uc3m | Universidad Carlos III de Madrid

Curso OpenCourseWare

Aprendizaje del Software Estadístico R: un entorno para simulación y computación estadística

Alberto Muñoz García

---

14. Clasificación de datos multivariantes: análisis discriminante

Introducción

La clasificación es una tarea del análisis de datos que suele conocerse en estadística bajo el nombre de "análisis discriminante". La situación en la que se utiliza el análisis discriminante se da cuando se tienen los siguientes ingredientes:

1. Una población de datos que se dividen en dos o más clases de acuerdo con una taxonomía determinada.
2. Un grupo de datos de los cuales se conoce a priori la clase a la que pertenecen.
3. Un conjunto de datos de los cuales deseamos saber a qué clase pertenencen.

Un problema de clasificación se asemeja a un problema de análisis de cluster en el hecho de que existen grupos, pero se diferencia en que en este caso concreto se sabe cuántos grupos hay y se conoce a qué grupo pertenece cada dato para un conjunto de datos etiquetado.

Ejemplos de aplicación de estas técnicas son la clasificación de medidas tomadas a partir de pacientes en sanos y enfermos. Por ejemplo, se dispone de una serie de medidas sobre una muestra de tejido de varias mujeres, y se sabe cuáles muestras corresponden a un tumor benigno, y cuáles corresponden a un tumor maligno. Se trataría de construir una función que, ante la presencia de una nueva serie de medidas de una persona no presentada anteriormente al sistema, nos dijese si la muestra corresponde a tumor benigno o maligno.

Otro ejemplo típico se da cuando tenemos una serie de datos de tipo económico por cada cliente de un banco, y se trata de saber si tiene riesgo de ser moroso o de no serlo.

La idea del método de clasificación estriba en que las medidas de los individuos bajo estudio forman grupos (nubes) de datos más o menos bien separados en el espacio de características, y en ese caso, es posible construir una función discriminante (de ahí el nombre del método en la literatura estadística) que permita separar los datos de un grupo de los datos de los demás grupos. Naturalmente, los problemas de clasificación pueden involucrar la presencia de dos clases, pero también la de un número más amplio de grupos.

Breve descripción de la terminología del problema de clasificación

En términos matemáticos, el problema de clasificación consiste en asignar un vector de p observaciones

$x = (x_1,x_2,...,x_p)$ a una de m clases $C_1,...,C_m$.

En este problema se suponen conocidas las llamadas probabilidades a priori de las clases $P(C_1)$,... , $P(C_m)$.

Tales probabilides indican la proporción de elementos existentes de cada clase. Por ejemplo, si el problema es de diagnosis médica y el 80% de las personas están sanas, y el 20% están enfermas (digamos gripe),

pues $P(C_1) = 0.8$ y $P(C_2) = 0.2$, donde C_1 = personas sin gripe , y C_2 = personas con gripe.

En este caso, $x = (x_1,x_2,...,x_p)$ sería un vector de medidas tales como la temperatura corporal y otras.

La regla de decisión que seguiría un análisis discriminante sería calcular (de alguna manera de entre varias posibles) las denominadas probabilidades a posteriori:

$P(C_1|x)$ = Probabilidad de que el individuo descrito por el vector x pertenezca a la clase C_1.

$P(C_1|x)$ = Probabilidad de que el individuo descrito por el vector x pertenezca a la clase C_2.

.....

$P(C_m|x)$ = Probabilidad de que el individuo descrito por el vector x pertenezca a la clase Cm.

La regla discriminante (y cualquier otro método sensato de clasificación) funcionará asignando el individuo descrito por el vector x a la clase que de la mayor probabilidad a posteriori.

En el problema de la gripe, si por ejemplo $P(C_1|x) = 0.9$ y $P(C_2|x) = 0.1$, entonces concluiremos que el individuo pertenece al grupo de los sanos. Naturalmente, hay una probabilidad de error al decidir de esta manera. Sin embargo está garantizado que la probabilidad de error, en promedio, es mínima con esta regla: Si se conocen las probabilidad esa posteriori (la probabilidad de pertenencia a una clase, una vez observadas las características x) , esta regla de clasificación, denominada "regla Bayes", es la mejor posible.

Para los que deseéis ampliar conocimientos sobre las técnicas de clasificación, podéis acudir a los siguientes links:

http://www.isip.msstate.edu/publications/reports/isip_internal/1998/linear_discrim_analysis/lda_theory.pdf(S. Balakrishnama, A. Ganapathiraju, Mississippi State University).

Clasificación del conjunto iris usando la función lda

El conjunto iris que ya hemos utilizado anteriormente, se almacena en una matriz 150 x 5. Las cuatro primeras variables contienen 4 medidas de 150 flores, y la última contiene una descripción de la especie de cada flor. Hay tres especies, "setosa", "virginica" y "versicolor". Podemos utilizar este conjunto de datos para resolver el problema de clasificar automáticamente las flores en las tres especies existentes.

Para resolver este problema utilizaremos la función lda (linear discriminant analysis) que está contenida en la librería MASS.

La idea de la técnica "linear discriminant analysis" consiste básicamente en buscar una dirección de proyección de los datos (en este caso de dimensión 4), de manera que los datos

proyectados sobre la recta correspondiente sean lo más separables posibles. Una vez encontrada una dirección de proyección, podremos buscar una segunda dirección de proyección, y así sucesivamente. Esto nos permitirá adicionalmente dibujar los datos sobre los dos primeros ejes de proyección para hacernos una idea más exacta de como son los grupos que estamos determinando.

```
> library(MASS)   # Cargamos la libreria que contiene a  lda
> data(iris)      # Cargamos los datos
> dim(iris)
[1] 150   5

> datos = data.frame(iris[,1:4],clase=as.vector(iris[,5]))
> lda(clase~.,datos)
Call:
lda.formula(clase ~ ., data = datos)

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

Como vemos, la función nos devuelve las probabilidades a priori de los grupos, que las calcula utilizando la proporción de elementos de cada clase (en este caso, 50 por clase, de un total de 150, hacen 50/150 = 1/3 = 0.3333333). Luego calcula la media de cada grupo, que daría la descripción media de la flor típica de cada grupo. Los coeficientes de los discriminantes lineales se usan por la función para decidir a qué clase pertenece cada ejemplar de flor. Y las últimas medidas dan una idea de la importancia de cada eje discriminante. Al ser LD1 mucho mayor que LD2, y casi cercano a 1, esto quiere decir que las flores se pueden clasificar muy bien utilizando un solo eje discriminante.

Ahora lo repetimos y guardamos el resultado bajo un nombre (iris.lda) para poder accederlo posteriormente:

> iris.lda = lda(clase~.,datos)

> attributes(iris.lda) # cosas que devuelve la función

$names

[1] "prior" "counts" "means" "scaling" "lev" "svd" "N"

[8] "call" "terms" "xlevels"

$class

[1] "lda"

> iris.lda$prior # Asi podemos ver,por ejemplo,las probabilidades a priori

setosa versicolor virginica

0.3333333 0.3333333 0.3333333

> predict(iris.lda,iris[1:4,])$class # predecimos por ejemplo, la clase

[1] setosa setosa setosa setosa # de las cuatro primeras flores

Levels: setosa versicolor virginica

Tal clase resulta ser setosa, lo cual es correcto. Para ver cómo de correctas fueron las predicciones, podemos hacer una tabla cruzando las verdaderas clases con las predicciones:

> table(iris[,5],predict(iris.lda,iris[,1:4])$class)

setosa versicolor virginica

setosa 50 0 0

versicolor 0 48 2

virginica 0 1 49

Como podemos ver, el clasificador ha cometido tan sólo 3 errores: dos datos de la clase versicolor han sido asignados a virginica, y un dato de virginica a versicolor. En la clase setosa no ha habido errores.

Otra manera alternativa de proporcionar el conjunto de datos a la función lda podría ser:

> especies = iris[,5]

> lda(iris[,1:4],especies)

Call:

lda.data.frame(iris[, 1:4], especies)

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

Para dibujar la proyección del conjunto de datos (que es de dimensión 4) sobre los dos primeros ejes discriminantes:

> iris.proy = predict(iris.lda,iris[,1:4])$x

> plot(iris.proy)

> plot(iris.proy,type="n")

> text(iris.proy,labels=especies)