Algoritmos de Clustering No Supervisados: Una Introducción

Los algoritmos no supervisados son técnicas de aprendizaje automático que detectan patrones en datos sin etiquetas previas. Entre ellos, el clustering destaca por su capacidad de agrupar datos similares sin supervisión.

Estos métodos son cruciales para segmentar clientes en marketing, analizar imágenes y organizar datos complejos. Comprender el clustering nos abre la puerta a descubrir estructuras ocultas y tomar decisiones basadas en datos no categorizados.

¿Qué es Clustering?

Definición Formal

El clustering agrupa objetos o datos similares basándose en características comunes, formando grupos llamados clusters.

Diferencia con Clasificación

La clasificación asigna datos a etiquetas conocidas, mientras que el clustering descubre grupos sin etiquetas previas.

Tipologías

- Particional: divide en grupos exclusivos
- Jerárquico: crea estructuras de clusters anidados
- Basado en densidad: detecta clusters separando áreas densas de ruido

Medidas de Proximidad en Clustering

- **Distancia Euclidiana:** Mide la distancia recta entre dos puntos en un espacio n-dimensional.

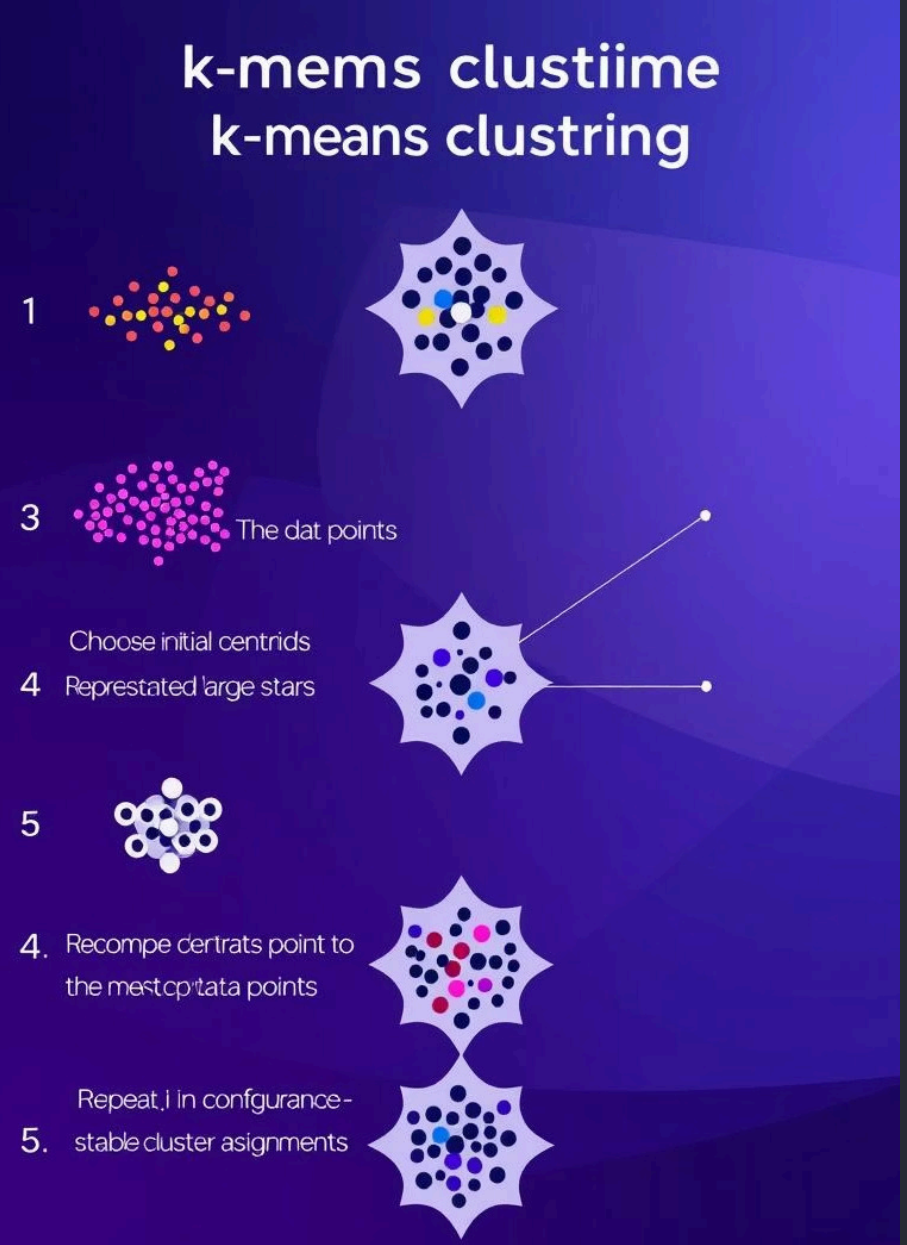
$$\text{dist}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

- **Distancia Manhattan:** Suma de las diferencias absolutas entre coordenadas; también llamada distancia de ciudad.

$$\text{distancia}_{\text{Manhattan}}(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

Medidas de Proximidad en Clustering

- **Distancia de Minkowski:** Generaliza Euclidiana y Manhattan, ajustando su parámetro para distintos cálculos.
- **Coeficiente de Correlación:** Evalúa similitud midiendo la relación lineal entre vectores.



Algoritmo K-Means

1

Inicialización

Selecciona k centroides aleatorios como puntos de partida.

2

Asignación

Asigna cada punto al cluster con el centroide más cercano.

3

Actualización

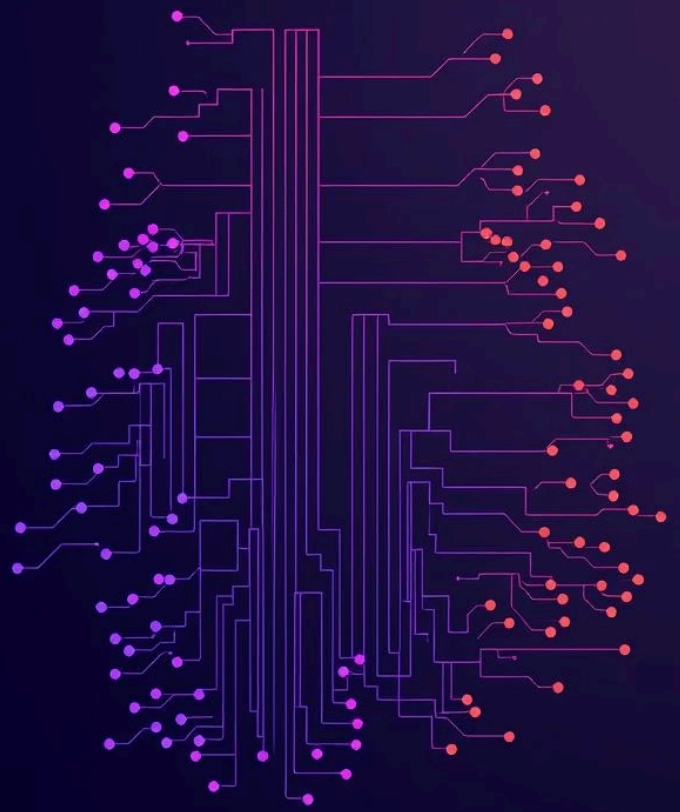
Recalcula los centroides como el promedio de puntos asignados.

4

Iteración

Repite la asignación y actualización hasta convergencia.

K-Means es eficiente con complejidad O(n*k*i), pero sensible a la inicialización y no maneja clusters de forma arbitraria.



Clustering Jerárquico

Aglomerativo

Empieza con cada punto como cluster individual y los fusiona.

Divisivo

Comienza con todos los datos y los divide progresivamente.

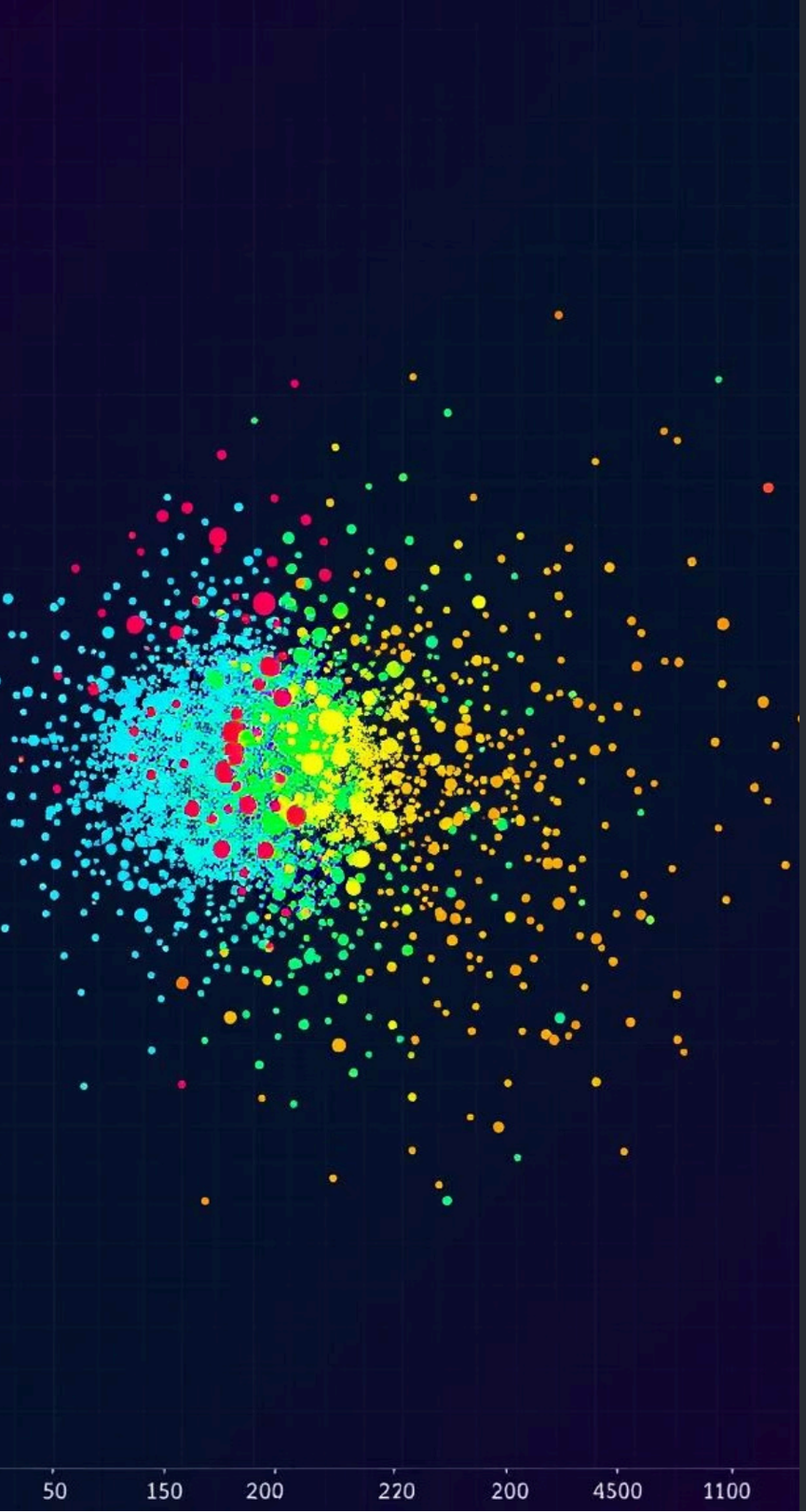
Tipos de Enlace

Simple, completo, promedio y centroide para calcular distancias entre clusters.

Interpretación

Los dendrogramas muestran relaciones y permiten definir clusters según corte deseado.

Este método es intuitivo pero costoso computacionalmente ($O(n^2 \log n)$).



DBSCAN: Clustering Basado en Densidad

Conceptos Clave

- ε: radio para vecindad
- MinPts: puntos mínimos
- Puntos centrales, frontera y ruido

Ventajas

- No requiere k predefinido
- Detecta clusters de forma arbitraria
- Manejo eficiente del ruido

Desventajas

- Parámetros sensibles
- Difícil optimización para conjuntos variables

Ejemplo Práctico

Identificación de anomalías espaciales en datos geográficos.

Evaluación del Clustering

Métricas Internas

Midiendo cohesión y separación sin etiquetas externas.

- WCSS: suma de cuadrados intra-cluster
- Coeficiente de Silueta: calidad de agrupamiento

Métricas Externas

Comparando clusters con etiquetas de referencia.

- Índice de Rand Ajustado (ARI)
- Información Mutua Normalizada (NMI)

Consideraciones

Elegir métricas según disponibilidad de datos y objetivos; algunas requieren etiquetas previas.

Aplicaciones Prácticas del Clustering

Segmentación de Clientes

Personalizando campañas y mejorando la experiencia del cliente.

Análisis de Imágenes Médicas

Detección temprana de tumores y anomalías.

Agrupación de Documentos

Facilitando búsqueda y organización eficiente.

Detección de Fraude

Identificando patrones atípicos en transacciones financieras.

Conclusión y Próximos Pasos

Resumen

Se abordaron los principales algoritmos: K-Means, jerárquico y DBSCAN con sus ventajas y limitaciones.

Consideraciones

Escoger algoritmo según tipo de datos, forma de clusters y objetivos de análisis.

Recursos

Explorar literatura, cursos y herramientas para profundizar en clustering no supervisado.

Interacción

Espacio abierto para preguntas, discusión y colaboración futura.