k-Medias Proyecto final

Edgar Ortiz Mota

2022-05-27

Introducción

El metodo de K-Medias se basa en dividir los datos en un número prefijado de grupos, estos van a estar definidos por un centroide, cuya distancia con los puntos será lo menor posible,el resultado del algoritmo depende de la asignación inicial y del orden de los elementos. Por eso siempre conviene repetir el algoritmo con distintos valores iniciales y permutando los elementos de la muestra.

Es importante mencionar que cada vez que se introduce un nuevo elemento en un grupo se recalculan las coordenadas de la nueva media de grupo.

De manera general el metodo de K-medias se basa en:

- Determinar un número k de centroides al azar
- Calcular las distancias de cada uno de los puntos con los centroides y se asignará al centroide cuya distancia sea menor.
- Se ira rotando hasta que cada uno de los puntos sea asignado a a un centroide.
- Estos pasos se repetiran hasta que los centroides no cambien de posición y por lo tanto la asignación de los puntos no cambie.

Se decidio adaptar esta matriz de datos a K-medias, ya que es un metodo que me parecio intereante, la forma en que se agrupan los datos y como es que se ve visualmente ademas de que fue un tema que entendí facilmente y comprendí su metodología rapidamente.

Para esto se utilizó la base de datos llamada USArrests, esta se encuentra cargada en R, en ella se encuentra el número de arrestos por cada 100,000 residentes en cada estado de EE. UU. En 1973 por asesinato , asalto y violación junto con el porcentaje de la población en cada estado que vive en áreas urbanas. , UrbanPop.

Metodología

Exploración de la matriz

```
data("USArrests")
arrestos =USArrests
```

Como se puede observar nuestra matriz tiene una dimensión de 50x4

```
dim(arrestos)
```

```
## [1] 50  4
```

El nombre de nuestras variables son los siguientes:

```
names(arrestos)
```

```
## [1] "Murder"   "Assault"  "UrbanPop" "Rape"
```

Nuestras variables son numericas

```
str(arrestos)
```

```
## 'data.frame':    50 obs. of  4 variables:
##  $ Murder  : num  13.2 10 8.1 8.8 9 7.9 3.3 5.9 15.4 17.4 ...
##  $ Assault : int  236 263 294 190 276 204 110 238 335 211 ...
##  $ UrbanPop: int  58 48 80 50 91 78 77 72 80 60 ...
##  $ Rape    : num  21.2 44.5 31 19.5 40.6 38.7 11.1 15.8 31.9 25.8 ...
```

No se encontraron datos faltantes o NA

```
anyNA(arrestos)
```

```
## [1] FALSE
```

Tratamiento de la matriz

se utilizó el metodo de k-medias para tratar esta matriz de datos

1.- Separamos filas y columnas

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

2.- Estandarizamos los datos

```
a.s<-scale(arrestos)
```

3.-Algoritmo k-medias (2 grupos)

nstar Es la cantidad de subconjunto aleatorios que se escojen para realizar los calculos de algoritmo.

2 Es el numero de clúster o de agrupaciones

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

Centroides

```
Kmeans.2$centers
```

```
##      Murder  Assault UrbanPop     Rape
## 1  4.841379 109.7586 64.03448 16.24828
## 2 11.857143 255.0000 67.61905 28.11429
```

cluster de pertenencia

```
Kmeans.2$cluster
```

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

Aquí se puede observar como es que cada dato esta asignado a algún grupo, ya se 1 o 2

4.- suma de cuadrados dentro del grupo

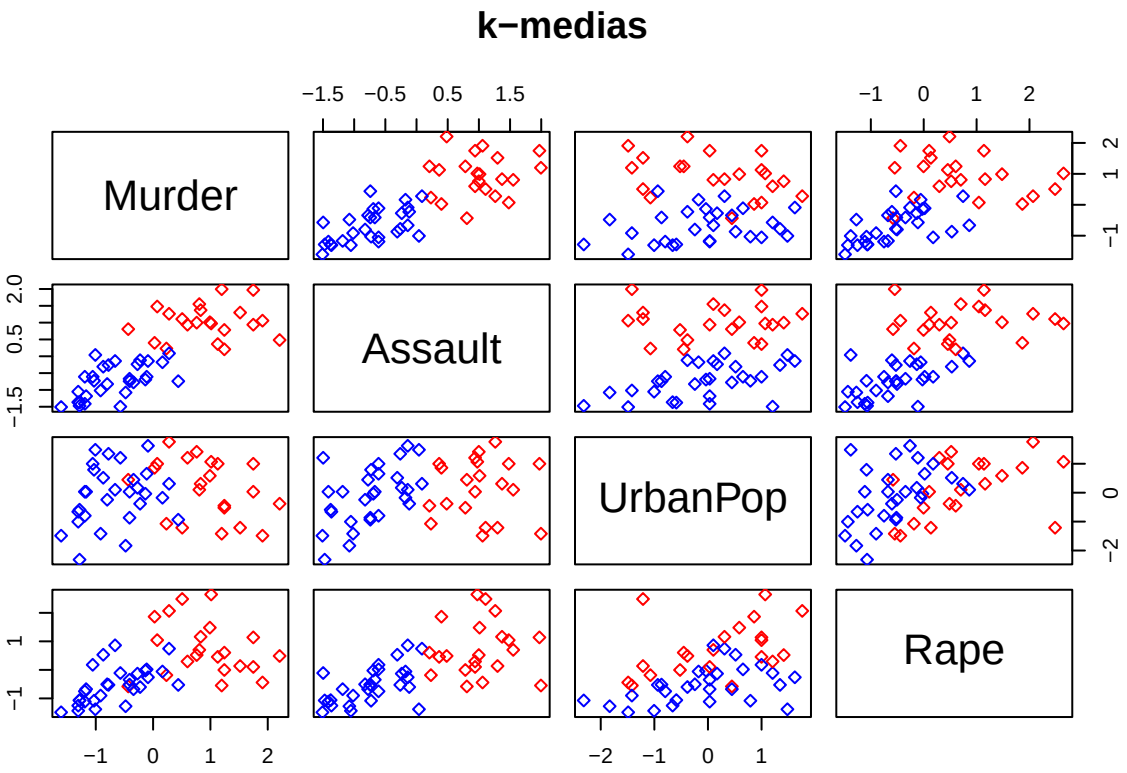
```
SCDG<-sum(Kmeans.2$withinss)
```

5.- Clusters

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

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

```
col.cluster<-c("blue", "red")[cl.kmeans]
pairs(a.s, col=col.cluster, main="k-medias", pch=23)
```

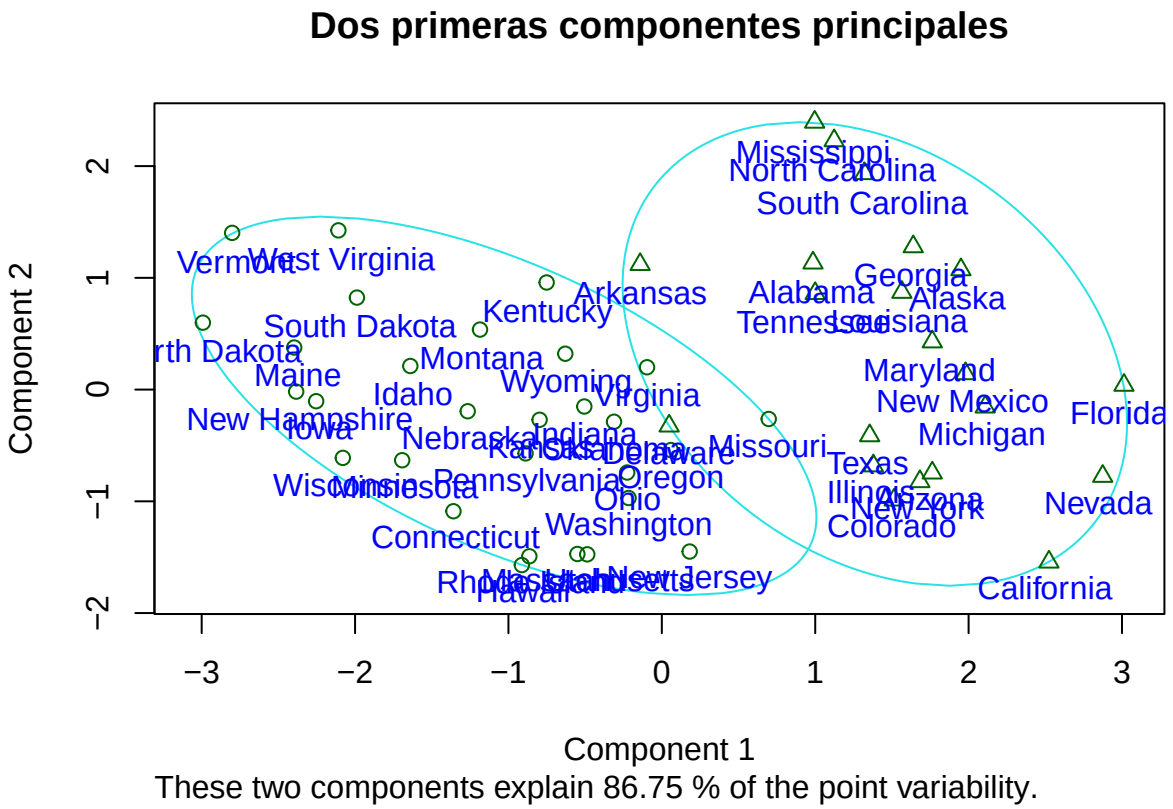


Visualización con las dos componentes principales

```
library(cluster)

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

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



En este gráfico podemos observar como es que se agrupan muy bien la mayoria de los datos, sin embargo tenemos algunos datos que nos afectara nuestro agrupamiento, como es el caso de "Missouri","Virginia" y "Delawer".

Silhouette

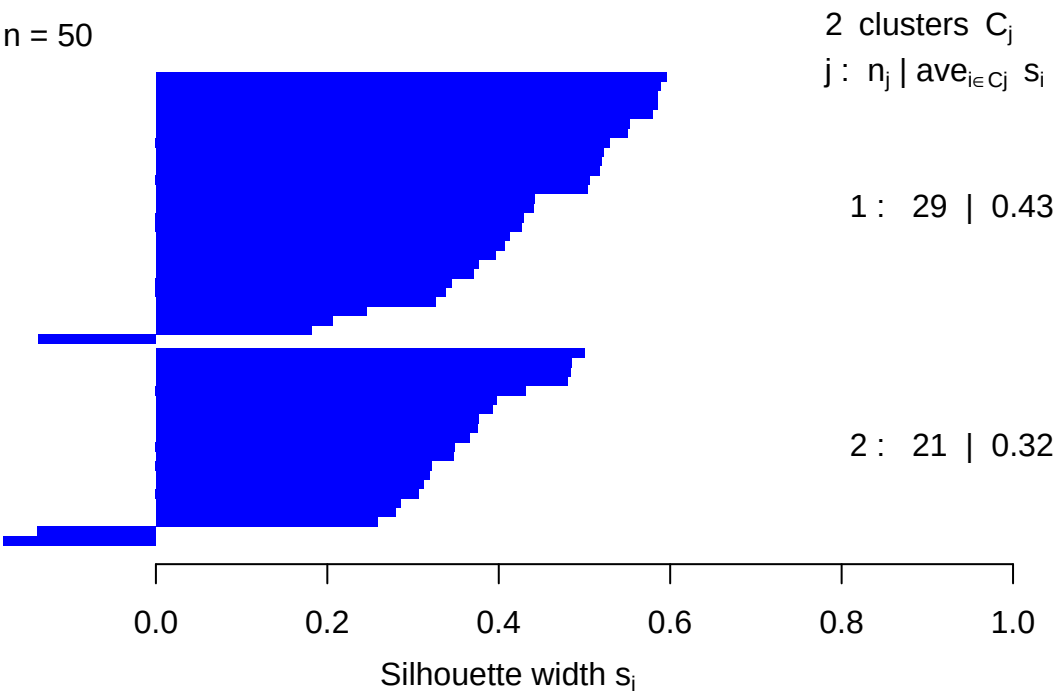
Silhouette sirve para que podamos representar graficamente la eficiencia de clasificación de una observación dentro de un grupo.

1.- Generación de los calculos

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

2.- Generación del grafico

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



Como se puede observar el ancho de Silhouette es de 0.38 y los porcentajes de clasificación de cada uno de los clústers es buena, se observa que hay datos fuera del rango, estos datos afectan nuestra capacidad de agrupamiento.

conclusiones

Para determinar el numero correcto de clústers fue necesario probar con distintos números de clústers, se intentó con 6 grupos, pero su capacidad de agrupamiento era de 0.03 también se intentó con 4 grupos pero la capacidad de agrupamiento fue de 0.09, es por eso que se decidió dejar en 2 ya que así la capacidad de agrupamiento subia a 0.38, esto de acuerdo al ancho de Silhouette.

Algo importante para mencionar es que en internet ofrecen algunos metodos para determinar el numero exacto de clísters que se deben utiizar, pero si se sigue esa metodología al momento de medir el ancho de Silhouette, se observara que no es una metodología del todo correcta, ya que la capacidad de agrupamiento es muy poco, por eso se recomienda que se pruebe con distintos numeros de clústers hasta encontrar el número de clusters donde su capacidad de agrupamiento sea la mayor.

Referencias

- https://statologos.com/k-medios-de-agrupacion-en-r/
- http://rpubs.com/rdelgado/399475
- El conjunto de datos utilizado esta precargado en R, se encuentra como “USArrests”

No lo creía hasta que lo viví :(