K-MEDIAS

Equipo 6

- 1728247 Castillo Cerda Manuel Orlando
- 1854568 Cedillo Hernandez Vanessa Nahomy
- 1847759 Lozano Rangel Antonio de Jesus
- 1795359 Vega Flores Blanca Janeth

CLASIFICACIÓN

\En la clasificación automática de datos se pueden considerar 3 tipos de algoritmos:

- **Clasificación no supervisada:** los datos no tienen etiquetas y estos se clasifican a partir de su estructura interna (propiedades, características).
- **Clasificación supervisada:** son conjuntos de datos llamados datos de entrenamiento, cada uno está asociado a una etiqueta, se crea un modelo utilizando dichas etiquetas que indica si los datos estan clasificados correctamente o incorrectamente
- **Clasificación semisupervisada:** algunos datos tienen etiquetas, pero no todos, este caso es muy típico en clasificación de imágenes, donde se disponen de muchas imágenes mayormente no etiquetadas.

CLUSTERING

Proceso de agrupar datos en clases o clusters de tal forma que los objetos de un cluster tengan una similaridad alta entre ellos, y se diferencien con objetos de otros clusters. También se conoce como segmentación

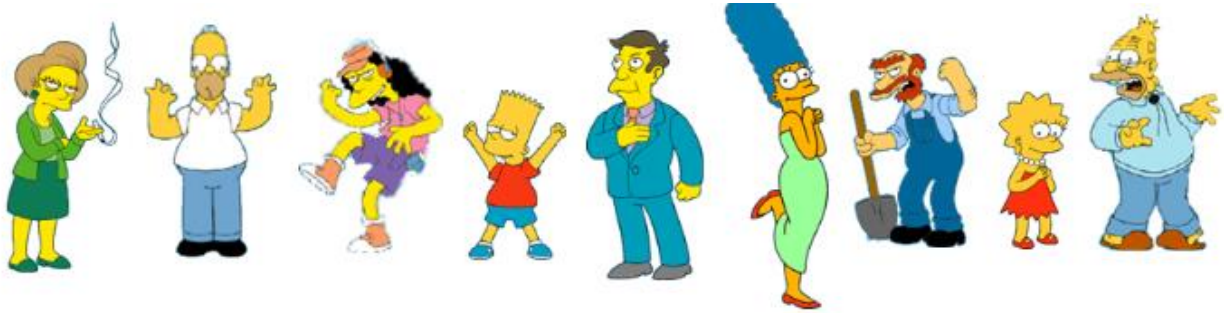
¿QUÉ ES K-MEDIAS?

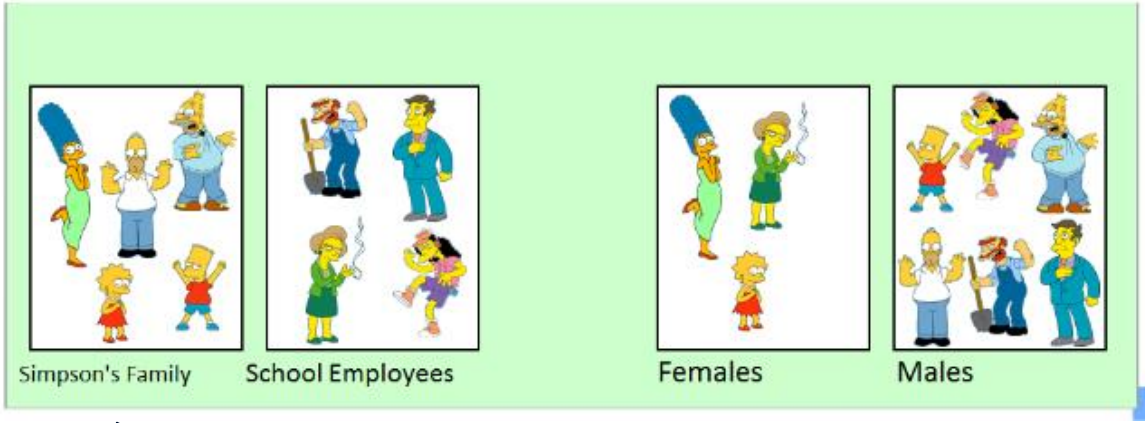
Es un algoritmo no supervisado de Clustering, se utiliza cuando tenemos demasiados datos sin etiquetar para con esto encontrar “K” grupos (clusters) entre los datos crudos. Este método de agrupamiento objetiva en la partición de un conjunto de n observaciones en k grupos en el que cada observación pertenece al grupo cuyo valor medio es más cercano

En general: K-means clustering son particiónes de las observaciones en un número predefinido de clústeres

EJEMPLO AGRUPAMIENTO

¿CUÁL ES EL AGRUPAMIENTO NATURAL ENTRE ESTOS OBJETOS



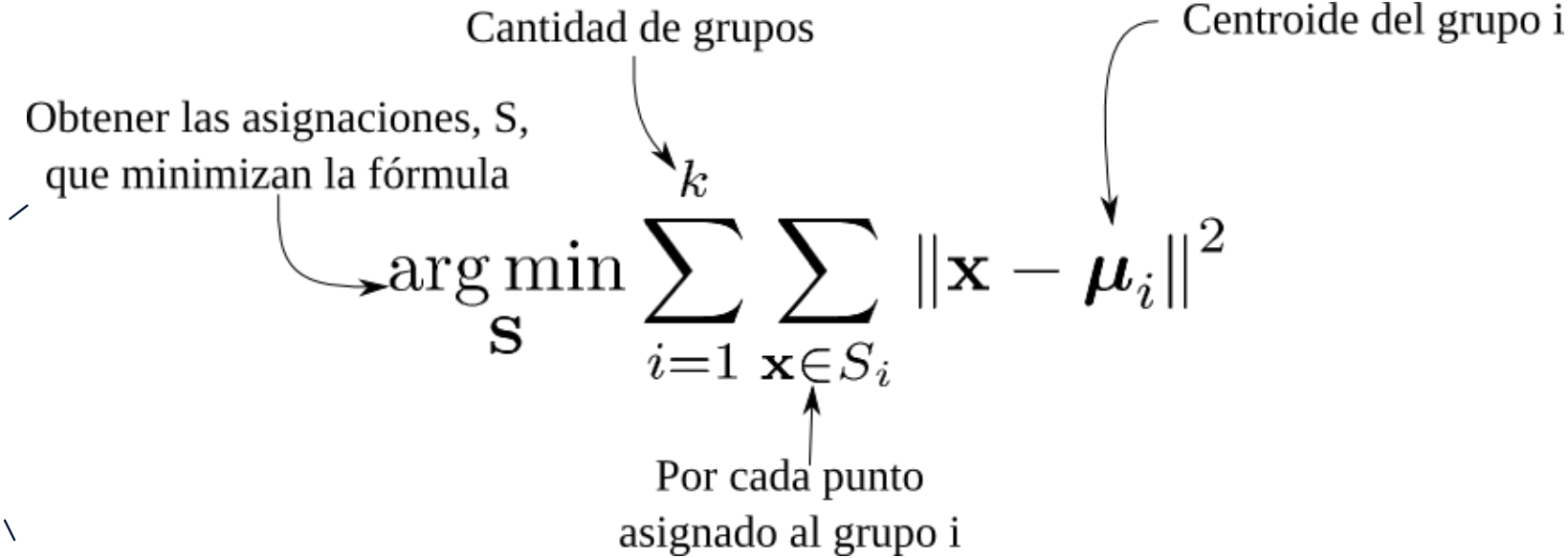


¿PARA QUÉ SIRVE K-MEDIAS?

Sirve para tener escalabilidad con la cantidad de datos. El algoritmo es de los **más usados** para encontrar grupos ocultos sobre un conjunto de datos no etiquetado, esto es de gran utilidad para confirmar o descartar alguna teoría asumida de nuestros datos. Y también puede ayudarnos a descubrir relaciones asombrosas entre conjuntos de datos, que deforma manual, no son reconocibles.

Una vez que se ejecuta y obtienen las etiquetas, es fácil clasificar nuevos valores o muestras entre los grupos obtenidos.

FORMA MATEMÁTICA

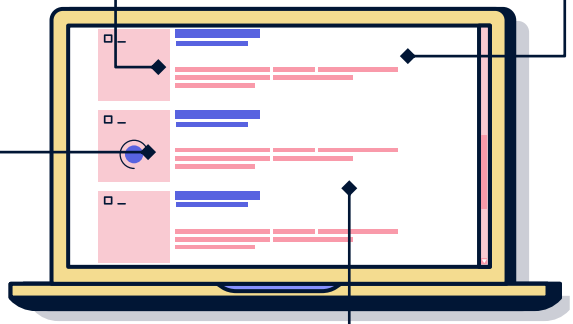


La forma en la que se opera es la suma de las varianzas de cada punto asignado respecto a la media de cierto grupo donde k es la cantidad de grupos y se busca obtener S asignaciones en busca de minimizar la formula para con esto poco a poco ir disminuyendo las medias de agrupaciones

ALGORITMO

1ER PASO

Aleatoriamente **seleccionar un K** (en nuestro caso K=3) de los puntos de datos de cada base que están siendo agrupados como el grupo inicial del centro geométrico en la cual estamos trabajando



2DO PASO

Asignar a cada punto con la agrupación **más cercana** la cual está representada como su centro geométrico basado en la distancia euclidiana

3ER PASO

Después de que todos los objetos han sido asignados, **recalcularemos** la posición de el número K de medias geométricas

4TO PASO

Repetir el paso 2 y 3 hasta que todos las medias de las agrupaciones no cambien nunca más

¿ELEMENTOS DEL K-MEDIAS?

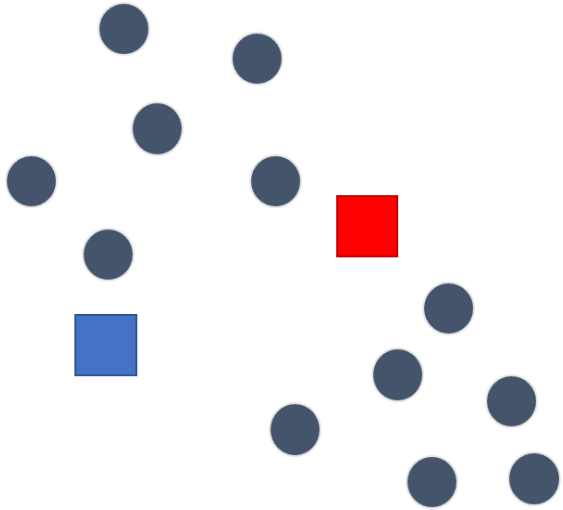
- **Inicialización**: se elige la localización de los centroides de los K grupos aleatoriamente

- **Asignación**: se asigna cada dato al centroide más cercano

- **Actualización**: se actualiza la posición del centroide a la media aritmética de las posiciones de los datos asignados al grupo

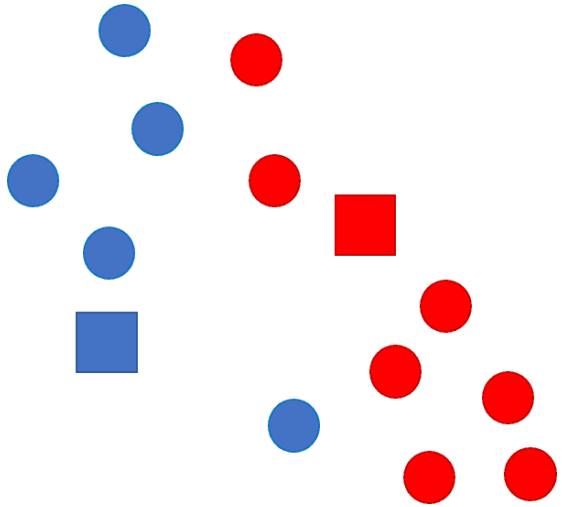
INICIALIZACIÓN

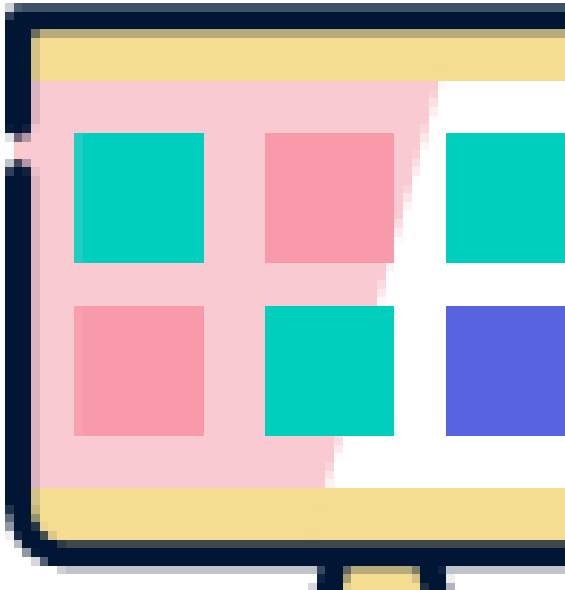
Se elige la localización de los centroides de los K grupos aleatoriamente. La figura muestra los datos como círculos y los centroides como cuadrados.



ASIGNACIÓN

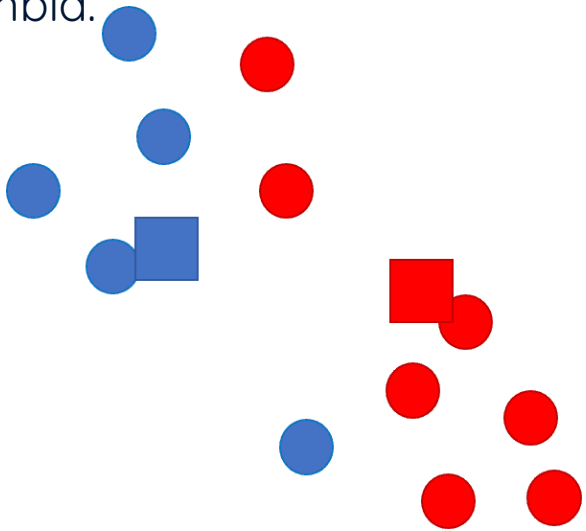
A continuación, se asigna cada dato al centroide más cercano. En el ejemplo, los círculos cambian de color para indicar a qué centroide han sido asignados.





ACTUALIZACIÓN

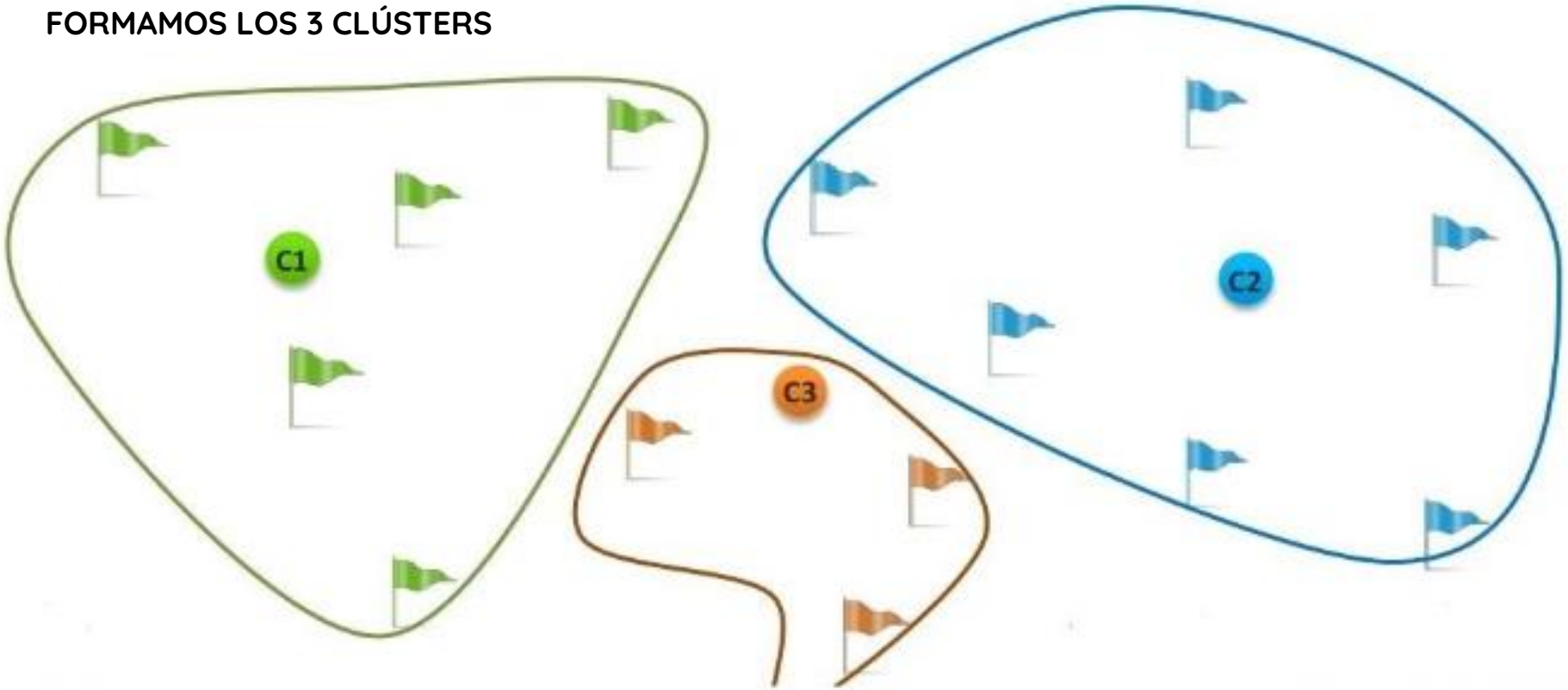
Ahora se actualiza la posición del centroide a la media aritmética de las posiciones de los datos asignados al grupo. Observa cómo la posición de los centroides (cuadrados) cambia.



A continuación irían las fases: asignación, actualización, asignación, actualización, etc. hasta que las posiciones de los centroides no cambien.

EJEMPLO (AGRUPACIÓN PIZZA HUT)

FORMAMOS LOS 3 CLÚSTERS



RECOMENDACIONES

01. SELECCIÓN DE CARACTERÍSTICAS

Al ser una técnica de aprendizaje automático no-supervisada, implica que **no es capaz** de establecer la relación entre los atributos de entrada y los resultado ya que no existen resultados.

Así que se debe **identificar** qué atributos son relevantes. Siempre es mejor usar el menor número atributos posible debido a que a medida que el número de dimensiones (atributos) aumenta, la distancia discrimina cada vez menos. Una práctica común antes de hacer clustering es reducir la dimensionalidad del problema.

RECOMENDACIONES

02. NORMALIZACIÓN

Siempre se debe normalizar nuestros datos, es decir que los valores de cada atributo estén en escalas similares, esto ayuda porque los grupos se forman a partir de distancias, si existen atributos con escalas muy diferentes, los atributos de escala mayor dominarán las distancias.

Las técnicas más comunes de normalización son:

- Re-escalar cada atributo en el rango [0, 1]
- Suponer que cada atributo sigue una distribución normal estándar

¿APLICACIONES DEL K-MEDIAS?

Relacionar el carrito de compras de un usuario, sus tiempos de acción e información del perfil.

SEGMENTACIÓN POR COMPORTAMIENTO

1

CATEGORIZACIÓN DE INVENTARIO

Agrupar productos por actividad en sus ventas

2

3

DETECTAR ANOMALÍAS

Según el comportamiento en una web reconocer un troll -o un bot- de un usuario normal

PROGRAMA EJEMPLO

https://github.com/VanessaCedillo19/Mineria_de_Datos/blob/main/Kmedias.ipynb

KAHOOT

https://create.kahoot.it/kahoots/my-kahoots

FUENTES

- https://www.iartificial.net/clustering-agrupamiento-kmeans-ejemplos-en-python/#Ejemplos_de_Clustering
- https://www.aprendemachinelearning.com/k-means-en-python-paso-a-paso/
- https://prezi.com/oo_q3nv4tsj4/k-means-presentation/