K-Means

Equipo 5

Introducción al algoritmo

-Método de agrupamiento (clustering)

-Algorítmo NO supervisado (no tiene variable dependiente)

-Busca maximizar la variación inter-cluster y maximizar la variación intra-cluster

-Usamos las bibliotecas dplyr, boom, tydiverse y ggplot2.

El funcionamiento general del algoritmo

Desarrollo:

1. Realiza selecciones aleatorias, tantas como clusters hayamos elegido.
2. Asignar cada observación al punto más cercano.
3. Calcular los centroides de cada uno de los grupos creados.
4. Reasignar las observaciones en función de los nuevos centroides. (Esta operación se repetirá tantas veces como sea necesario mientras haya variaciones entre los clusters)

Evaluación:

- Inercia total $totss (Inercia de los grupos (clusters) con respecto al centroide de todas las observaciones)
- Inercia entre grupos $betweenss (Debemos procurar que sea la mayor posible, lo que nos definirá el número “óptimo” de clusters, asegura heterogeneidad entre los grupos)
- Inercia dentro de los grupos $withinss/$tot.whitinss (Indica las inercias individuales de cada grupo/suma de las “n” inercias)
- Inercia total = inercia entre grupos + inercia dentro de los grupos

Ventajas / desventajas

Ventajas:

- Almacenamiento rápido.
- Almacenamiento económico (sólo necesita guardar los k centroides)

Desventajas:

- Hay que probar número de clusters para encontrar la mejor inercia inter grupos.
- Es débil cuando existen outliers.

Problemas en los que se aplica

Se utiliza cuando tenemos muchos datos sin etiquetar. Responde a preguntas de tarifas o targets de usuarios.

Ejemplo/Ejemplo práctico: Compañías aseguradoras quienes quieren determinar los precios de sus seguros dependiendo de la edad, número de siniestros registrados, y target de seguros.

Ahora a R Studio!

!Gracias!

Referencias:

Roberto Caride. (2017). Ejemplo básico algoritmo K-means con R studio. 25/09/2020, de Youtube Sitio web: https://www.youtube.com/watch?v=w_aUCJHRv0Y&feature=youtu.be