k-means

Irma Eunice Martínez de la Cruz

2022-05-26

K-MEANS

Cargar la matriz de datos "state.x77"

```
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

1.- Separacion de filas y columnas.

```
dim(X)
```

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

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

2.- Estandarizacion univariante.

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

3.- Algoritmo k-medias (3 grupos)

nstart=cantidad de subconjuntos aleatorios que se escogen para realizar los calculos de algoritmo.

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

Centroides

```
Kmeans.3$centers
```

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

Cluster de pertenencia

```
Kmeans.3$cluster
```

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

4.- SCDG

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

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

5.- Clusters

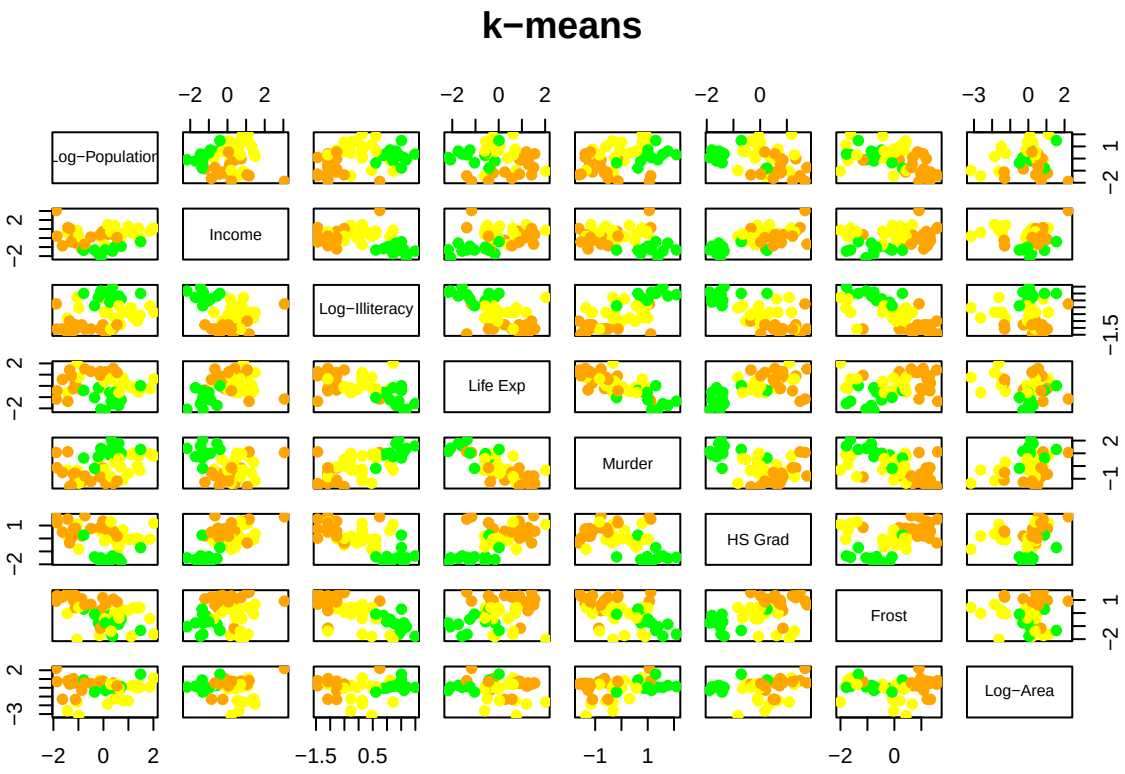
```
cl.kmeans<-Kmeans.3$cluster
cl.kmeans
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              3              2              1              2
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              3              2              2              2              1
```

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

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

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

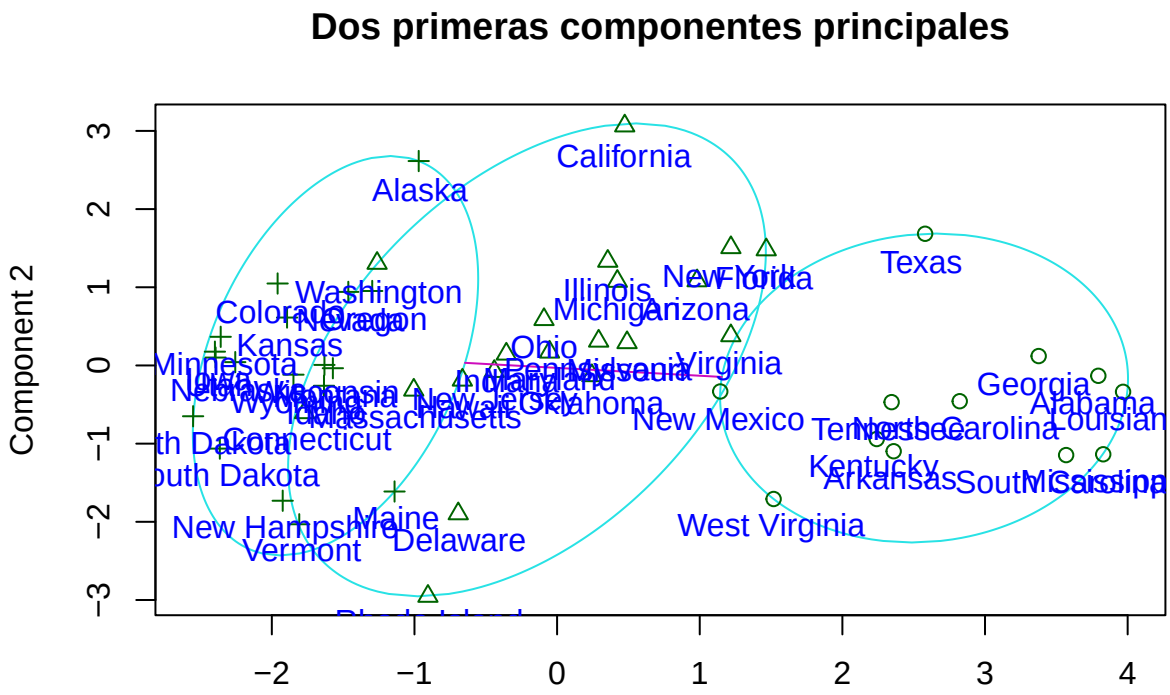


Visualizacion con las dos componentes principales

```
install.packages("cluster")
library(cluster)
```

```
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

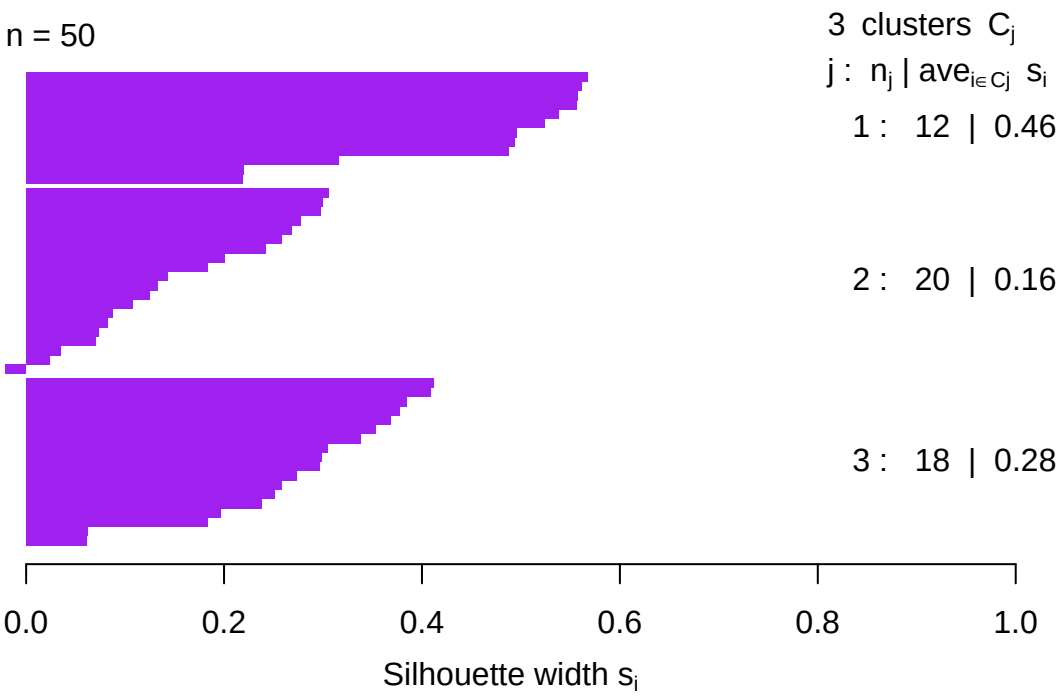
Representacion grafica de la eficacia de clasificacion de una observacion dentro de un grupo.

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="purple")
```



EJERCICIO

SE REPLICA EL SCRIPT PERO SE VA A SUGERIRA UN NÚMERO DE CLUSTERS DIFERENTE A 3 Y 1

INCLUIR LA INTERPRETACION DEL SILHOUETTE

#_____ K-MEANS_____

Cargar la matriz de datos "state.x77"

```
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 x5 y x8 con la funcion de logaritmo.

```
X[,5]<-log(X[,5])
colnames(X)[5]<-"Log-Murder"

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

Metodo k-means

1.- Separacion de filas y columnas.

```
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)

nstart=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
```

```
##     Population     Income Illiteracy   Life Exp Log-Murder    HS Grad
## 1  0.002728194  0.3862098 -0.5320231  0.3982013 -0.3814467  0.4680664
## 2 -0.006365787 -0.9011561  1.2413873 -0.9291364  0.8900424 -1.0921550
##        Frost    Log-Area
## 1  0.3759077 -0.09618613
## 2 -0.8771179  0.22443430
```

Cluster de pertenencia

```
Kmeans.2$cluster
```

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

4.- SCDG

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

```
## [1] 262.9755
```

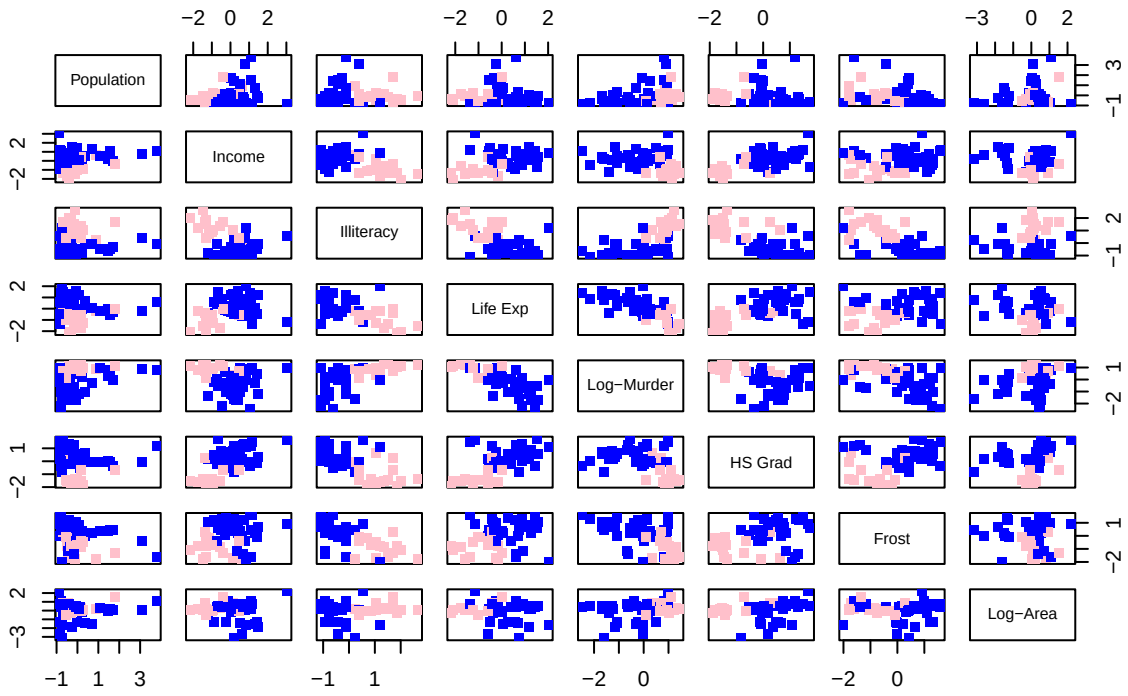
5.- Clusters

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

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

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

```
col.cluster<-c("blue", "pink")[cl.kmeans]
pairs(X.s, col=col.cluster, main="k-means", pch=15)
```

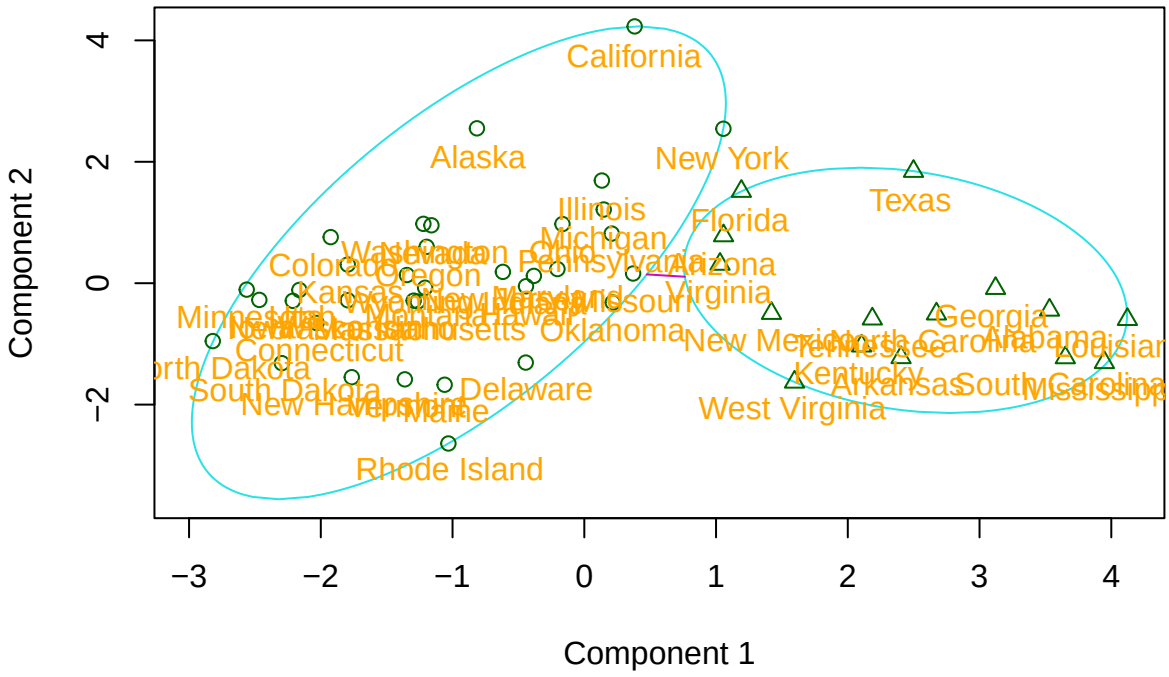


Visualizacion con las dos componentes principales

```
install.packages("cluster")
library(cluster)
```

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

text(princomp(X.s)$score[,1:2],
     labels=rownames(X.s), pos=1, col="orange")
```



These two components explain 62.92 % of the point variability.

Silhouette

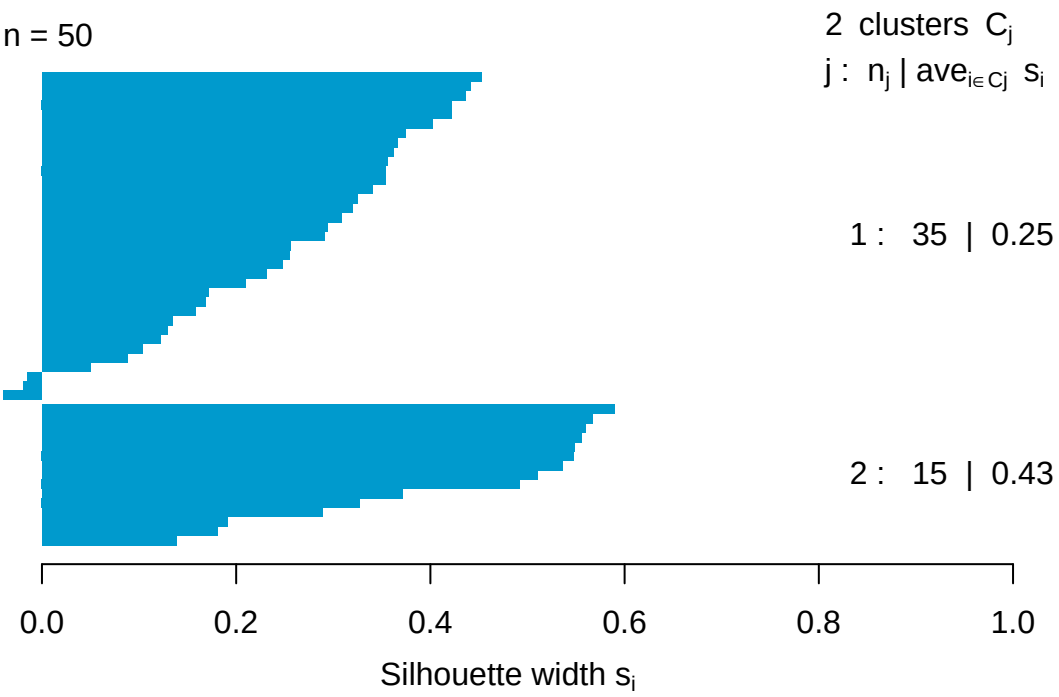
Representacion grafica de la eficacia de clasificacion de una observacion dentro de un grupo.

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="deepskyblue3")
```



Average silhouette width : 0.31