DMC ONLINE

#YoMeQuedoEnCasa

TALLER:

ANÁLISIS CON MÉTODOS NO SUPERVISADOS

Perú

Soy

Daniel Chávez Gallo

Científico de datos CVM en Entel Perú

Me puedes encontrar como:

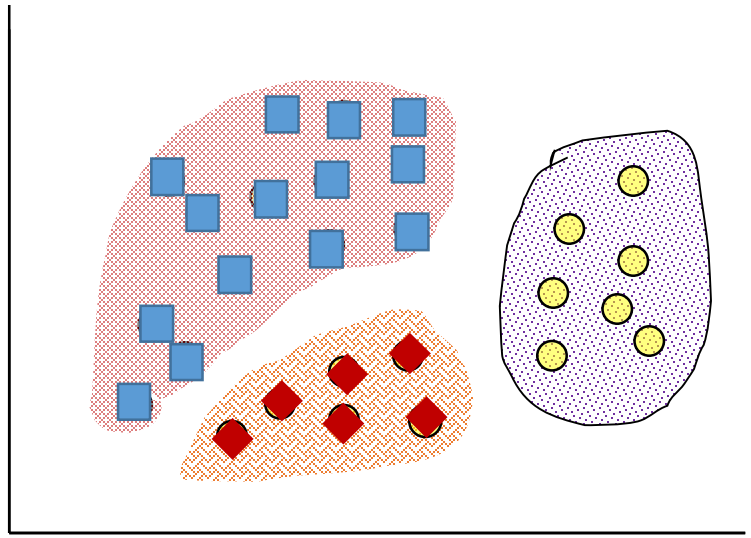
dacg160381@hotmail.com

www.dmc.pe

Ideas básicas

Agrupar objetos similares entre sí que sean distintos a los objetos de otros agrupamientos

- Datos dentro del mismo grupo deben tener caracteristicas similares.
- Datos de grupos diferentes deben tener características diferentes.

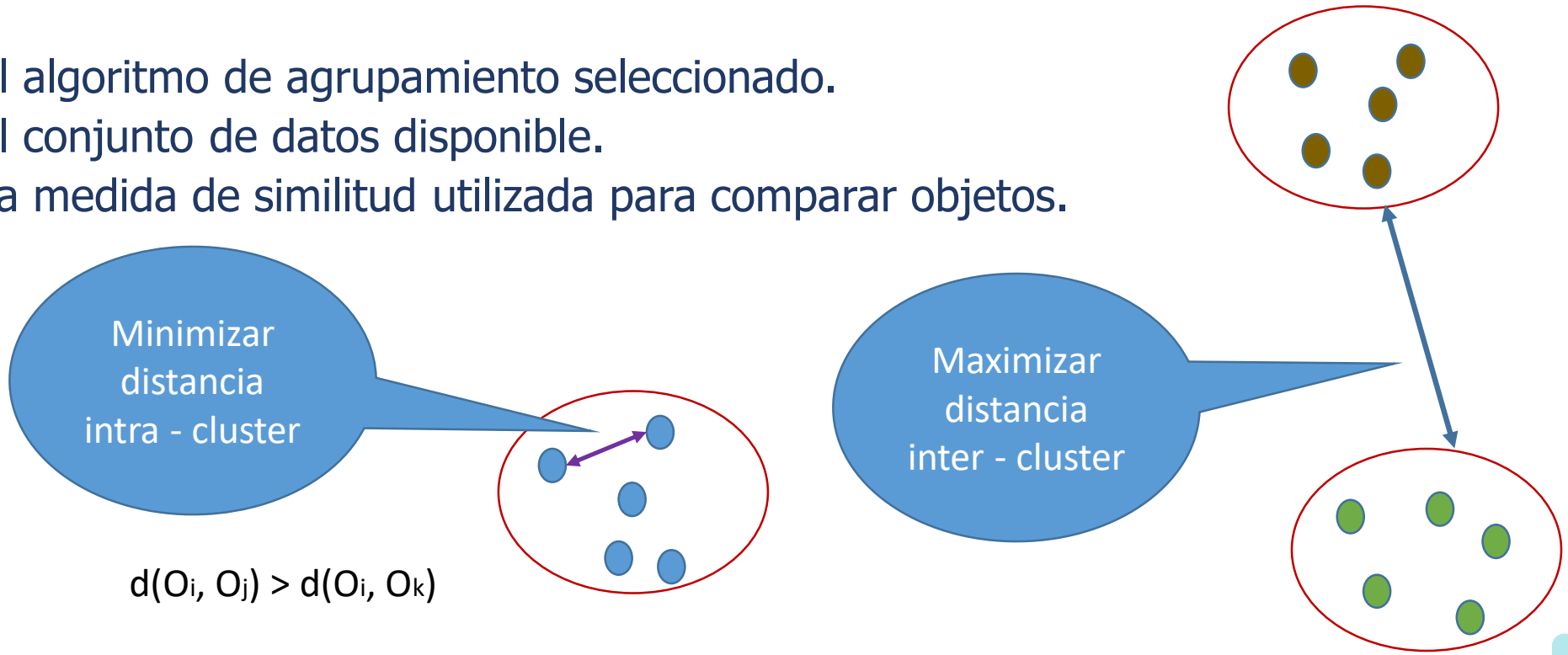


Ideas básicas

Agrupar objetos similares entre sí que sean distintos a los objetos de otros agrupamientos

Los resultados obtenidos dependerán de:

- El algoritmo de agrupamiento seleccionado.
- El conjunto de datos disponible.
- La medida de similitud utilizada para comparar objetos.

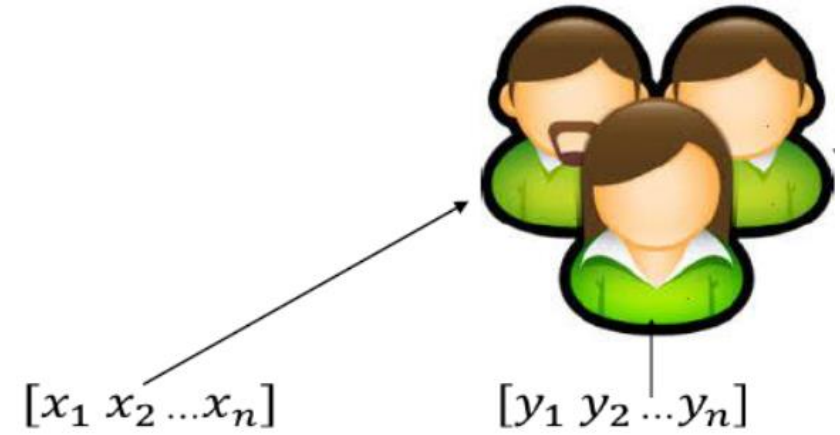


$d(O_i, O_j) > d(O_i, O_k)$

Ideas básicas

Noción de similitud

Dada una representación vectorial de dos clientes ***x*** y ***y,*** podemos determinar el grado de similitud entre ellos a través del uso de una **métrica.**



$$d(x,y) = \sqrt[2]{\sum_{i=1}^{n} |x_i - y_i|^2}$$

Ideas básicas

- **¿Cuántos grupos?**

  Grupos o clusters no definidos a priori. Diferencia con los métodos supervisados.

- **¿Cómo buscarlos?**

  Los objetos dentro de un cluster sean similares o cercanos entre sí en algún sentido (gran similaridad intra-clase) y diferentes o alejados a los objetos de otro cluster (baja similaridad inter-clase)

Ideas básicas

Para medir la distancia entre las instancias, es necesario que todos los atributos estén en la misma escala.

- **Normalización**: escala los valores numéricos en el Rango [0,1]

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

- **Estandarización**: hace que la distribución de los datos sea normal

$$x_{new} = \frac{x - \mu}{\sigma}$$

Agrupación k-means

- Probablemente el más utilizado y conocido
- Asigna cada observación a uno de los k clusters
- K es un número definido a priori
- Minimizar las distancias intra cluster y maximizar las inter clase

Agrupación k-means

- ¿Cómo funciona el algoritmo?
  1. Elegir el valor de K (número de clusters).
  2. Asignar cada objeto al grupo más cercano (por ejemplo distancia euclídea)
  3. Re-estimar los centros de los k clusters, asumiendo que las asignaciones a los grupos están bien.
  4. Repetir el paso 3 hasta que no haya más cambios

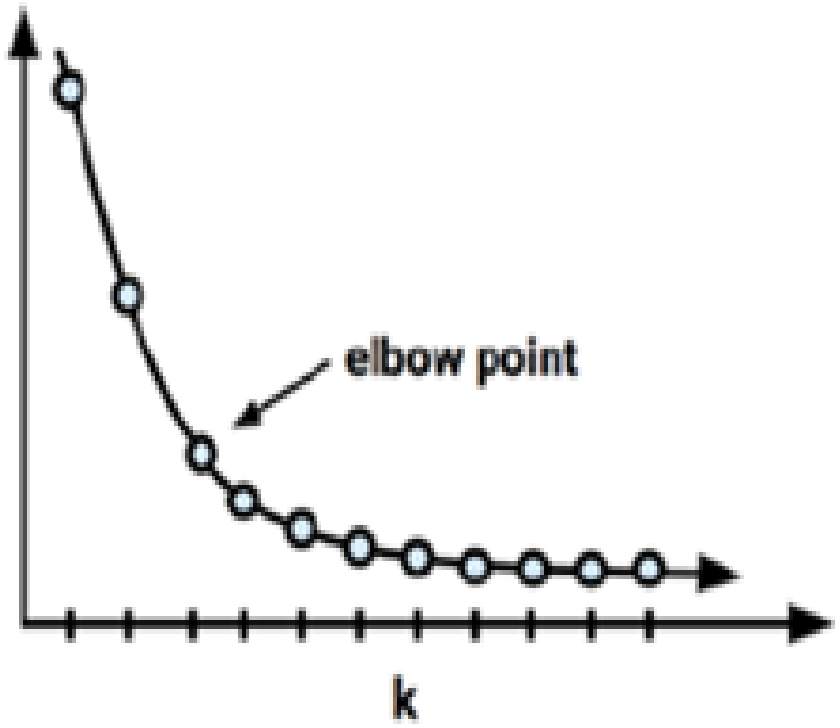
- Se puede cambiar el punto 2, empezando con k centroides iniciales

- La mayor parte de las reasignaciones ocurren en la primera iteración del algoritmo

Agrupación k-means

Elegir el **número de clústers:**

- **Conocimiento a priori**: por ejemplo, si clasificamos películas, k = nº de géneros
- **Dirigidos por el negocio**: por ejemplo, el departamento de Marketing sólo tiene recursos para hacer 3 campañas distintas de marketing
- Sin nada de lo anterior: k = raíz(n/2), valor inicial



Agrupación k-means

- **Ventajas**
  - Principios no estadísticos
  - Muy flexible
  - Funciona bien en casos de la vida real
  - Rápido: no hay calcular las distancias entre todas y cada una de las observaciones

- **Desventajas**
  - No muy sofisticado
  - No está garantizado encontrar en número de clusters óptimo
  - Sensible a outliers que pueden formar clusters propios
  - La solución final depende del punto de partida

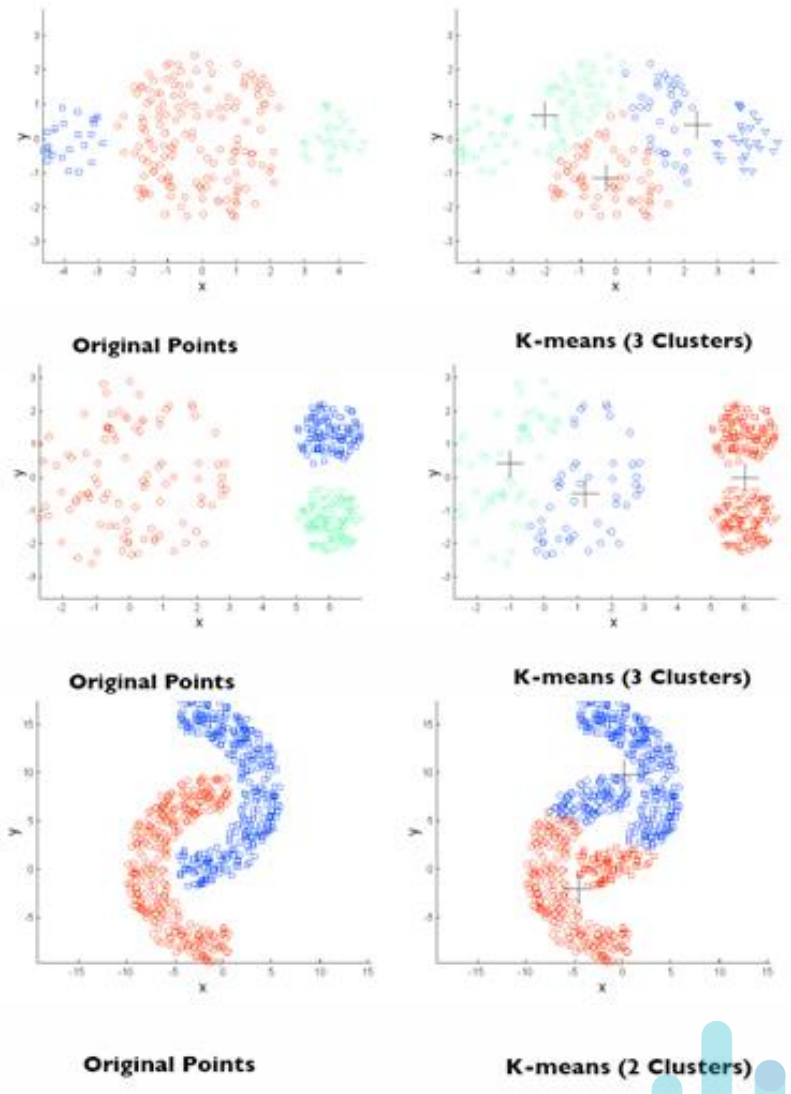
Agrupación k-means

Limitaciones

- Principalmente, su desempeño se ve mermado cuando los clusters tienen
  - Diferentes tamaños
  - Diferentes densidades
  - Formas no globulares

- (Al igual que casi todos) También presenta problemas cuando los datos contienen outliers

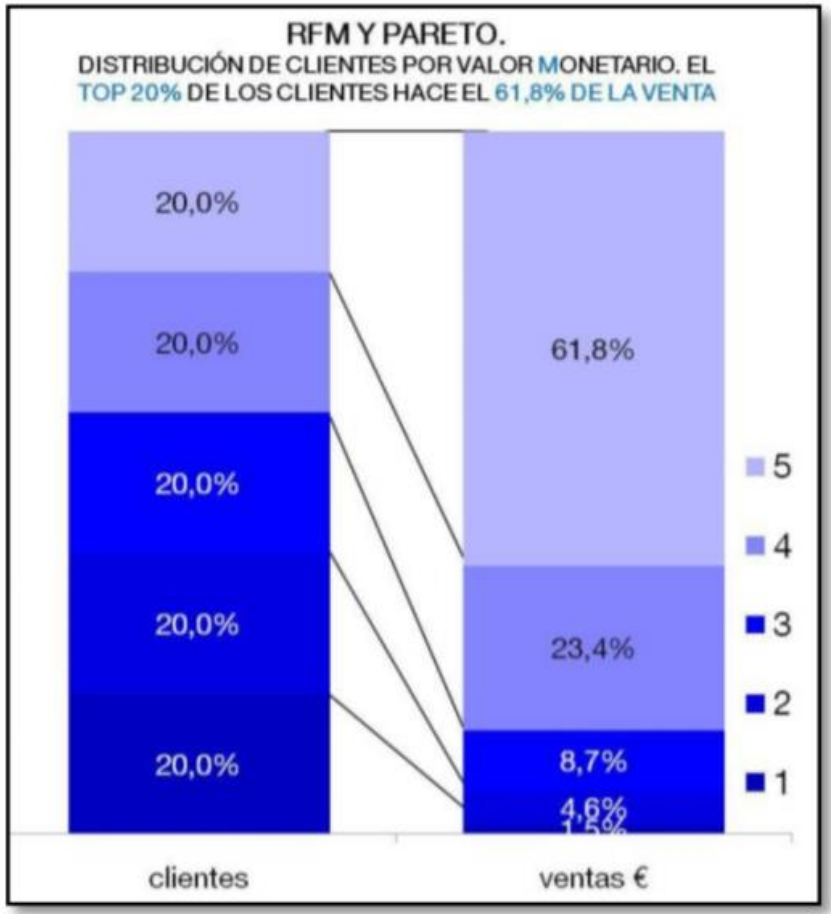
- Una solución puede ser hacer un número superior de clusters, y luego “unir las partes”



Agrupación RFM

- Los más propensos a comprar son aquellos que han comprado más recientemente, con más frecuencia y gastan más dinero.

- Se aplica sobre esta “Ley de Pareto” y se refiere a que “el 80% de las compras las realizan el 20% de los clientes”.

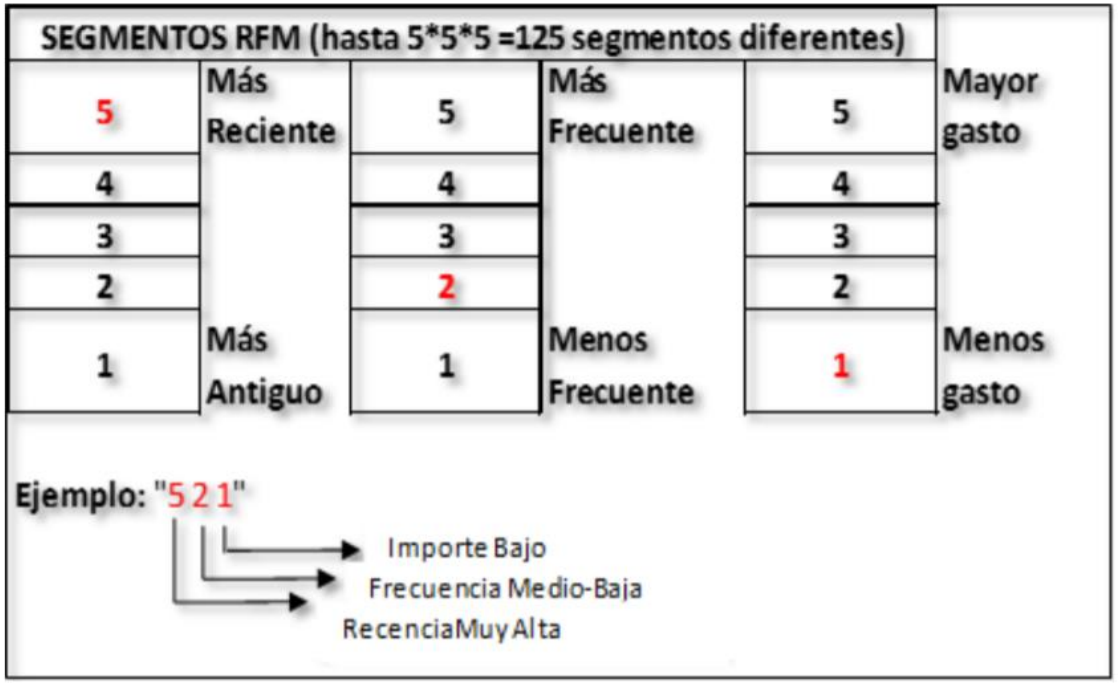


Agrupación RFM

- Con la **Recencia**, medimos los días que han pasado desde hoy (o cualquier fecha a futuro) hasta la fecha en que realizó su última compra.

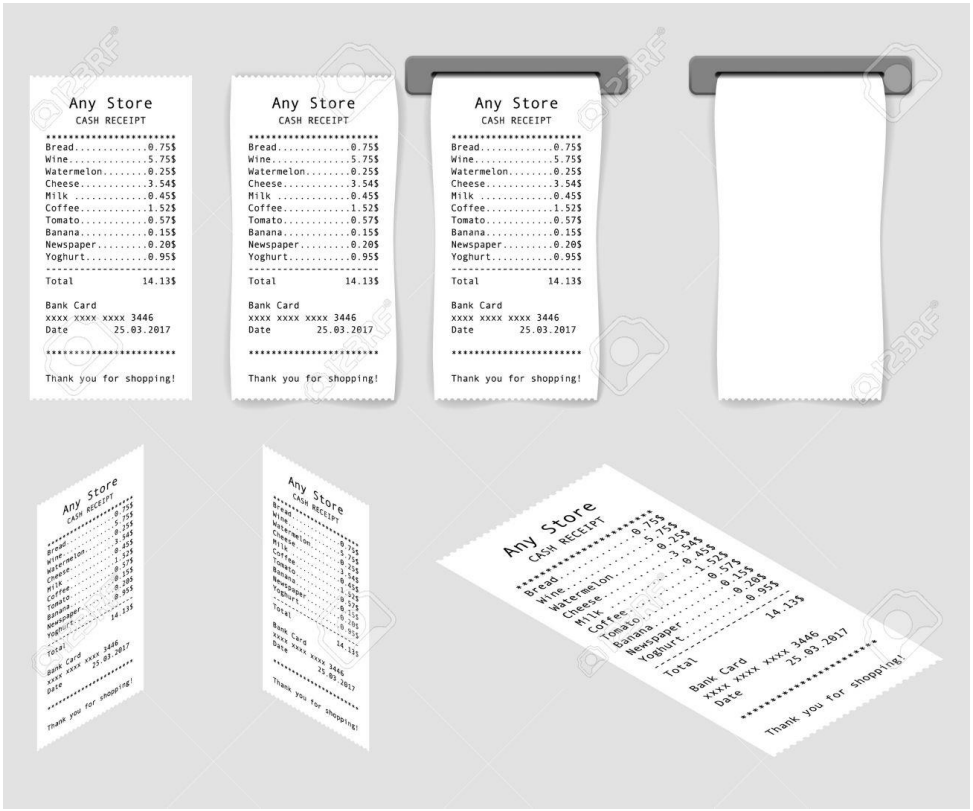
- Con la **Frecuencia**, medimos el número de compras que ha hecho cada cliente en total.

- Y el Valor **Monetario**, es la suma total de cantidad de dinero que el cliente lleva gastado en sus compras.



Reglas de asociación

- Los ejemplos para este tipo de problema están constituidos por **Transacciones** las cuales constan de un TID y un conjunto de Items o Itemset:
  - TID: Identificador de la transacción.
  - Itemset: Artículos de la transacción.



(4235, {Leche, Pan, Huevos, Jamón})

Reglas de asociación

- **Soporte**: Indica el porcentaje de transacciones que llevan juntos el antecedente y el consecuente, con respecto al total de transacciones realizadas.

- **Confianza**: Indica el porcentaje de transacciones que llevan juntos el antecedente y el consecuente, con respecto al total de transacciones donde sólo aparece el antecedente.

- **Lift**: Indica el aumento en la probabilidad de selección del consecuente, al ser comprado en conjunto con el antecedente.

Reglas de asociación

- Luego de obtener las reglas, se deben evaluar los indicadores estudiados.
- Las mejores reglas son aquellas en los que los tres indicadores son altos.
- El Soporte varía entre 0 y 1, al igual que la confianza, porque finalmente son probabilidades.
- El indicador de Lift es bueno si es más alto. Son recomendables los valores por encima de 1, siempre y cuando los otros valores sean también altos.
- Si A,B->C es una buena regla, significa que si el cliente compra A y B, tiene una probabilidad alta de comprar C.

¡GRACIAS!

dacg160381@hotmail.com