Análisis de Datos Multivariados 8-Diciembre-2014

Semestre 2015-1

Profra.: M. en E. Leticia E. Gracia-Medrano Valdelamar

Ayudante: José Antonio Aranda Román

Tarea 6- Análisis de Discriminantes

Equipo:

* Avelino Palacios Cynthia Nizeida
* Rosales Córdova Aldebarán

El análisis discriminante es un conjunto de métodos y herramientas utilizados para distinguir entre grupos de poblaciones $π\_{j}$conocidos a priori, el objetivo principal es determinar la forma de asignar nuevas observaciones en los grupos. Dichos grupos podrían haber sido formadas por un análisis de conglomerados realizado con datos pasados.

#Leemos la base de datos obtenida desde internet

salmon<-read.table("https://www.cs.tcd.ie/~houldinb/Index/MVA\_files/salmon.txt",head=TRUE)

#Se carga la librería MASS

library(MASS)

#Se separan los grupos con distinción de color de acuerdo a su origen(Alaska o Canadá)

plot(salmon[,-1],col=as.factor(salmon[,1]), main="Gráfica de dispersión por procedencia del salmon")



En la gráfica se pueden observar la formación de grupos.

Cabe mencionar que el análisis de discriminantes asume que los datos dentro de cada grupo sigue una distribución normal multivariada y que la distribución de cada grupo comparte la misma matriz de covarianza.

Previamente para utilizar la función *LDA* (Linear DiscriminantAnalysis), primero se dividen los datos en una primera parte que se usa para entrenar al clasificador, y una segunda parte que se utiliza para probar el clasificador. En este ejercicio se hace unadivisión 80:20.

#Se declara la variable strain,concatenando una columna numérica ascendente para reasignar un número a cada salmón, (con esto se entrena el clasificador).

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

#Se declara la variable stest para que los 20 salmones restantes de cada nivel del factor se les reasigne un número (con estos se prueba el clasificador).

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

#Con la función de Análisis de Discriminantes lineal (lna) se forman grupos.

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

lsol

Call:

lda(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

Coefficients of linear discriminants:

LD1

Freshwater 0.04390178

Marine -0.01806237

#Se obtienen las probabilidades apriori por grupo (Alaska, Canadá), dichas probabilidadesse toma como las proporciones de clase en el entrenamiento de datos.

lsol$prior

Alaska Canada

 0.5 0.5

#Se obtienen las medias por grupo

lsol$means

Freshwater Marine

Alaska 100.550 422.275

Canada 138.625 368.650

#Se crea una variable únicamente con los datos de los salmones originarios de Alaska.

alaskasalmon=salmon[c(1:40),c(2,3)]

#Se obtiene la media de las variables agua dulce y agua salada del grupo se salmones originarios de Alaska.

medialaska<-apply(alaskasalmon,2,mean)

medialaska

Freshwater Marine

 100.550 422.275

#Se crea una variable únicamente con los datos de los salmones originarios de Canadá.

canadasalmon=salmon[c(51:90),c(2,3)]

#Se obtiene la media de las variables agua dulce y agua saladadel grupo de salmones originarios de Canadá.

mediacanada<-apply(canadasalmon,2,mean)

mediacanada

Freshwater Marine

 138.625 368.650

#Se utiliza una única matriz de covarianza que se utiliza para las agrupaciones.

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

singlecov

 Freshwater Marine

Freshwater 322.22147 -15.24744

Marine -15.24744 1087.44968

#Se realiza la función para el cálculo de discriminantes.

discriminante<- function(x,m1,m2,matvar){

dato<-x

discriminante<-(dato-.5\*(m1+m2))%\*%solve(matvar)%\*%(m1-m2)

return(discriminante)

}

#Se concatenan las medias de Alaska de las dos variables tipo de agua

puntoalaska<-(c( 100.550, 422.275))

#Se concatenan las medias de Canadá de las dos variables tipo de agua

puntocanada<-(c( 138.625, 368.650))

#Se obtienen los puntos medios por tipo de agua

puntomedio<-(c( 119.5875,395.4625 ))

#Se obtiene la discriminante del grupo de salmones de Alaska con la variable de sumedia

discriminante(puntoalaska,medialaska,mediacanada,singlecov)

 [,1]

[1,] 3.485209

#Se obtiene la discriminante del grupo de salmones de Canadá con la variable de su media

discriminante(puntocanada,medialaska,mediacanada,singlecov)

 [,1]

[1,] -3.485209

#Se obtiene la discriminante de todos los salmones con la variable de medias por tipo de agua

discriminante(puntomedio,medialaska,mediacanada,singlecov)

 [,1]

[1,] 0

#Se obtiene la discriminante de los salmones con una grabación de agua dulce de 120 y de marina de 380.

discriminante(c(120,380),medialaska,mediacanada,singlecov)

 [,1]

[1,] -0.7851791

#Se predice la clasificación con el centro del grupo de los salmones con origen en Alaska y se obtienen las probabilidades aposterioris.

predict(lsol,c( 100.550, 422.275))#centro alaska

$class

[1] Alaska

Levels: Alaska Canada

$posterior

 Alaska Canada

[1,] 0.970264 0.02973602

$x

 LD1

[1,] -1.320077

#Se predice la clasificación con el centro del grupo de los salmones con origen en CanadáAlaska y se obtienen las probabilidades aposterioris.

predict(lsol,c( 138.625, 368.650))# centro canada

$class

[1] Canada

Levels: Alaska Canada

$posterior

 Alaska Canada

[1,] 0.02973602 0.970264

$x

 LD1

[1,] 1.320077

#Se predice la clasificación para una observación con una grabación de agua dulce de 120 y de marina de 380 y se obtienen las probabilidades aposterioris.

predict(lsol,c(120,380))

$class

[1] Canada

Levels: Alaska Canada

$posterior

 Alaska Canada

[1,] 0.3132047 0.6867953

$x

 LD1

[1,] 0.2973989

#Se hace la predicción de los datos de prueba

predict(lsol,stest[,c(2,3)])

$class

[1] Alaska AlaskaAlaskaAlaskaAlaskaAlaskaAlaskaAlaskaAlaskaAlaskaCanada

[12] CanadaCanadaCanadaCanadaCanadaCanadaCanadaCanadaCanada

Levels: Alaska Canada

$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

$x

 LD1

41 -3.6492357

42 -2.5833037

43 -3.0045114

44 -2.2345978

45 -1.7837951

46 -1.7574513

47 -1.3653479

48 -2.4488379

49 -2.8489705

50 -2.9575969

91 0.9584339

92 0.5944260

93 0.9890377

94 1.4774783

95 1.0309356

96 2.3550095

97 1.1774384

98 1.1310284

99 2.2519041

100 0.5111346

#Se carga la libreriaklarR

library(klaR)

#Se graficala partición de los grupos con el discriminante lineal

partimat(SalmonOrigin~ Marine+Freshwater,data=strain,method="lda")



#Se grafica la partición de los grupos con el discriminante cuadrático

partimat(SalmonOrigin~ Marine+Freshwater,data=strain,method="qda")



#Se carga la librería ellipse

library(ellipse)

#Se grafican los grupos de origen del salmón

plot(salmon[,c(2,3)],col=as.factor(salmon[,1]),xlim=c(50,190),ylim=c(290,530),main="Gráfica de dispersión por procedencia del salmon")

#Se marca la masa de puntos de cada grupo Alaska

lines(ellipse(cov(salmon[c(1:50),c(2,3)]),centre=apply(salmon[c(1:50),c(2,3)],2,mean),level=0.5))

# Se marca la masa de puntos de cada grupo Canadá

lines(ellipse(cov(salmon[c(51:100),c(2,3)]),centre=apply(salmon[c(51:100),c(2,3)],2,mean),level=0.5),col=2)

