Modelos de Soporte No Supervisado

Implementación de k-medias en R y SAS

Objetivos

- Implementar la metodología K-medias en R y SAS
- Discutir las diferencias entre K-medias y K-medoids

Práctica II: K-medias

- La función kmeans () realiza clustering de K-means en R.
- Simularemos una serie de datos con dos grupos predefinidos.
- Las primeras 25 observaciones tienen un cambio medio en relación con las siguientes 25 observaciones.

Práctica II: K-medias

1. Simulación de los datos a agrupar

- set.seed (2)
- x=matrix (rnorm (50*2) , ncol =2)
- x[1:25 ,1]=x[1:25 ,1]+3
- x[1:25 ,2]=x[1:25 ,2] -4

Práctica II: K-medias

2. Realizamos el análisis de cluster con k-medias, con una k=2

- km.out =kmeans (x,2, nstart =20)

centros=número de grupos

Nstart: si los centros son un número, ¿cuántos semillas aleatorias se deben elegir?

Práctica II: K-medias

3. Para ver los resultados de la clasificación usamos

km.out$cluster

Y graficamos

plot(x, col =(km.out$cluster +1) , main="K-Means Clustering Results with K=2", xlab ="", ylab="", pch =20, cex =2)

Práctica II: K-medias

Sin embargo, para datos reales, en general no sabemos la verdadera cantidad de clusters. En cambio, podríamos haber realizado k-medoas con K = 3.

- set.seed (4)
- km.out =kmeans (x,3, nstart =20)
- km.out

Se generaron tres grupos con tamaños de 10, 23 y 17.

Práctica II: K-medias

- km.out
- km.out$centers
- km.out$cluster
- km.out$totss
- km.out$withinss
- km.out$tot.withinss
- km.out$betweenss
- km.out$size
- km.out$iter
- km.out$ifault
- plot(x, col =(km.out$cluster +1) , main="K-Means Clustering Results with K=3", xlab ="", ylab="", pch =20, cex =2)

Práctica II: K-medias

¿Cuál es la diferencia en los resultados?

```
set.seed (3)
 km.out =kmeans (x,3, nstart =1)
 km.out$tot .withinss
[1] 104.3319

 km.out =kmeans (x,3, nstart =20)
 km.out$tot .withinss
[1] 97.9793
```

Práctica II: K-medias

- Tenga en cuenta que km.out $ tot.withinss es la suma total de cuadrados dentro del cluster, que buscamos minimizar mediante la realización de clustering K-means.

- La suma de cuadrados individuales dentro del grupo está contenida en el vector km.out $ withinss.

- Es recomendable ejecutar siempre K-means clustering con un gran valor de nstart, como 20 o 50, ya que de lo contrario podría obtenerse un óptimo local no deseado.

- Al realizar K-means clustering, además de usar múltiples asignaciones iniciales de clúster, también es importante establecer una semilla aleatoria utilizando el comando

- set.seed () función. De esta manera, las asignaciones de clúster iniciales en el paso ánterior pueden ser replicadas, y la salida K-means será completamente reproducible.

Práctica II: ¿Cómo definir cuántos clusters utilizar?

- Una solución simple es calcular la agrupación k-means usando diferentes valores de clusters k.

- Utilizamos el WSS y lo graficamos respecto a la cantidad de clusters.

- El punto donde se estabiliza la curva es el número de clusters apropiado.

- La función R fviz_nbclust () [en el paquete factoextra] proporciona una solución conveniente para estimar el número óptimo de clusters.

Ejercicio. Aplique K-means a los datos USArrests

- **data**("USArrests") *# Cargamos la base de datos*
- df <- **scale**(USArrests) *# Estandarizamos*
- *# Muestra las primeras filas de los datos*
- **head**(df, n = 3)
- **library**(factoextra)
- **fviz_nbclust**(df, kmeans, method = "wss") +
- **geom_vline**(xintercept = 4, linetype = 2)

Ejercicio. Aplique K-means a los datos USArrests

- **set.seed**(123)
- km.res <- **kmeans**(df, 4, nstart = 25)
- *# Print the results*
- **print**(km.res)
- **aggregate**(USArrests, by=**list**(cluster=km.res$cluster), mean)
- dd <- cbind(USArrests, cluster = km.res$cluster)
- head(dd)

Ejercicio. Aplique K-means a los datos USArrests

¿Cómo visualizamos los clusters generados?

- **fviz_cluster**(km.res, data = df, palette = **c**("#2E9FDF", "#00AFBB", "#E7B800", "#FC4E07"),
  ellipse.type = "euclid", *# Concentration ellipse*
  star.plot = TRUE, *# Add segments from centroids to items*
  repel = TRUE, *# Avoid label overplotting (slow)*
  ggtheme = **theme_minimal**()
)

Ejercicio. Aplique K-means a los siguientes datos

Utilice la siguiente base de datos descargando la librería ISLR en R

- > library (ISLR)
- > nci.labs=NCI60$labs
- > nci.data=NCI60$data

Ejemplo en SAS

Data Set BUYTEST

The variables in this data set are as follows. There are 10,000 records (rows) in this data set.

Variable Name	Type	Description
Age	Numeric	Age in years
Income	Numeric	Yearly income in thousands of dollars
Married	Numeric	(binary) 1 if married, 0 otherwise
Sex	Category	(binary) M, F

Ejemplo en SAS

Variable Name	Type	Description
Coa6	Numeric	(binary) 1 if change of address in last 6 months, 0 otherwise
Ownhome	Numeric	(binary) 1 if own home, 0 otherwise
Loc	Category	Location of residence code: A-H
Climate	Category	Climate code for residence, 10, 20, and 30
Buy6	Numeric	Number of purchases in last 6 months
Buy12	Numeric	Number of purchases in last 12 months
Buy18	Numeric	Number of purchases in last 18 months
Value24	Numeric	Total value of purchases in past 24 months
Fico	Numeric	Credit score
Orgsrc	Category	Original customer source code: (C, D, I, O, P, R, U)
Discbuy	Numeric	(binary) 1 if a discount buyer, 0 otherwise
Return24	Numeric	(binary) 1 if product was returned in past 24 months, 0 otherwise
Respond	Numeric	(binary) 1 if responder to test mailing, 0 otherwise
Purchtot	Numeric	Test mailing purchase total
C1 – C7	Numeric	Test mailing total by product category
ID	Category	Unique ID number for each customer