K-Medias

Omar Sanchez Hernandez

4/6/2022

INTRODUCCIÓN

El metodo de agrupación K-medias pertenece al conjunto de metodos de aprendizaje **NO** supervisados el cual agrupa a los sujetos de la matriz de datos en K grupos dependiendo en base a sus similitudes.

En el presente ejercicio se llevara a cabo el desarrollo de esta tecnica usando la matriz de datos **state.x77** que se encuentra en la libreria de ***datasets*** que contiene informacion sobre indicadores de 50 estados de Estados Unidos.

Matriz de datos.

```
# Librerias
library(cluster)
library(knitr)
X<-as.data.frame(state.x77)
head(X)
```

```
##            Population Income Illiteracy Life Exp Murder HS Grad Frost   Area
## Alabama          3615   3624        2.1    69.05   15.1    41.3    20  50708
## Alaska            365   6315        1.5    69.31   11.3    66.7   152 566432
## Arizona          2212   4530        1.8    70.55    7.8    58.1    15 113417
## Arkansas         2110   3378        1.9    70.66   10.1    39.9    65  51945
## California      21198   5114        1.1    71.71   10.3    62.6    20 156361
## Colorado         2541   4884        0.7    72.06    6.8    63.9   166 103766
```

Sobre la matriz de datos

```
# Librerias
dim(X)
```

```
## [1] 50  8
```

```
str(X)
```

```
## 'data.frame':    50 obs. of  8 variables:
##  $ Population: num  3615 365 2212 2110 21198 ...
##  $ Income    : num  3624 6315 4530 3378 5114 ...
##  $ Illiteracy: num  2.1 1.5 1.8 1.9 1.1 0.7 1.1 0.9 1.3 2 ...
##  $ Life Exp  : num  69 69.3 70.5 70.7 71.7 ...
##  $ Murder    : num  15.1 11.3 7.8 10.1 10.3 6.8 3.1 6.2 10.7 13.9 ...
##  $ HS Grad   : num  41.3 66.7 58.1 39.9 62.6 63.9 56 54.6 52.6 40.6 ...
##  $ Frost     : num  20 152 15 65 20 166 139 103 11 60 ...
##  $ Area      : num  50708 566432 113417 51945 156361 ...
```

La matriz de datos contiene 50 casos, 8 variables y todas las variables son de tipo numericas.

Transformacion de datos

Debido a las grandes diferencias entre las escalas de la matriz de datos es que se propone realizar una transformación logartimica sobre las variables 1,3 y 8.

```
X[,1]<-log(X[,1])
colnames(X)[1]<-"Log-Population"

X[,3]<-log(X[,3])
colnames(X)[3]<-"Log-Illiteracy"

X[,8]<-log(X[,8])
colnames(X)[8]<-"Log-Area"
```

Metodo k-means

```
dim(X)
```

```
## [1] 50  8
```

```
n<-dim(X)[1]
p<-dim(X)[2]
```

Estandarizacion univariante.

Se aplica una estandarización a la matriz para homogeneizarla y que las variables con valores muy altos o valores muy bajos no tengan problemas por el tipo de escala.

```
X.s<-scale(X)
```

Aplicación del algoritmo k-medias

Se plantean 3 grupos (cantidad de subconjuntos **K**) aleatorios que se escogen para realizar los calculos de algoritmo y luego se muestran los centroides de cada grupo.

```
Kmeans.3<-kmeans(X.s, 3, nstart=25)
Kmeans.3$centers
```

```
##   Log-Population     Income Log-Illiteracy   Life Exp      Murder    HS Grad
## 1      0.5693805  0.5486843     0.05412021  0.1388564 -0.01977495  0.1203417
## 2     -0.7900149  0.2080926    -0.93960948  0.5642988 -0.71791785  0.7707484
## 3      0.2360549 -1.2266128     1.31921387 -1.0778757  1.10983501 -1.3566922
##        Frost   Log-Area
## 1 -0.3291597 -0.4878988
## 2  0.8803670  0.4093602
## 3 -0.7719510  0.1991243
```

Cluster de pertenencia

```
Kmeans.3$cluster
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              3              2              1              3              1
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              1              1              1              3
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              1              2              1              1              2
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              2              3              3              2              1
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              1              1              2              3              1
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              2              2              2              2              1
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              3              1              3              2              1
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              1              2              1              1              3
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              2              3              3              2              2
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              1              1              3              2              2
```

SCDG

```
SCDG<-sum(Kmeans.3$withinss)
SCDG
```

```
## [1] 203.2068
```

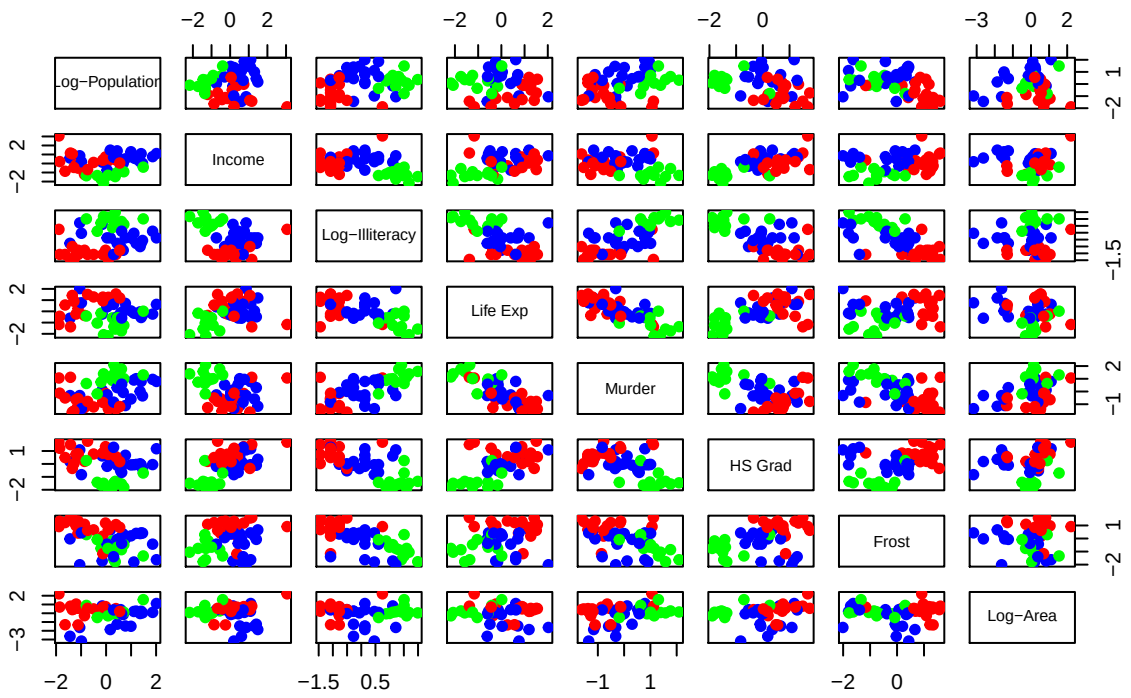
Clusters

```
cl.kmeans<-Kmeans.3$cluster
cl.kmeans
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              3              2              1              3              1
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              1              1              1              3
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              1              2              1              1              2
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              2              3              3              2              1
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              1              1              2              3              1
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              2              2              2              2              1
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              3              1              3              2              1
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              1              2              1              1              3
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              2              3              3              2              2
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              1              1              3              2              2
```

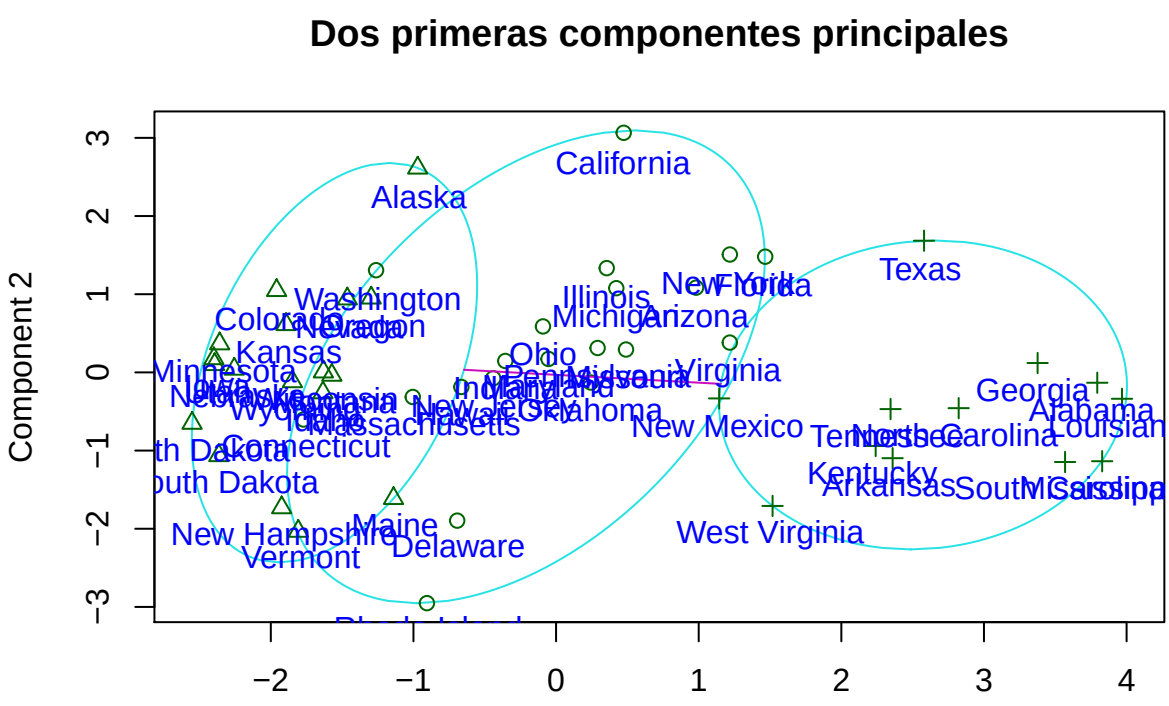
Scatter plot por grupos (delimitados por el K-medias)

```
col.cluster<-c("blue", "red", "green")[cl.kmeans]
pairs(X.s, col=col.cluster, main="k-means", pch=19)
```



Visualizacion con las dos componentes principales

```
clusplot(X.s, cl.kmeans,
         main="Dos primeras componentes principales")
text(princomp(X.s)$score[,1:2],
     labels=rownames(X.s), pos=1, col="blue")
```



These two components explain 62.5 % of the point variability.

Silhouette

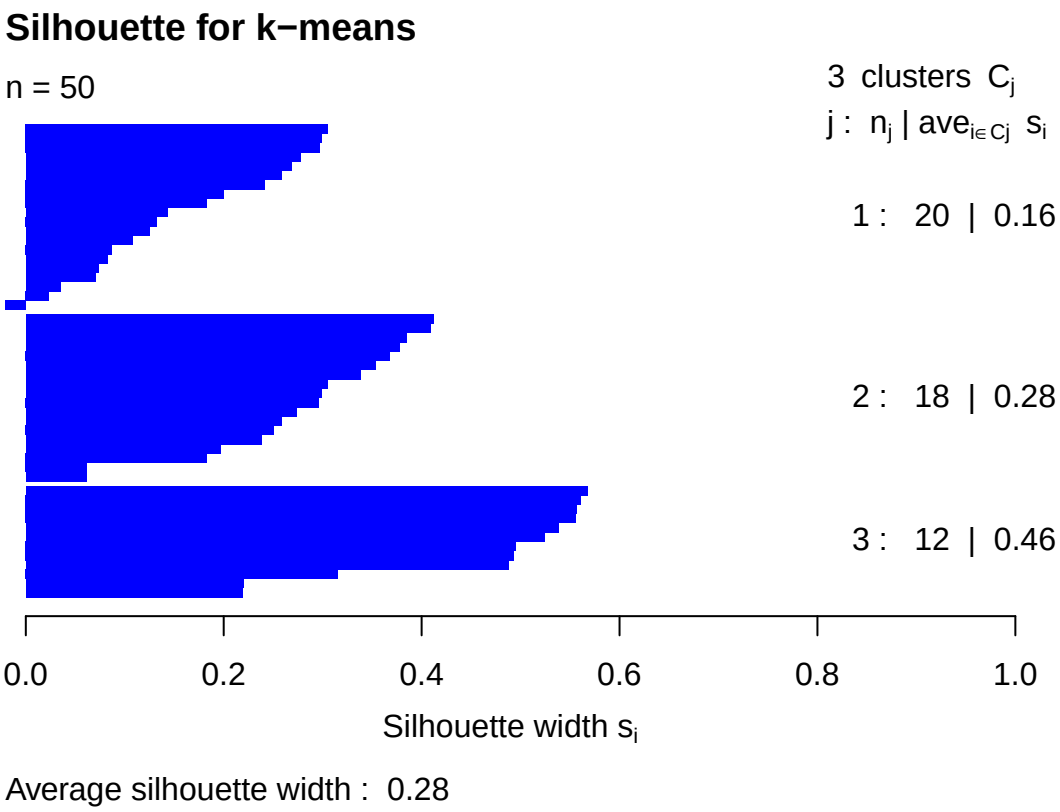
Representacion grafica de la eficacia de clasificacion de una observacion dentro de un grupo.

Generacion de los calculos

```
dist.Euc<-dist(X.s, method = "euclidean")
Sil.kmeans<-silhouette(cl.kmeans, dist.Euc)
```

Generacion del grafico

```
plot(Sil.kmeans, main="Silhouette for k-means",
col="blue")
```



Average silhouette width : 0.28