K-MEANS

Saraí Campos Varela

2022-06-01

INTRODUCCIÓN

K-medias es un método de agrupamiento, que tiene como objetivo la partición de un conjunto de n observaciones en k grupos en el que cada observación pertenece al grupo cuyo valor medio es más cercano. Es un método utilizado en minería de datos. k-medias tiende a encontrar grupos de extensión espacial comparable, mientras que el mecanismo expectation-maximization permite que los grupos tengan formas diferentes.

Librerias

```
library(cluster)
```

Matriz de datos.

```
X<-as.data.frame(state.x77)
colnames(X)
```

```
## [1] "Population" "Income"     "Illiteracy" "Life Exp"   "Murder"
## [6] "HS Grad"    "Frost"      "Area"
```

Transformacion de datos

1.- Transformacion de las variables x1,x3 y x8 con la funcion de logaritmo.

```
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]
```

2.- Estandarizacion univariante.

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

3.- Algoritmo k-medias (2 grupos) cantidad de subconjuntos aleatorios que se escogen para realizar los calculos de algoritmo.

```
Kmeans.2<-kmeans(X.s, 2, nstart=25)
```

Centroides

```
Kmeans.2$centers
```

```
##   Log-Population     Income Log-Illiteracy   Life Exp     Murder    HS Grad
## 1      0.3921592 -0.7973132      1.1635825 -0.8863645  0.9913208 -1.0270524
## 2     -0.1845455  0.3752062     -0.5475682  0.4171127 -0.4665039  0.4833188
##        Frost   Log-Area
## 1 -0.8493032  0.2164565
## 2  0.3996721 -0.1018619
```

Cluster de pertenencia

```
Kmeans.2$cluster
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              2              1              1              2
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              2              2              1              1
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              2              2              2              2              2
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              2              1              1              2              2
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              2              2              2              1              2
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              2              2              2              2              2
```

```
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              1              1              1              2              2
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              2              2              2              2              1
##   South Dakota      Tennessee          Texas          Utah        Vermont
##              2              1              1              2              2
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              1              2              1              2              2
```

4.- SCDG

```
SCDG<-sum(Kmeans.2$withinss)
SCDG
```

```
## [1] 257.0639
```

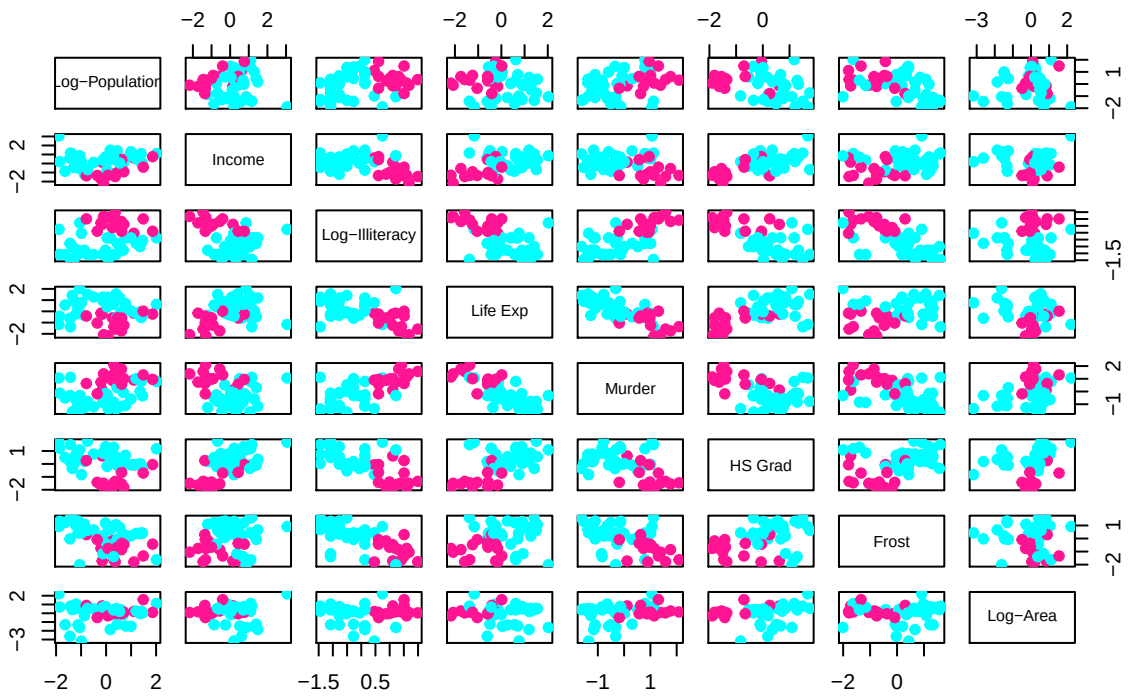
5.- Clusters

```
cl.kmeans<-Kmeans.2$cluster
cl.kmeans
```

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

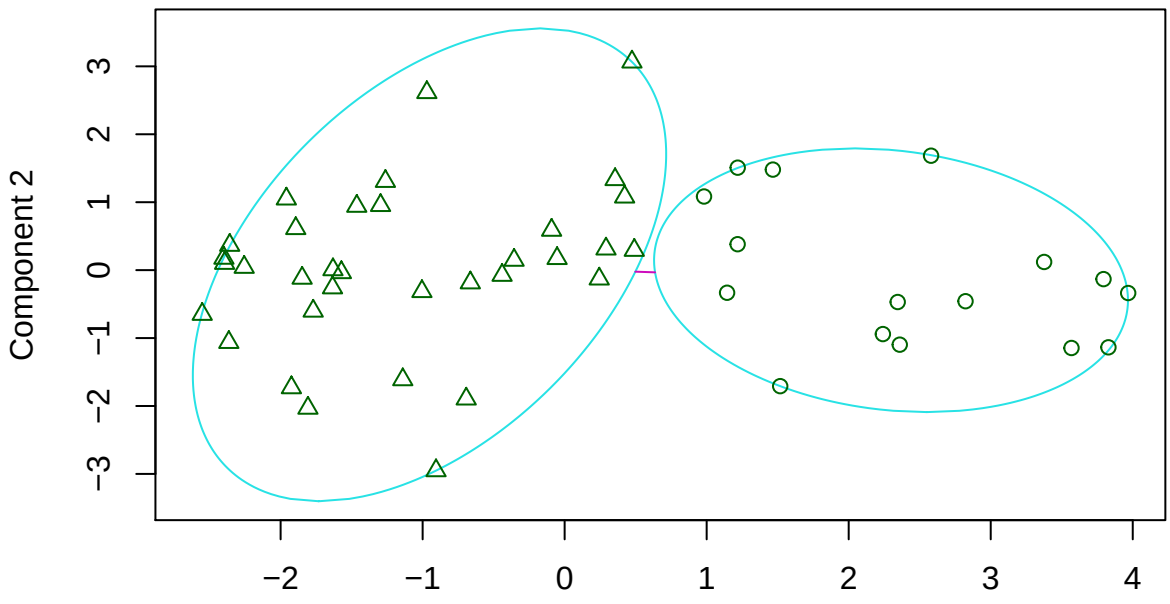
6.- Scatter plot con la division de grupos obtenidos (se utiliza la matriz de datos centrados).

```
col.cluster<-c("deeppink1", "cyan")[cl.kmeans]
pairs(X.s, col=col.cluster, main="k-means", pch=19)
```

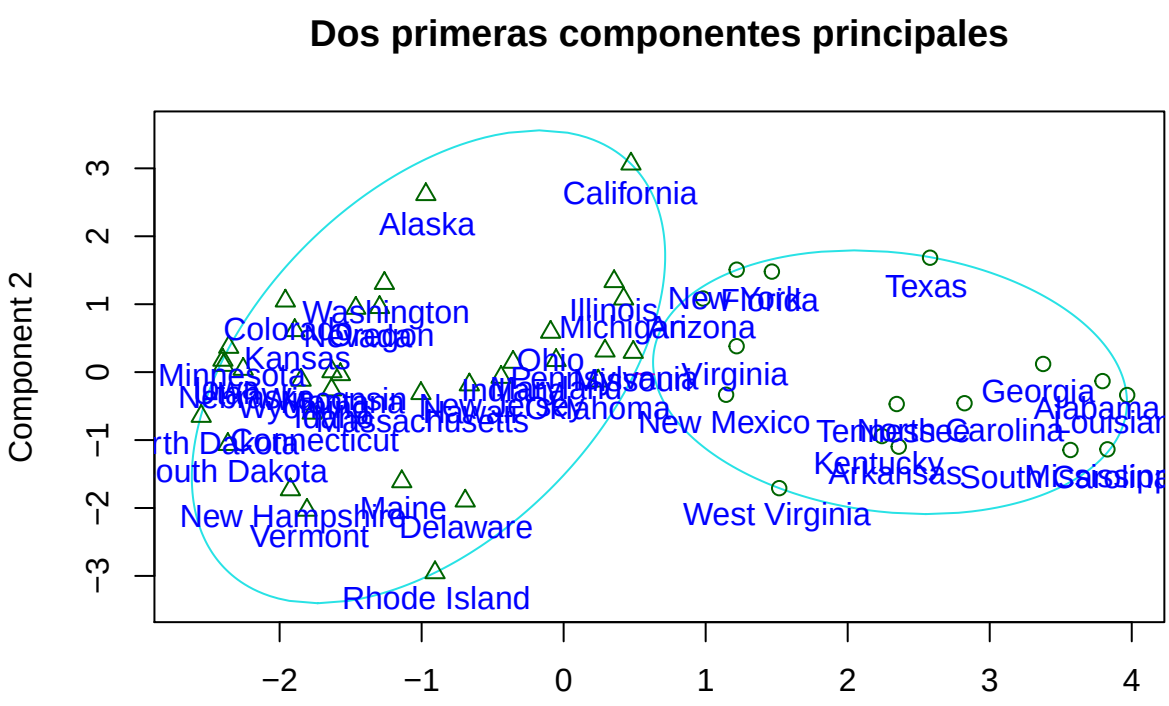


Visualizacion con las dos componentes principales

```
clusplot(X.s, cl.kmeans,
         main="Dos primeras componentes principales")
```




```
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

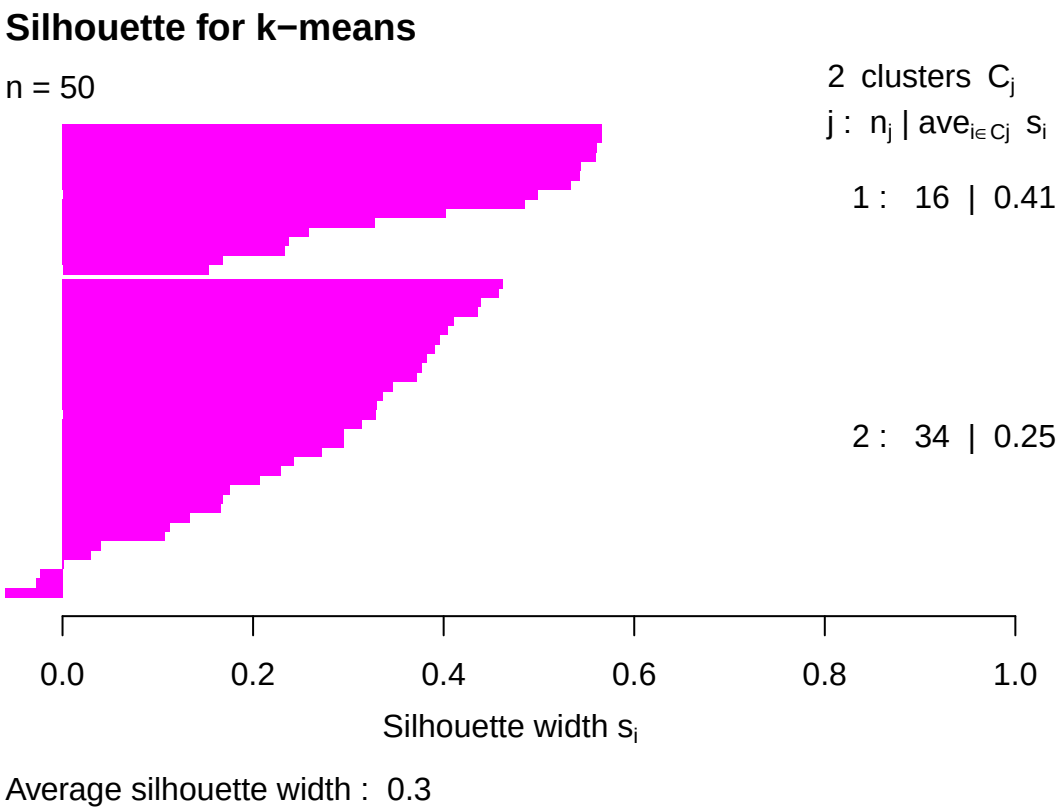
Representacion grafica de la eficacia de clasificacion de una observacion dentro de un grupo.

1.- Generacion de los calculos

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

2.- Generacion del grafico

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



INTERPRETACION

Se eligío tomar unicamente dos grupos ya que al elaborar el gráfico podemos notar que las variables se encuentran en su mayoria dentro de su grupo correspondiente. A pesar de que se presentan variables negativas, son minimas, creando mayor numero de grupos notamos que se presentan mayor numero de variables negativas.