UTEC Posgrado

UTEC Posgrado

UTEC Posgrado

Una escuela para
Early Adopters

MACHINE LEARNING

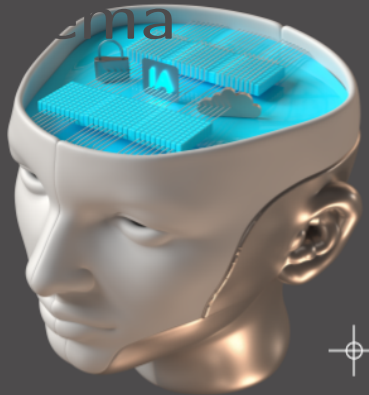
PRÁCTICA DE K-MEANS

UTEC Posgrado

Una escuela para
Early Adopters

Presentación del Pro ma

Dataset y contexto **K-Means Clustering**



Dataset: Clientes de una tienda retail

Una cadena de tiendas desea segmentar a sus clientes según su comportamiento de compra. Se recopilaron los siguientes datos de 6 clientes representativos:

Cliente	Gasto por visita (x_1)	Visitas al mes (x_2)
A	1	9
B	2	6
C	3	9
D	7	1
E	8	4
F	9	1

Objetivo: Identificar grupos de clientes con patrones de compra similares usando **K-Means**.

¿Por qué K-Means para este problema?

Justificación del uso de K-Means:

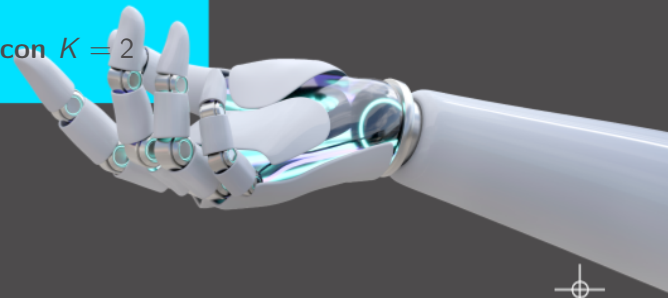
- **Datos numéricos continuos:** ambas variables (x_1, x_2) son cuantitativas, ideales para calcular distancias euclidianas
- **Grupos naturales esperados:** se observan patrones diferenciados (clientes frecuentes vs. clientes de alto gasto)
- **Escalabilidad:** K-Means es eficiente incluso con grandes volúmenes de datos, muy útil en retail
- **Interpretabilidad:** los centroides representan el "cliente típico" de cada segmento

Aplicaciones en el negocio:

- Campañas de marketing personalizadas por segmento
- Programas de fidelización diferenciados
- Optimización de inventario según perfil de cliente

Ejercicio 1

Segmentación de clientes **K-Means con** $K = 2$



Ejercicio 1: Enunciado

Aplique el algoritmo K-Means con $K = 2$ al dataset de clientes presentado anteriormente. Use la **distancia euclidiana** y los siguientes centroides iniciales:

$$\mu_1^{(0)} = A = (1,\ 9) \qquad \mu_2^{(0)} = F = (9,\ 1)$$

Se pide:

1. Calcular las distancias de cada punto a los centroides
2. Asignar cada punto al cluster más cercano
3. Recalcular los centroides como la media de cada cluster
4. Repetir hasta que las asignaciones no cambien (convergencia)
5. Interpretar los resultados obtenidos

Iteración 1: Cálculo de distancias

Centroides actuales: $\mu_1 = (1, 9), \quad \mu_2 = (9, 1)$

Fórmula: $d(\mathrm{x}, \boldsymbol{\mu}) = \sqrt{(x_1 - \mu_1)^2 + (x_2 - \mu_2)^2}$

Ejemplo: $d(B, \mu_1) = \sqrt{(2-1)^2 + (6-9)^2} = \sqrt{1+9} = \sqrt{10} \approx 3{,}16$

Punto	(x_1, x_2)	$d(\cdot, \mu_1)$	$d(\cdot, \mu_2)$	**Cluster**
A	$(1, 9)$	0	$\sqrt{128} \approx 11{,}31$	C_1
B	$(2, 6)$	$\sqrt{10} \approx 3{,}16$	$\sqrt{74} \approx 8{,}60$	C_1
C	$(3, 9)$	$\sqrt{4} = 2{,}00$	$\sqrt{100} = 10{,}00$	C_1
D	$(7, 1)$	$\sqrt{100} = 10{,}00$	$\sqrt{4} = 2{,}00$	C_2
E	$(8, 4)$	$\sqrt{74} \approx 8{,}60$	$\sqrt{10} \approx 3{,}16$	C_2
F	$(9, 1)$	$\sqrt{128} \approx 11{,}31$	0	C_2

Iteración 1: Actualización de centroides

Cluster C_1: $\{A, B, C\}$

$$\mu_1^{(1)} = \left(\frac{1+2+3}{3}, \frac{9+6+9}{3} \right) = \left(\frac{6}{3}, \frac{24}{3} \right) = (2,\ 8)$$

Cluster C_2: $\{D, E, F\}$

$$\mu_2^{(1)} = \left(\frac{7+8+9}{3}, \frac{1+4+1}{3} \right) = \left(\frac{24}{3}, \frac{6}{3} \right) = (8,\ 2)$$

Los centroides cambiaron: $(1, 9) \rightarrow (2, 8)$ y $(9, 1) \rightarrow (8, 2)$.

$\Rightarrow$ Se requiere otra iteración.

Iteración 2: Verificación de convergencia

Centroides actuales: $\mu_1 = (2, 8), \quad \mu_2 = (8, 2)$

Punto	(x_1, x_2)	$d(\cdot, \mu_1)$	$d(\cdot, \mu_2)$	Cluster
A	$(1, 9)$	$\sqrt{2} \approx 1{,}41$	$\sqrt{98} \approx 9{,}90$	C_1
B	$(2, 6)$	$2{,}00$	$\sqrt{52} \approx 7{,}21$	C_1
C	$(3, 9)$	$\sqrt{2} \approx 1{,}41$	$\sqrt{74} \approx 8{,}60$	C_1
D	$(7, 1)$	$\sqrt{74} \approx 8{,}60$	$\sqrt{2} \approx 1{,}41$	C_2
E	$(8, 4)$	$\sqrt{52} \approx 7{,}21$	$2{,}00$	C_2
F	$(9, 1)$	$\sqrt{98} \approx 9{,}90$	$\sqrt{2} \approx 1{,}41$	C_2

Las asignaciones **no cambiaron** respecto a la iteración anterior.

$\Rightarrow$ **Convergencia alcanzada en 2 iteraciones.**

Ejercicio 1: Interpretación de resultados

Cluster C_1 — Clientes frecuentes: $\{A, B, C\}$

- Centroide: $\mu_1 = (2, 8)$ — bajo gasto, alta frecuencia
- Clientes que visitan seguido pero gastan poco por visita
- **Estrategia:** Programas de fidelización, descuentos por volumen

Cluster C_2 — Clientes premium: $\{D, E, F\}$

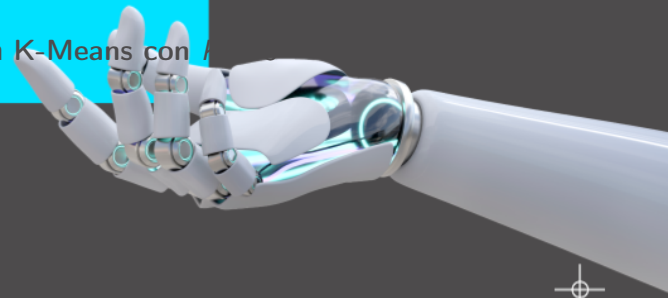
- Centroide: $\mu_2 = (8, 2)$ — alto gasto, baja frecuencia
- Clientes que compran poco frecuente pero gastan mucho
- **Estrategia:** Atención personalizada, ofertas exclusivas

UTEC Posgrado

Una escuela para
Early Adopters

Ejercicio 2

Ubicación de centros de distribución K-Means con A



Ejercicio 2: Enunciado

Una empresa de logística debe ubicar **3 centros de distribución** para atender 9 puntos de entrega en una ciudad. Las coordenadas (en km) son:

Punto	x_1 (km)	x_2 (km)
P_1	1	2
P_2	3	1
P_3	2	3
P_4	6	7
P_5	7	8
P_6	8	6
P_7	9	1
P_8	10	2
P_9	8	3

Aplique K-Means con $K = 3$ y centroides iniciales:

Ejercicio 2: Instrucciones

Resuelva el ejercicio siguiendo los mismos pasos del Ejercicio 1:

1. Calcule la distancia euclidiana de cada punto $(P_1, \ldots, P_9)$ a los 3 centroides iniciales
2. Asigne cada punto al centroide más cercano
3. Recalcule los centroides como la media de los puntos de cada cluster
4. Repita los pasos 1–3 hasta que las asignaciones no cambien
5. Interprete los clusters obtenidos en el contexto del problema de logística

Sugerencia: Organice sus cálculos en una tabla con columnas para cada distancia y la asignación final, como se mostró en el Ejercicio 1.

Conclusiones

Resumen de la práctica

Resumen de la práctica

- K-Means agrupa datos minimizando la distancia intra-cluster
- El algoritmo itera: **asignar** $\rightarrow$ **actualizar centroides** $\rightarrow$ **verificar convergencia**
- La elección de K y los centroides iniciales afectan el resultado final
- Los centroides representan el "punto típico" de cada grupo
- Aplicaciones: segmentación de clientes, ubicación de centros, compresión de datos