Análisis Multivariado

ID 033521 - Clase 4901

Lina Maria Acosta Avena

Ciencia de Datos
Departamento de Matemáticas
Pontificia Universidad Javeriana

Semana 16: 04/11/24 – 09/11/24

Introducción

Ya vimos que:

- ✓ **El análisis de conglomerado (cluster o de clasificación)** tiene como propósito definir la estructura de los datos colocando las **unidades más parecidas en grupos**.
- ✓ Ese objetivo envuelve la determinación de: medida de **similitud o disimilitud**, procedimientos para **conformar los cluster** y el criterio para el **número de cluster**.
- ✓ Dentro de las medidas de **similitud/disimilitud** se encuentran las medidas de distancias (recomendado para variables cuantitativas): **Euclidiana, distancia generalizada, Manhatan y Minkowski.**

Introducción

- ✓ Revisamos el coeficiente de **concordancia/asociación** (presencia o ausencia - variables dummy): coeficiente de concordancia **simple y Jaccard**, entre otros.
- ✓ Dentro de las técnicas para **construir** los **cluster** vimos las **técnicas jerárquicas aglomerativas**: enlace simple (**vecino mas cercano**), enlace completo (**vecino mas lejano**), **distancias medias** y el método de **Ward**.
- ✓ Para definir el **número de cluster** revisamos: **Dendrograma, criterio de la silueta y wss** (Within Cluster Sum of Square)

Pontificia Universidad JAVERIANA Bogotá

Métodos no Jerárquicos

- ✓ Las **técnicas no jerárquicas** tienen como objetivo, encontrar directamente una partición de n elementos en k grupos o cluster, que envuelva "similitud" interna y separación entre los cluster.

Observaciones:

- *Computacionalmente es practicamente imposible, crear todas las posibles particiones de orden k, y apartir de éstas decidir cual es la "más adecuada". En este caso, surge la necesidad de investigar cuál de las (algunas) posibles particiones es "la más optima".*
- *En la técnicas no jerárquicas requieren del conocimiento o determinación previa del número de cluster.*

Métodos no Jerárquicos

- *En cada etapa de agrupamiento, se pueden formar nuevos grupos, esto es, pueden juntar/dividir grupos que fueron previamente combinados en pasos anteriores.*
- *En comparación con las técnicas jerárquicas, en las no jerárquicas los métodos computacionales son iterativos, tienen mayor capacidad de análisis en conjuntos de datos grandes.*

Existen varios **métodos de agrupamiento no jerárquicos**. Dentro de los **más populares** y destacados están:

- ✓ k-medias (**k-means**),
- ✓ Fuzzy e-medias (**Fuzzy e-means**)

K-Medias

- ✓ **redes neuronales**
- ✓ **escalonamiento multidimendional**

En la práctica, K-means es, probablemente el más utilizado.

K-Medias

- ✓ **redes neuronales**
- ✓ **escalonamiento multidimendional**

En la práctica, K-means es, probablemente el más utilizado.

K-means fue desarrollado por **MacQueen (1967)** y la idea básica es localizar/ubicar cada observación multivariada al cluster cuyo centroide (vector de medias) es el "mas cercano" del vector de valores observado para la respectiva observación.

K-Medias

El método consiste en:

- Paso 1: escoja k centroides, llamados "semillas" o "prototipos"
- Paso 2: usando una medida de distancia, cada observación multivariada es comparada con cada uno de los k centroides iniciales.
- Paso 3: las observaciones de k grupos son usadas para recalcular los centroides y se repite el paso 2
- Paso 4: repite paso 2 y 3 hasta que no haya cambios en las alteraciones de las observaciones de los grupos que están siendo localizados (los que estan el grupo 1 son siempre los mismos, etc) .

K-Medias

No todos los software o paquetes usan el mismo algoritmo, esto es, tienen variaciones; sin embargo, proporcionan resultados que no son muy diferentes, aunque la elección de la semilla inicial del agrupamiento sí influye en el agrupamiento final.

Algunas recomendaciones para elegir los centroides iniciales:

1. usar técnicas jerárquicas aglomerativas para definir el número de cluster. Entonces se calcula la media con las observaciones que quede en cada cluster, éstas seran las estimaciones inciales.
2. selección aleatoria
3. selección con base en alguna de las variables.

K-Medias

4. usar los datos discrepantes como centroides.
5. los k primeros valores del conjunto de datos.
6. conocimiento apriori basado en trabajos anteriores.

Evaluar el rendimiento de k-means es crucial para asegurar que los clusters que hayan sido conformados sean útiles y significativos. Dentro de los métodos más relevantes estan:

K-Medias

- ✓ la **suma de cuadrados dentro del cluster** (*within-cluster sum square* - ***WSS***), la cual mide la variabilidad dentro (intra) de los clusters. Generalmente se grafica el WSS contra el número de clusters y se busca el codo que indica donde la reducción en WSS se vuelve menos significativa.
- ✓ el **método de la silueta** que mide qué tan similar es un punto a su propio cluster en comparación con otros clusters. El valor de la silueta (score) varía entre -1 y 1. Valor cercano a 1 sugieren que los puntos están bien agrupados. Generalmente se calcula para diferentes números de clusters y buscar el máximo.

Estos son criterios internos. Validación cruzada y F1-score son dos de los criteriores de naturaleza externa, más populares.

K-Medias

Example

Los datos `USArrests` disponibles en `R` contiene información sobre el número de detenciones realizadas (por cada 100.000 habitantes) en 50 estados americanos en 1973. Específicamente:

- **Murder:** tasa de arrestos por asesinato.
- **Assault:** tasa de arrestos por asalto.
- **UrbanPop:** porcentaje de la población que vive en áreas urbanas en cada estado.
- **Rape:** tasa de arrestos por violación.

K-Medias

```
# ----- USArrests Data ----- #
data("USArrests")
head(USArrests)
Datos<-scale(USArrests)

# --- No. de cluster
require(factoextra)
fviz_nbclust(USArrests,
             kmeans,
             method = "wss")
```

K-Medias

```
# --- Usamos K-means con k=4 cluster
km<-kmeans(Datos,
           centers = 4,
           nstart = 25)

aggregate(USArrests,
          by=list(cluster=km$cluster),
          mean)

fviz_cluster(km,
             USArrests,
             ellipse.type = "norm")
```