CURSO DE FORMACIÓN
"MINERÍA DE DATOS CON 'R' "
UNIVERSIDAD DE VIGO

RESOLUCIÓN ORIENTATIVA DE LA TAREA 1

1) Lea el archivo de datos "vinos.RData". Aplique el método k-medias para obtener dos clases utilizando como variables explicativas las concentraciones de los distintos ácidos orgánicos. Relacione la clasificación obtenida con la variedad de uva, con el fin de averiguar hasta que punto esa variedad es un criterio principal de la clasificación "natural" de los vinos o es necesario buscar criterios adicionales (como la zona, el año de la vendimia, u otros).

Cargamos los datos:

load("D:/CURSO DM/vinos.RData")

Es conveniente inspeccionar las variables:

names(vinos)

```
[1] "var" "gal" "tar" "mal" "shi" "cit" "suc"
```

summary(vinos)

```
 var         gal              tar              mal        
 A:35   Min.   :0.0000   Min.   :0.000   Min.   :0.000  
 G:19   1st Qu.:0.4550   1st Qu.:1.555   1st Qu.:1.054  
        Median :0.5695   Median :1.988   Median :2.503  
        Mean   :0.5781   Mean   :1.936   Mean   :2.193  
        3rd Qu.:0.6845   3rd Qu.:2.394   3rd Qu.:3.291  
        Max.   :1.0980   Max.   :3.904   Max.   :4.891  

      shi               cit              suc         
 Min.   :0.00000   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:0.01900   1st Qu.:0.1015   1st Qu.:0.2352  
 Median :0.02850   Median :0.1410   Median :0.3110  
 Mean   :0.02563   Mean   :0.1517   Mean   :0.3506  
 3rd Qu.:0.03375   3rd Qu.:0.1770   3rd Qu.:0.4660  
 Max.   :0.06000   Max.   :0.3700   Max.   :0.7630  
```

Aplicamos el método K-medias con dos clases:

```
set.seed(12345)
modelo <- kmeans(subset(vinos, select = -var), centers = 2)
modelo
```

```
> modelo
K-means clustering with 2 clusters of sizes 28, 26

Cluster means:
        gal      tar       mal        shi       cit       suc
1 0.6236786 2.003143 3.3466429 0.02700000 0.1713214 0.3823214
2 0.5290385 1.863231 0.9507692 0.02415385 0.1306154 0.3163462

Clustering vector:
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27
 1  1  1  1  2  1  2  2  2  2  2  2  1  1  1  2  2  2  1  1  1  1  1  1  1  1  1
28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54
 2  2  1  1  2  2  2  2  1  1  2  1  2  2  2  2  1  1  2  2  1  2  1  2  1  1  2

Within cluster sum of squares by cluster:
[1] 23.68309 20.95674
 (between_SS / total_SS =  63.6 %)
```

k-medias ha construido dos clases homogéneas con 28 y 26 casos respectivamente. Se muestra el valor medio (Cluster means) de cada variable en ambas clases, así como cual es la clase (1/2) asignada a cada uno de los 54 elementos de nuestra muestra (clustering vector). La varianza entre las clases es el 63,6% del total; el 36,4% restante es la varianza interna de las clases (cuanto más pequeña sea la varianza interna, más homogéneas serán las clases).

Veamos ahora si la clasificación está relacionada con la variedad de vino:

```
table(modelo$cluster, vinos$var)
```

```
   A  G
1 21  7
2 14 12
```

21 de los 35 albariños están en la clase 1, y 12 de los 19 godellos en la 2. La clasificación está ligeramente relacionada con la variedad de uva, ya que permite reconocer solamente 33 de los 54 vinos, el 61,1% de la muestra:

```
(21+12)/(21+7+14+12)      #  Se puede utilizar R como calculadora
```

```
[1] 0.6111111
```

2) Lea el conjunto de datos "deudas.RData". Se trata de una muestra de 100 clientes de un banco, algunos de los cuales han presentado impagos, de los que se dispone de información relativa a su nivel de ingresos, relación entre deudas e ingresos, importe de las deudas por tarjeta de crédito, e importe de otras deudas, entre otras variables. Utilice esas 4 variables (columnas 5 a 8) para obtener con el método EM una clasificación con dos grupos o clases. Averigue si la clasificación obtenida está relacionada con la variable "Impago".

(Nota: la parte del conjunto "deudas" que contiene las 4 variables de interés puede representarse como deudas[,5:8])

Leemos los datos:

load("C:/CURSO DM/deudas.RData")

names(deudas)

```
[1] "Edad"       "Formacion"  "Empleo"     "Residencia" "Ingreso"
[6] "Deud_ing"   "Deud_tarj"  "Deud_otr"   "Impago"
```

summary(deudas)

```
   Edad               Formacion              Empleo             Residencia
 entre 30 y 40:36  bachillerato       :28  entre 5 y 10 años:24  entre 5 y 10:21
 mas de 40    :37  elemental          :56  más de 10 años   :41  más de 10   :39
 menos de 30  :27  estudios universit.:16  menos de 5 años  :35  menos de 5  :40

   Ingreso          Deud_ing        Deud_tarj          Deud_otr      Impago
 Min.   : 15.00   Min.   : 0.600   Min.   : 0.0300   Min.   : 0.090   No:68
 1st Qu.: 27.00   1st Qu.: 3.550   1st Qu.: 0.3675   1st Qu.: 0.940   Si:32
 Median : 38.50   Median : 6.400   Median : 0.8700   Median : 1.780
 Mean   : 46.27   Mean   : 8.915   Mean   : 1.2622   Mean   : 2.448
 3rd Qu.: 57.00   3rd Qu.:12.725   3rd Qu.: 1.5225   3rd Qu.: 3.027
 Max.   :176.00   Max.   :35.300   Max.   :11.3600   Max.   :16.670
```

Aplicamos el método EM:

```
library(mclust)       #  carga el paquete mclust.
modelo <-  Mclust(deudas[,5:8], G=2)    # construimos el modelo con 2 grupos
```

Veamos ahora si la clasificación obtenida está relacionada con los impagos:

table(modelo$classification, deudas$Impago)

```
   No Si
 1 17 32
 2 51  0
```

Todos los impagos están en la clase 1; 51 de los 68 clientes que no tienen impago están en la clase 2.

La clasificación obtenida con el método EM parece reconocer razonablemente a los clientes menos solventes, todos ellos incluidos en la clase 1 (la clase 2 no tiene ninguno con impagos). Sin embargo los clientes sin impagos están en ambas clases, aunque mayoritariamente en la clase 2.