Tema9\_Ejercicio\_kmeans

Fran Camacho

2025-03-16

# Tema 9 - Ejercicio de K-Means

La base de datos Acidosis.txt contiene datos de 40 pacientes de un hospital con 6 medidas relativas a la acidez de diferentes fluidos corporales. En particular:

-La columna 1 es el pH del líquido cefalorraquídeo (nanomoles/litro). -La columna 2 es el pH de la sangre (nanomoles/litro). -La columna 3 es la cantidad de HCO3 en el líquido cefalorraquídeo (milimoles/litro). -La columna 4 contiene los valores de HCO3 en sangre (milimoles/litro). -La columna 5 es la presión del CO2 en el líquido cefalorraquídeo (mmHg). -La columna 4 es la presión del CO2 en sangre (mmHg).

Considerando este dataset, obtenga los coeficientes de silueta y obtenga el valor de k previamente a la ejecución del algoritmo K-Means.

Con los clústeres estimados, realice una representación gráfica de los mismos usando fviz\_cluster() e interprete los resultados.

## Paso 1: Carga de los datos

# import the CSV file  
acidosis\_raw <- read.csv(file.path("Chapter09", "Acidosis.txt"), sep = "")

## Paso 2 – Exploración y preparación de los datos

Carga de librerías necesarias.

if (!require(cluster)) install.packages('cluster', dependencies = T)

## Cargando paquete requerido: cluster

library(cluster)  
  
if (!require(factoextra)) install.packages('factoextra', dependencies = T)

## Cargando paquete requerido: factoextra

## Cargando paquete requerido: ggplot2

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

library(factoextra)  
  
if (!require(hms)) install.packages('hms', dependencies = T)

## Cargando paquete requerido: hms

library(hms)

Estructura y resumen estadístico de los datos:

# structure  
str(acidosis\_raw)

## 'data.frame': 40 obs. of 6 variables:  
## $ X1: num 39.8 53.7 47.3 41.7 44.7 47.9 48.4 48.4 48.4 41.7 ...  
## $ X2: num 38 37.2 39.8 37.6 38.5 39.8 36.7 35.1 45.7 81.3 ...  
## $ X3: num 22.2 18.7 23.3 22.8 24.8 22 21 23.9 18.6 9.8 ...  
## $ X4: num 23.2 18.5 22.1 22.3 24.4 23.3 21.3 24 14.9 4.2 ...  
## $ X5: num 38.8 45.1 48.2 41.6 48.5 46.2 44.5 50.6 39.4 17.8 ...  
## $ X6: num 36.5 28.3 36.4 34.6 38.8 38.5 32.6 35 28.8 12.9 ...

# summary  
summary(acidosis\_raw)

## X1 X2 X3 X4   
## Min. :38.50 Min. :32.80 Min. : 9.80 Min. : 4.20   
## 1st Qu.:45.67 1st Qu.:37.20 1st Qu.:19.43 1st Qu.:19.00   
## Median :48.05 Median :39.55 Median :22.65 Median :23.30   
## Mean :47.52 Mean :41.55 Mean :21.91 Mean :22.72   
## 3rd Qu.:49.45 3rd Qu.:41.98 3rd Qu.:24.52 3rd Qu.:26.93   
## Max. :54.90 Max. :81.30 Max. :30.40 Max. :34.80   
## X5 X6   
## Min. :17.80 Min. :12.90   
## 1st Qu.:40.38 1st Qu.:30.25   
## Median :45.25 Median :35.70   
## Mean :45.57 Mean :36.44   
## 3rd Qu.:52.58 3rd Qu.:42.30   
## Max. :69.00 Max. :61.40

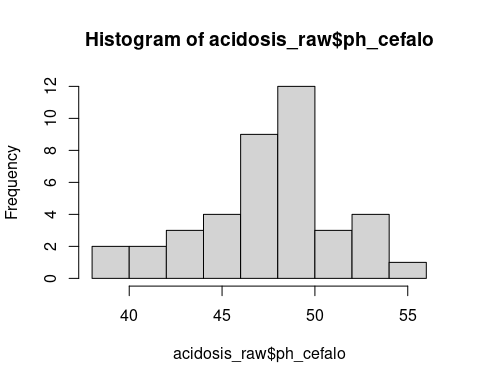
Creo que las variables no difieren demasiado en sus valores (todos están entre 0 y 100). De todas maneras vamos a cambiar los nombres de las variables por unos que sean descriptivos:

# rename column names  
colnames(acidosis\_raw) <- c("ph\_cefalo", "ph\_sangre", "hco3\_cefalo", "hco3\_sangre", "pres\_co2\_cefalo", "pres\_co2\_sangre")

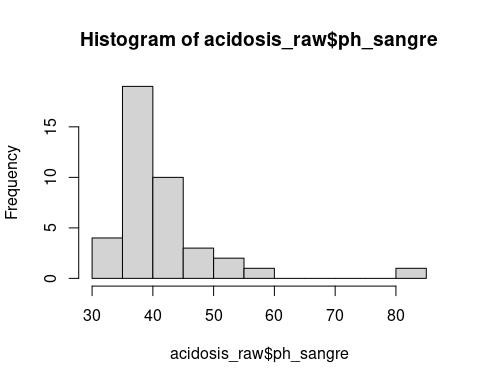
[

Comprobamos las distribuciones de las variables:

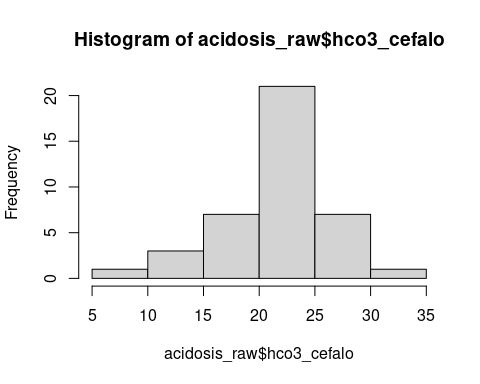
hist(acidosis\_raw$ph\_cefalo)



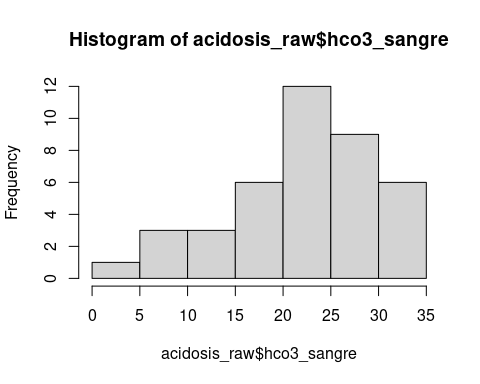
hist(acidosis\_raw$ph\_sangre)



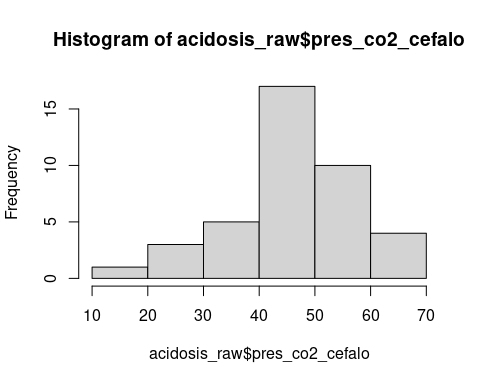
hist(acidosis\_raw$hco3\_cefalo)



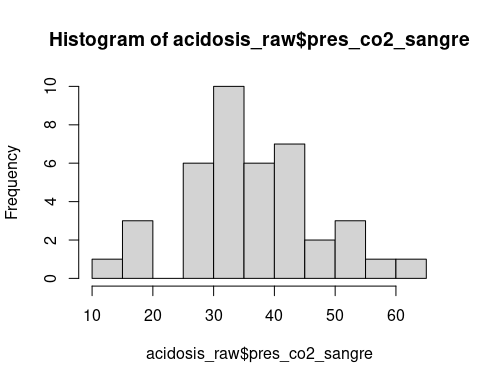
hist(acidosis\_raw$hco3\_sangre)



hist(acidosis\_raw$pres\_co2\_cefalo)



hist(acidosis\_raw$pres\_co2\_sangre)

 ]

Y también aplicamos la función scale():

# scale the dataset  
datos\_scaled <- as.data.frame(scale(acidosis\_raw))  
  
summary(datos\_scaled)

## ph\_cefalo ph\_sangre hco3\_cefalo hco3\_sangre   
## Min. :-2.4062 Min. :-1.01403 Min. :-2.5299 Min. :-2.49793   
## 1st Qu.:-0.4932 1st Qu.:-0.50441 1st Qu.:-0.5200 1st Qu.:-0.50147   
## Median : 0.1400 Median :-0.23223 Median : 0.1535 Median : 0.07858   
## Mean : 0.0000 Mean : 0.00000 Mean : 0.0000 Mean : 0.00000   
## 3rd Qu.: 0.5132 3rd Qu.: 0.04865 3rd Qu.: 0.5450 3rd Qu.: 0.56757   
## Max. : 1.9663 Max. : 4.60340 Max. : 1.7718 Max. : 1.62987   
## pres\_co2\_cefalo pres\_co2\_sangre   
## Min. :-2.41696 Min. :-2.19918   
## 1st Qu.:-0.45215 1st Qu.:-0.57829   
## Median :-0.02785 Median :-0.06913   
## Mean : 0.00000 Mean : 0.00000   
## 3rd Qu.: 0.60968 3rd Qu.: 0.54746   
## Max. : 2.03923 Max. : 2.33184

## Paso 3 - Ejecución del algoritmo k-means

Vamos a de determinar el número óptimo de clústeres, utilizando como se explica en el material complementario la métrica denominada **coeficiente de silueta**, la cual permite evaluar la calidad del agrupamiento obtenido.

El valor de este coeficiente es un valor comprendido entre -1 y +1. Cuanto más cercano sea su valor a +1, mejor será la calidad del agrupamiento.

Creamos un pequeño bucle variando el parámetro k y obtenemos los coeficientes de silueta:

# coeficientes de silueta  
set.seed(1234)  
  
d <- dist(datos\_scaled) # This function computes and returns the distance matrix computed   
 # by using the specified distance measure to compute the distances between the rows of a data matrix.  
#d

[

Output:

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

2 3.9453991  
3 2.1877121 2.1923064  
4 0.6165141 3.4165874 1.6321668  
5 1.6703134 3.0143345 0.8651709 1.1915932  
6 2.2713136 2.0720159 0.4407694 1.7710169 1.0748526  
7 2.4056835 1.5929585 0.8288452 1.8600744 1.5092490 0.7673272  
8 2.5655752 2.0991379 0.7242077 1.9987717 1.1514652 0.8567462 0.9305192  
9 2.8968229 1.8575578 1.8842250 2.4833464 2.5469819 1.8558759 1.5511897 2.3504904  
…

]

set.seed(911)  
  
avgS <- c()  
  
for(k in 2:9) {  
 cl <- kmeans(datos\_scaled, centers = k, iter.max = 200)  
 s <- silhouette(cl$cluster, d)  
 avgS <- c(avgS, mean(s[,3]))  
}  
  
data.frame(nClus = 2:9, Silh = avgS)

## nClus Silh  
## 1 2 0.5171241  
## 2 3 0.3890987  
## 3 4 0.2849339  
## 4 5 0.3201198  
## 5 6 0.3288259  
## 6 7 0.3010176  
## 7 8 0.3088594  
## 8 9 0.2922981

El resultado obtenido muestra que la mejor opción se da para k=2, por ser esta opción la que presenta un coeficiente de silueta más elevado. (Es positivo, y está a medio camino entre el 0 y el 1 …)

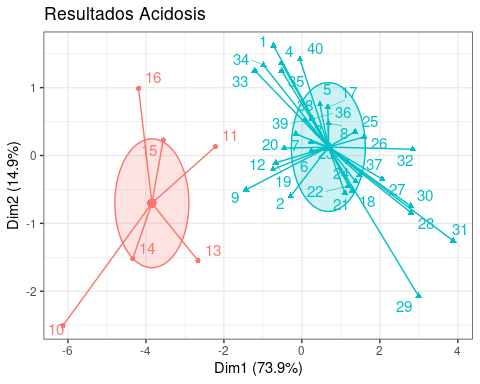
Habiendo determinado el número óptimo de clústeres, pasamos a ejecutar el algoritmo K-Means, especificando centers=2 para establecer el valor del número de clústeres:

#   
clusters <- kmeans(datos\_scaled, centers = 2, nstart = 50)  
  
clusters

## K-means clustering with 2 clusters of sizes 6, 34  
##   
## Cluster means:  
## ph\_cefalo ph\_sangre hco3\_cefalo hco3\_sangre pres\_co2\_cefalo pres\_co2\_sangre  
## 1 -0.8509424 1.6595567 -1.8163916 -1.7852282 -1.7569480 -1.5094034  
## 2 0.1501663 -0.2928629 0.3205397 0.3150403 0.3100497 0.2663653  
##   
## Clustering vector:  
## [1] 2 2 2 2 2 2 2 2 2 1 1 2 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2  
## [39] 2 2  
##   
## Within cluster sum of squares by cluster:  
## [1] 23.52067 102.26958  
## (between\_SS / total\_SS = 46.2 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss" "tot.withinss"  
## [6] "betweenss" "size" "iter" "ifault"

Pasamos a representar gráficamente los clústeres obtenidos. Para ello, usaremos la función fviz\_cluster() incluida en la librería factoextra:

#   
fviz\_cluster(object = clusters, data = datos\_scaled, show.clust.cent = TRUE, ellipse.type = "euclid", star.plot = TRUE, repel = TRUE) +  
 labs(title = "Resultados Acidosis") +  
 theme\_bw() +  
 theme(legend.position = "none")



Se ve que hay dos grupos de pacientes, uno pequeño con solo 6 miembros, para el que todas las variables a excepción del ph en sangre tienen valores inferiores a los del otro grupo más numeroso.

Dado que no es un grupo numeroso, podemos seleccionarlos para examinarlos con más detalle:

# group 1  
acidosis\_raw[c(10,11,13,14,15,16), ]

## ph\_cefalo ph\_sangre hco3\_cefalo hco3\_sangre pres\_co2\_cefalo pres\_co2\_sangre  
## 10 41.7 81.3 9.8 4.2 17.8 12.9  
## 11 46.2 42.7 15.5 15.0 31.3 26.8  
## 13 49.6 55.0 14.6 11.3 31.8 25.7  
## 14 47.3 59.4 10.4 7.5 21.5 18.6  
## 15 42.7 49.0 15.3 9.5 26.9 19.0  
## 16 38.5 47.9 13.7 9.4 23.0 18.7

También podemos comparar los valores estadísticos para ambos grupos:

# group 1 summary  
summary(acidosis\_raw[c(10,11,13,14,15,16), ])

## ph\_cefalo ph\_sangre hco3\_cefalo hco3\_sangre   
## Min. :38.50 Min. :42.70 Min. : 9.80 Min. : 4.200   
## 1st Qu.:41.95 1st Qu.:48.17 1st Qu.:11.22 1st Qu.: 7.975   
## Median :44.45 Median :52.00 Median :14.15 Median : 9.450   
## Mean :44.33 Mean :55.88 Mean :13.22 Mean : 9.483   
## 3rd Qu.:47.02 3rd Qu.:58.30 3rd Qu.:15.12 3rd Qu.:10.850   
## Max. :49.60 Max. :81.30 Max. :15.50 Max. :15.000   
## pres\_co2\_cefalo pres\_co2\_sangre  
## Min. :17.80 Min. :12.90   
## 1st Qu.:21.88 1st Qu.:18.62   
## Median :24.95 Median :18.85   
## Mean :25.38 Mean :20.28   
## 3rd Qu.:30.20 3rd Qu.:24.02   
## Max. :31.80 Max. :26.80

# group 2 summary  
summary(acidosis\_raw[-c(10,11,13,14,15,16), ])

## ph\_cefalo ph\_sangre hco3\_cefalo hco3\_sangre   
## Min. :39.80 Min. :32.80 Min. :18.60 Min. :14.90   
## 1st Qu.:46.30 1st Qu.:36.83 1st Qu.:21.93 1st Qu.:21.88   
## Median :48.40 Median :38.55 Median :22.90 Median :24.60   
## Mean :48.09 Mean :39.03 Mean :23.45 Mean :25.05   
## 3rd Qu.:49.85 3rd Qu.:40.88 3rd Qu.:24.75 3rd Qu.:28.20   
## Max. :54.90 Max. :53.80 Max. :30.40 Max. :34.80   
## pres\_co2\_cefalo pres\_co2\_sangre  
## Min. :38.60 Min. :26.80   
## 1st Qu.:43.23 1st Qu.:32.52   
## Median :46.90 Median :37.95   
## Mean :49.13 Mean :39.29   
## 3rd Qu.:53.40 3rd Qu.:43.05   
## Max. :69.00 Max. :61.40

Ahora ya sólo quedaría preguntarle a un médico qué hacer con estos pacientes :)

NOTA:

He investigado un poco sobre el ph en sangre. Quería convertir los valores de esa columna en valores de 0 a 14. He visto que los valores entre 20 y 80 nanomoles/l se corresponden con un ph entre 7.1 y 7.7, que es un nivel normal.

<https://acid-base.com/ph-playground>

ph <- -log10( (acidosis\_raw$ph\_sangre)\*1e-9)  
ph

## [1] 7.420216 7.429457 7.400117 7.424812 7.414539 7.400117 7.435334 7.454693  
## [9] 7.340084 7.089909 7.369572 7.374688 7.259637 7.226214 7.309804 7.319664  
## [17] 7.442493 7.402305 7.396856 7.383000 7.389340 7.383000 7.403403 7.388277  
## [25] 7.414539 7.420216 7.377786 7.404504 7.269218 7.379864 7.358526 7.413413  
## [33] 7.429457 7.469800 7.449772 7.469800 7.471083 7.428291 7.438899 7.484126

Entiendo por tanto que todos estos pacientes tienen un ph en sangre correcto.

#ph<7.1  
acidosis\_raw[ph<7.1, ]

## ph\_cefalo ph\_sangre hco3\_cefalo hco3\_sangre pres\_co2\_cefalo pres\_co2\_sangre  
## 10 41.7 81.3 9.8 4.2 17.8 12.9

El paciente número 10 es precisamente el que sale en el gráfico en la esquina inferior izquierda, muy separado de todos los demás. Incluso de los otros miembros de su clúster.

Serán los otros indicadores los que indiquen algún tipo de problema … (sobre todo para este paciente número 10).