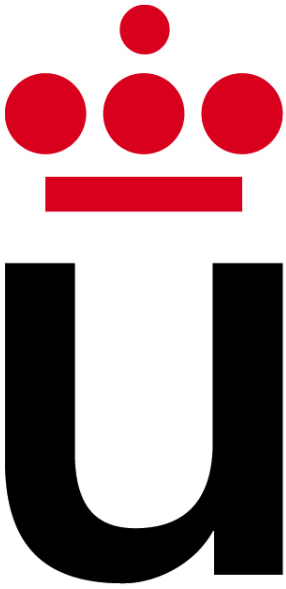
Caso Práctico I

Minería de datos

Análisis de clústers

José Ignacio Escribano



Móstoles, 7 de diciembre de 2015

Índice de figuras

1. Clusters mediante la representación de las variables dos a dos 2
2. Clusters mediante el análisis de componentes principales 2
3. Dendograma con el método ward.D 4
4. Dendograma con el método ward.D2 5
5. Dendogramas con los métodos faltantes 5

Índice de tablas

Índice

Índice de figuras **b**

Índice de tablas **c**

1. Introducción **1**

2. Resolución de las cuestiones de evaluación **1**
2.1. Cuestión 1 1
2.2. Cuestión 2 4

3. Conclusiones **6**

4. Código R utilizado **7**

1. Introducción

En este caso práctico aplicaremos la teoría vista para obtener grupos de países a partir de datos socioeconómicos usando el algoritmo de las k-medias. Por último, usaremos dendogramas para identificar tres tipos de iris a partir de la anchura y longitud de los pétalos y sépalos de sus hojas.

2. Resolución de las cuestiones de evaluación

A continuación resolveremos las cuestiones de evaluación planteadas.

2.1. Cuestión 1

En esta cuestión, utilizaremos los datos del fichero "SocioeconomicDatasets.csv", que contiene 6 variables que describen a 91 países. Estas variables son Birth.Rate (tasa de natalidad), Mortality.Rate (tasa de mortalidad), Infant.mortality.Rate (tasa de mortalidad infantil), Life.expectency.man (esperanza de vida en los hombres), Life.expectency.woman (esperanza de vida en mujeres) y GNP (Producto Interior Bruto).

Como en el caso resuelto, transformamos la variable GNP aplicando logaritmos. Renombramos la variable como logGNP. Además escalamos los datos, ya que cada variable tiene escalas distintas y afectará negativamente a la hora de calcular la matriz de distancias.

Aplicamos el algoritmos de las k-medias con $k = 5$, es decir, agruparemos los países en 5 grupos, de acuerdo a las variables explicativas.

Los 5 centroides son los siguientes:

	[,1]	[,2]	[,3]	[,4]	[,5]
Group.1	1.00000000	2.000000	3.00000000	4.0000000	5.0000000
Birth.Rate	0.43148667	1.273042	-0.69750278	-1.0993725	1.0732598
Mortality.Rate	-0.52272683	2.194461	-0.50506881	-0.4983267	0.7883010
Infant.mortality.Rate	0.25309057	1.925793	-0.72708079	-1.0011662	0.9011199
Life.expectency.man	-0.05682204	-2.024511	0.57326819	1.0367668	-0.9676394
Life.expectency.woman	-0.13881053	-1.925379	0.64835233	1.0498822	-1.0104112
logGNP	-0.24674302	-1.387819	-0.04384888	1.3387075	-0.7900894

Representamos los clusters, representando las variables dos a dos, como se puede ver en la Figura 1.

Este gráfico tiene difícil interpretabilidad, por lo que procedemos reduciendo la dimensión de los datos, haciendo uso del análisis de componentes principales. Así, los cinco clústers se pueden ver en la Figura 2.

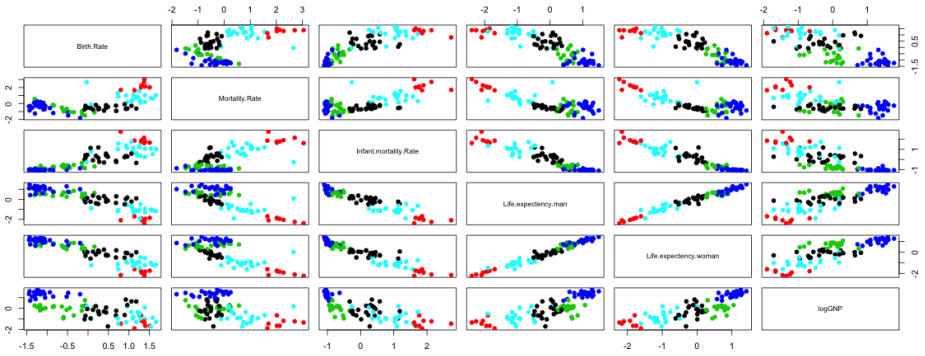


Figura 1: Clusters mediante la representación de las variables dos a dos

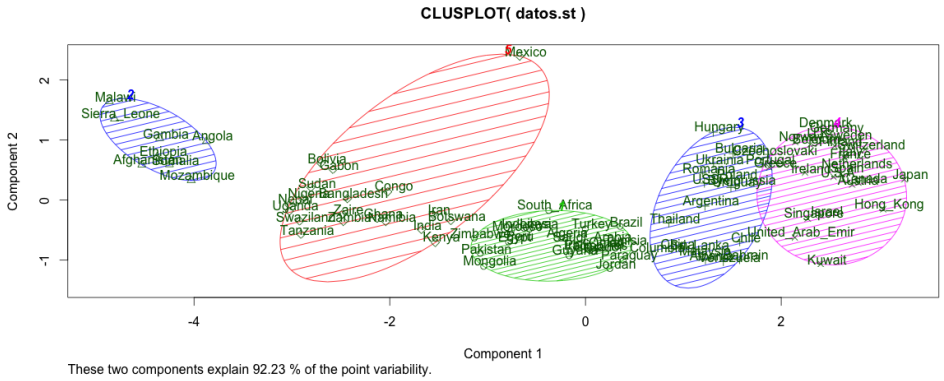


Figura 2: Clusters mediante el análisis de componentes principales

Este gráfico es mucho más explicativo que el anterior. Los países por grupos son los siguientes:

```
#-----------------------------------------------------------
# Grupo 1
#-----------------------------------------------------------
Brazil      Ecuador       Guyana     Paraguay      Peru
Oman        Saudi_Arabia  Turkey     Indonesia     Mongolia
Algeria     Egypt         Libya      Morocco       South_Africa
Iraq        Jordan        Pakistan   Philippines   Tunisia
Zimbabwe

#-----------------------------------------------------------
# Grupo 2
#-----------------------------------------------------------
Afghanistan  Angola  Ethiopia   Gambia  Malawi   Mozambique
Sierra_Leone Somalia

#-----------------------------------------------------------
# Grupo 3
#-----------------------------------------------------------
Albania    Bulgaria  Czechoslovaki  Hungary      Poland
Romania    USSR      Byelorussia    Ukrainia     Argentina
Chile      Columbia  Uruguay        Venezuela    Bahrain
China      Malaysia  Sri_Lanka      Thailand

#-----------------------------------------------------------
# Grupo 4
#-----------------------------------------------------------
Belgium    Finland   Denmark        France        Germany
Ireland    Italy     Netherlands    Norway        Portugal
Greece     Spain     Sweden         Switzerland   U.K.
Austria    Japan     Canada         U.S.A.        Israel
Kuwait     Hong_Kong Singapore      United_Arab_Emir

#-----------------------------------------------------------
# Grupo 5
#-----------------------------------------------------------
Bolivia     Mexico     Iran      Bangladesh India    Nepal
Botswana    Congo      Gabon     Ghana      Kenya    Namibia
Nigeria     Sudan      Swaziland Uganda     Tanzania  Zaire
Zambia
```

2.2. Cuestión 2

En esta cuestión usaremos la base de datos `iris`, disponible en el paquete datasets de R. La base de datos describe tres tipos de iris (setosa, virginica y versicolor) a partir de las longitud y anchura de los pétalos y de los sépalos.

De esta base de datos, elegimos 50 para obtener 3 clusters. En este caso, usamos clusters jerárquicos.

Usamos la distancia euclídea para obtener la matriz de distancias.

Tenemos varios métodos de agrupación disponibles en R para obtener el dendograma. Usaremos todos para ver con cuál se obtiene una mejor clasificación.

Comenzamos con el método ward.D. El el dendograma obtenido con este método se puede ver en la Figura 3

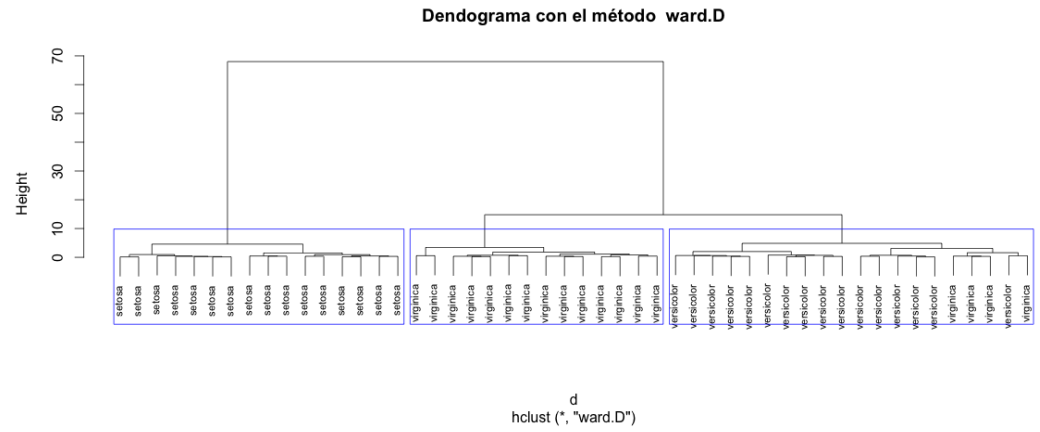


Figura 3: Dendograma con el método ward.D

Con este método se detecta perfectamente la especie setosa, pero no se distingue la especie versicolor de la virginica.

Usamos ahora el método ward.D2, cuyo dendograma se muestra en la Figura 4.

De nuevo, se obtiene que se detecta perfectamente la especie setosa, pero se mezclan las especies versicolor y virginica en el tercer cluster, el que está a la derecha.

Si repetimos lo anterior con los métodos restantes (single, complete, average, mcquitty, median y centroid), obtenemos los mismos resultados: se detecta a la perfección la especie setosa, pero en un cluster se mezclan las especies virginica y versicolor. Ver Figura 5.

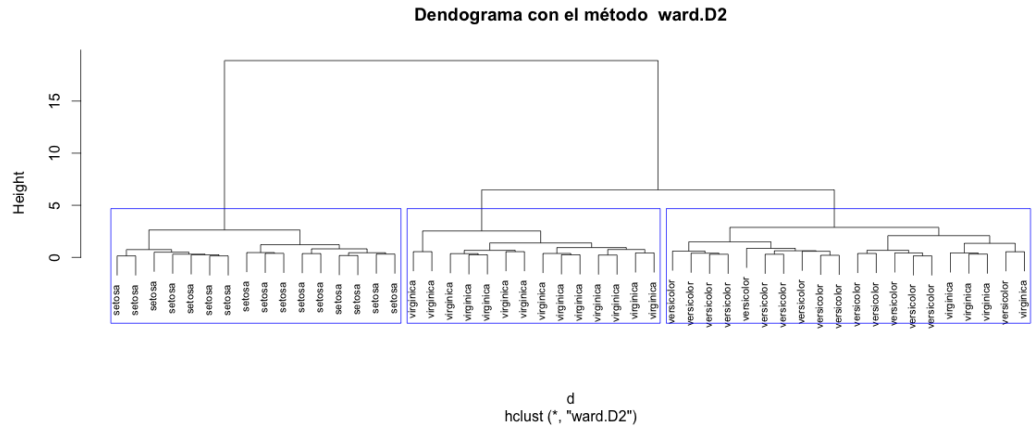
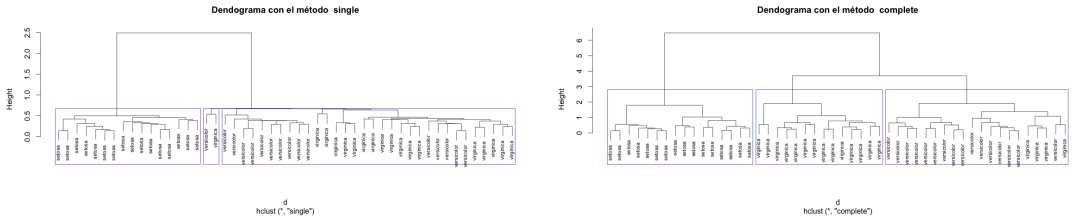
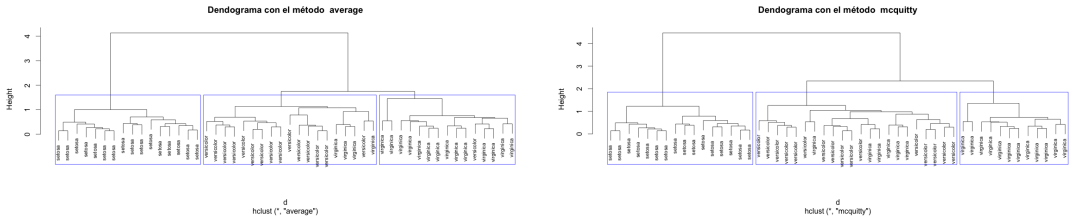


Figura 4: Dendograma con el método ward.D2



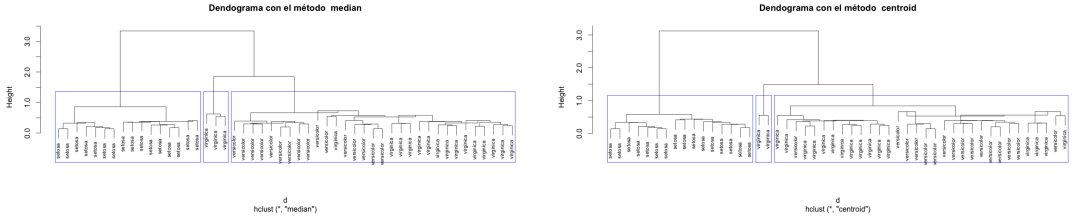
(a) Método single

(b) Método complete



(c) Método average

(d) Método Mcquitty



(e) Método median

(f) Método centroid

Figura 5: Dendogramas con los métodos faltantes

3. Conclusiones

En este caso práctico hemos visto cómo hacer un análisis de clústers. En la primera cuestión hemos identificado países en cinco grupos distintos a partir de datos socioeconómicos como el PBI, la esperanza de vida o la tasa de natalidad. En la segunda cuestión hemos visto cómo obtener dendogramas para obtener tres grupos que se debían corresponder cada uno con un tipo de iris. Dos clusters sí detectaban perfectamente el tipo de iris, pero uno de ellos mezclaba dos tipos de iris.

4. Código R utilizado

```
#-------------------------------------------------------------
# Cuestiones de evaluación
#-------------------------------------------------------------

#-------------------------------------------------------------
# Cuestión 1
#-------------------------------------------------------------

# Cargamos los datos
datos <- read.csv2(file.choose(), header=TRUE,
  row.names=1, dec=",")
datos <- as.matrix(datos)

# Número de países
n <- dim(datos)[1]

# Número de variables
p <- dim(datos)[2]

# Tomamos logaritmos de la variable GPB y
# renombramos la variable a logGNP
datos[,6] <- log(datos[,6])
colnames(datos)[6] <- "logGNP"

# Estandarizamos los datos
datos.st <- scale(datos)

# Aplicamos el algoritmo de las k-medias para 5 clusters
clusters5.datos <- kmeans(datos.st, 5, nstart=25)

# Calculamos los centroides
centroides <- aggregate(datos.st,
  by=list(clusters5.datos$cluster), FUN=mean)

# Vemos los centroides
t(centroides)

# Dibujamos las variables dos a dos
nk <- 5
pairs(datos.st, col=clusters5.datos$cluster, pch=19)
points(clusters5.datos$centers, col=1:nk, pch=19, cex=2)
```

```
# Cargamos la librería cluster
library(cluster)

datos.clusters5 <- clusters5.datos$cluster
clusplot(datos.st, datos.clusters5, color=TRUE, shade=TRUE,
  labels=2, lines=0)

# Países por cluster
grupo_1 = which(clusters5.datos$cluster==1)
grupo_2 = which(clusters5.datos$cluster==2)
grupo_3 = which(clusters5.datos$cluster==3)
grupo_4 = which(clusters5.datos$cluster==4)
grupo_5 = which(clusters5.datos$cluster==5)

#-------------------------------------------------------------
# Cuestión 1
#-------------------------------------------------------------

# Cargamos la dase de datos iris del paquete datasets
library(datsets)

# Dimensiones de los datos. Se encuentran en la variable iris
dim(iris)

# Elegimos 50 datos al azar
ind <- sample(1:150, 50)
iris.cl <- iris[ind, 1:4]

# Guardamos las etiquetas de cada dato
etiquetas <- iris[ind, 5]

# Calculamos el dendograma

# Matriz de distancia de los datos
d <- dist(iris.cl, method="euclidean")

# Calculamos el dendograma con distintos métodos
metodos <- c("ward.D", "ward.D2", "single", "complete",
  "average", "mcquitty", "median", "centroid")

for(metodo in 1:length(metodos)){
  # Usamos el método del vector ''metodos''
```

```
  # definido anteriormente
  fit <- hclust(d, method=metodos[metodo])

  # Representamos el dendograma
  plot(fit, labels=etiquetas, cex=0.7,
    main=paste("Dendograma con el método ", metodos[metodo]))

  # Grupo de cada
  groups <- cutree(fit, k=3)

  # Recuadramos los cluesters
  rect.hclust(fit, k=3, border="blue")
}
```