o *misclassification rate*.

<sup>10</sup> Este ejercicio se ha extraído del capítulo suplementario sobre regresión múltiple y logística del libro *La práctica de la estadística para las ciencias de la vida* de Baldi y Moore.

- b)* La investigación sugiere que el lugar donde se almacena la grasa corporal es un factor importante en la predicción de la diabetes. Calcular el modelo con ambas variables, **weight** y **waist**, como explicativas. ¿Son ambas significativas? ¿Como queda la tabla de clasificación ahora?
- c)* Entre la lista de los factores de riesgo potenciales para la diabetes de tipo 2 se halla la razón de colesterol personal (**cholratio**), es decir, la razón entre el colesterol bueno o HDL y el total. Ajustar un modelo logístico con las variables **waist** y **cholratio** y estudiar el resultado.
- d)* Realizar un análisis LDA con las tres variables explicativas y estudiar el resultado.