Machine learning avanzado

Unsupervised learning

No supervisado

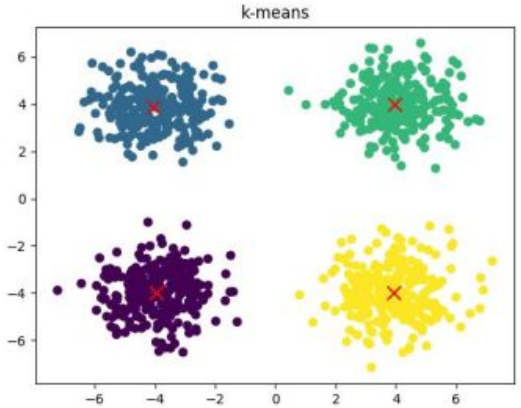
- no contamos con las etiquetas
- Qué podemos hacer? agruparlos

Clustering

- Técnica para análisis y visualización de datos
- Consiste en agrupar elementos desde un conjunto de datos no etiquetados.
- Permite identificar grupos en los datos, también posibles outliers

K-Means

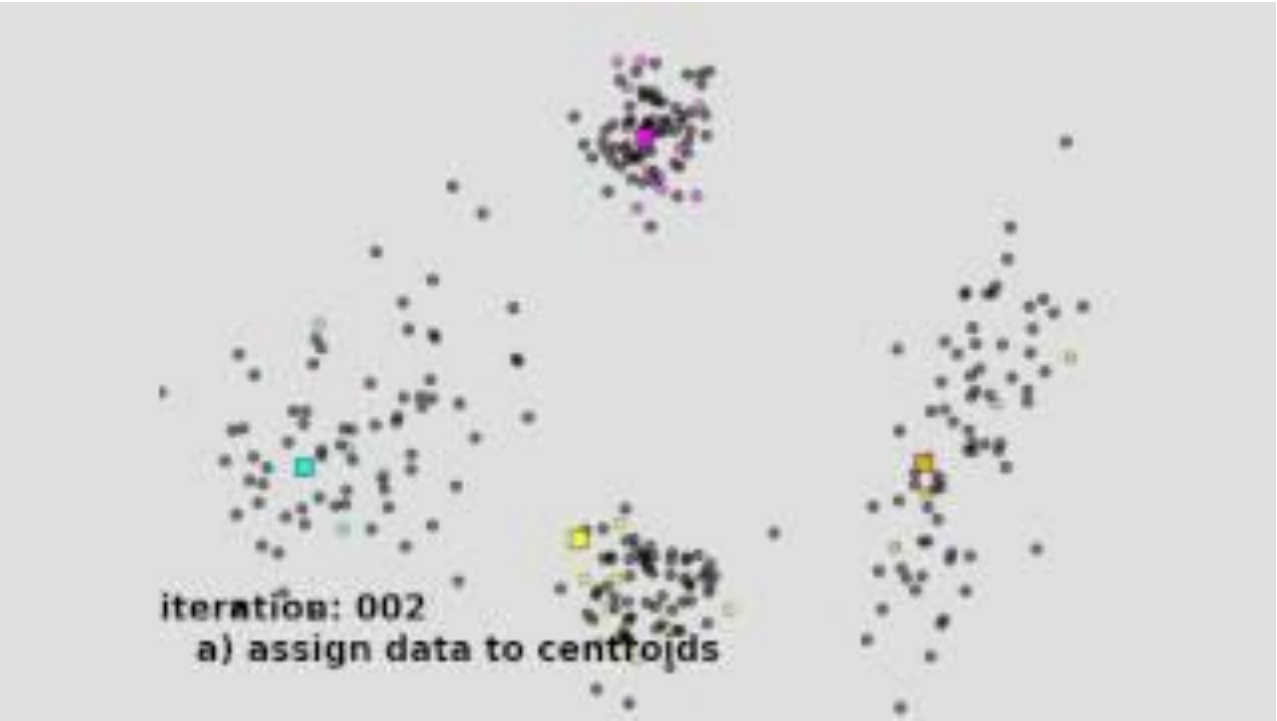
Buscar K centros y clusters, tales que cada centro sea la media o centroide de su respectivo cluster y cada elemento pertenezca a al cluster de su centro más cercano



K-Means

1. Define K elementos aleatorios como centros de los clusters
2. Asigna cada elemento del dataset al cluster de su centro más cercano
3. Recalcula los centros de cada cluster haciendo la media de la distancia a su centroide más cercano.
4. Repite 2 y 3, hasta que los centros dejen de moverse ó un número máximo de iteraciones

Funcionamiento de K-Means



Inercia

El objetivo de la clusterización es encontrar la menor inercia con el menor K número de clusters.

$$inercia = \frac{1}{N}\sum_{i=0}^{N}(x_i - u)^2$$

Sin embargo, la inercia y el K son inversamente proporcionales.

Ejercicio

Suponiendo que tenemos 4 clientes de un supermercado, de los cuales conocemos dos características para cada uno de ellos (x1 y x2).

Agrupe a los clientes en dos grupos.

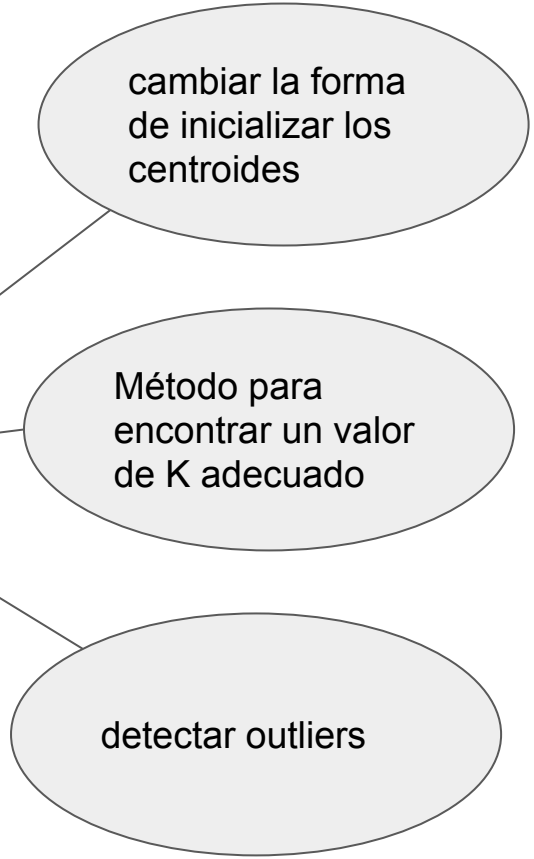
	x1	**x2**
cliente 1	1	1
cliente 2	2	1
cliente 3	4	3
cliente 4	5	4

Considere que los centros son (1,1) y (2,1)

Problemas?

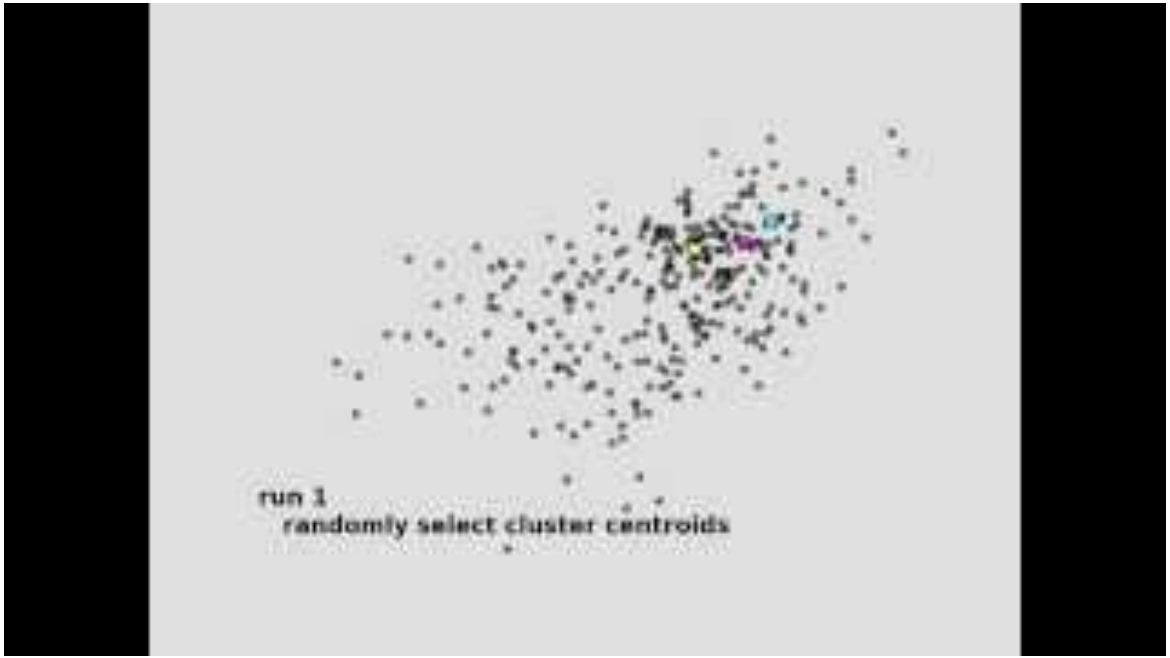
¿Qué problemas observa del algoritmo K-means?

- Dependiente de la inicialización de los clusters
- Dependiente del valor de K definido
- Los outliers afectan la ubicación de los centroides



Inicialización de clusters

random: Selecciona aleatoriamente K elementos del conjunto como centroide

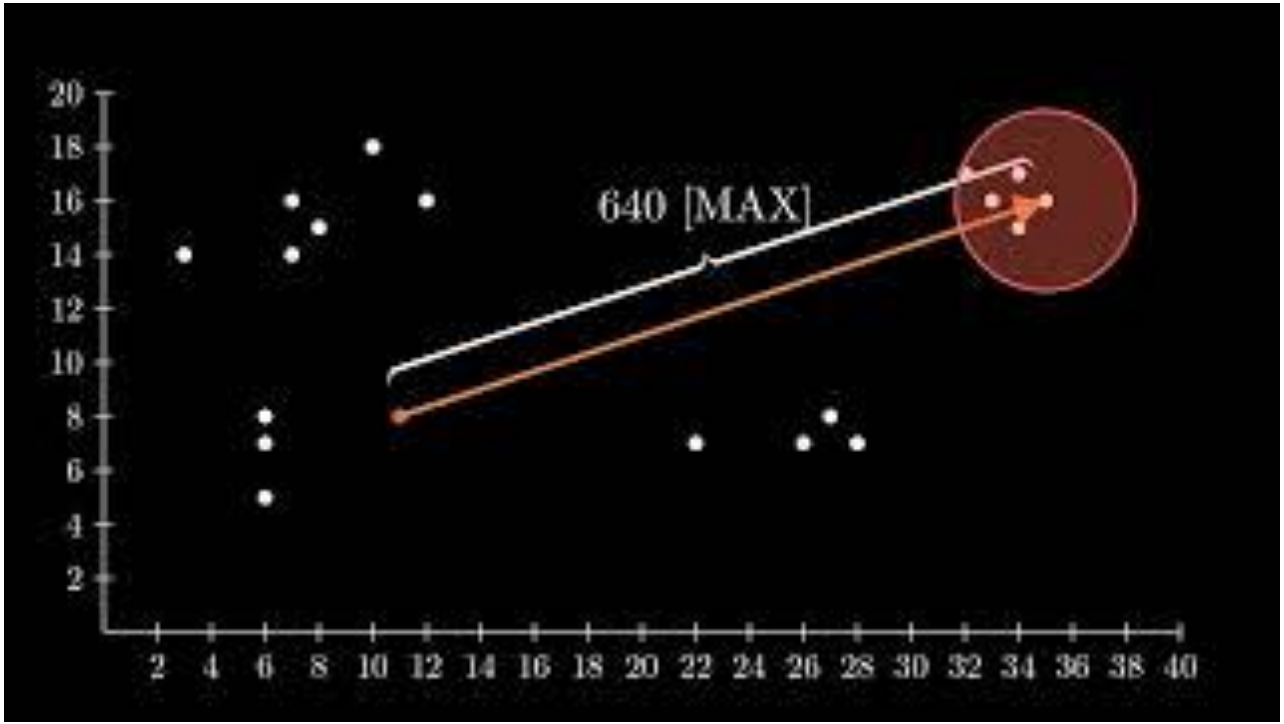


Inicialización de clusters

K-means++: Método de inicialización de centroides que realiza un preprocesado previo al algoritmo K-Means.

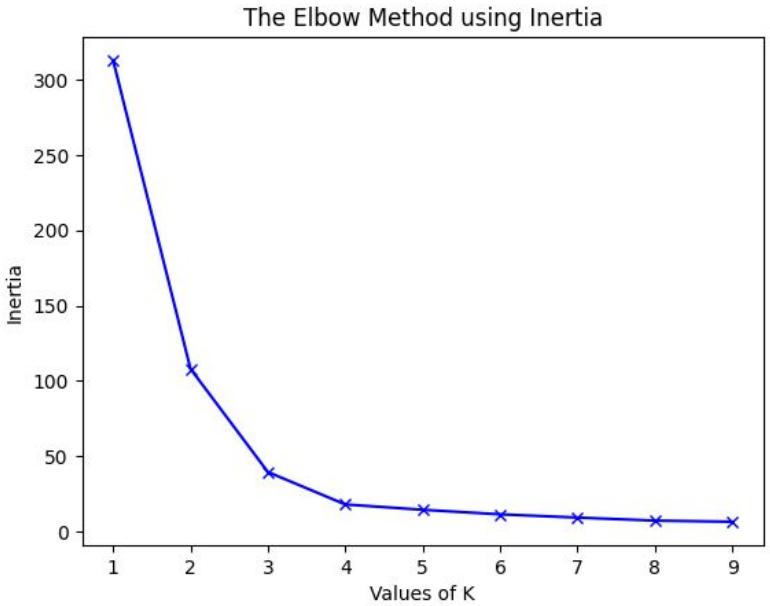
1. Elige 1 centroide aleatoriamente
2. Calcula la distancia d desde cada elemento hacia su centroide más cercano
3. Elige un nuevo centroide con probabilidad equivalente a la distancia desde el punto hacia su centroide más cercano (es decir, es más probable que aquel elemento que está más lejos de su centroide sea elegido como un nuevo centroide)
4. Repite hasta haber encontrado K puntos.

K-Means++


640 [MAX]
20
18
16
14
12
10
8
6
4
2
2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40

Problema: definir K

Elbow method: Encontrar un valor de K en el cual la medida de inercia comience a comportarse linealmente. Debe ser afinado como hiperparámetro.



Problema: Outliers

Los valores atípicos son un problema para los métodos de clusterización porque afectan el cálculo de los centroides. Para abordar este problema en K-Means, se debe hacer en la etapa de preprocesado una detección de outliers con la intención de excluirlos del conjunto sobre el cual se calcularán los nuevos centroides.

Otra forma de clusterizar: Clustering jerárquico

Es una familia de métodos de clusterización que agrupa los elementos por pares, de acuerdo a los que están más cercanos.

Se utiliza un dendrograma para visualizar los clusters, el cual es un árbol cuyas hojas son cada uno de los elementos del conjunto.

El algoritmo consiste en ir construyendo grupos que se combinan con sus grupos más cercanos, hasta encontrar el número de clusters deseado.

Clustering jerárquico

Ejemplo: usemos el ejercicio anterior y construyamos el dendrograma en la pizarra.

	x1	**x2**
cliente 1	1	1
cliente 2	2	1
cliente 3	4	3
cliente 4	5	4

DBScan

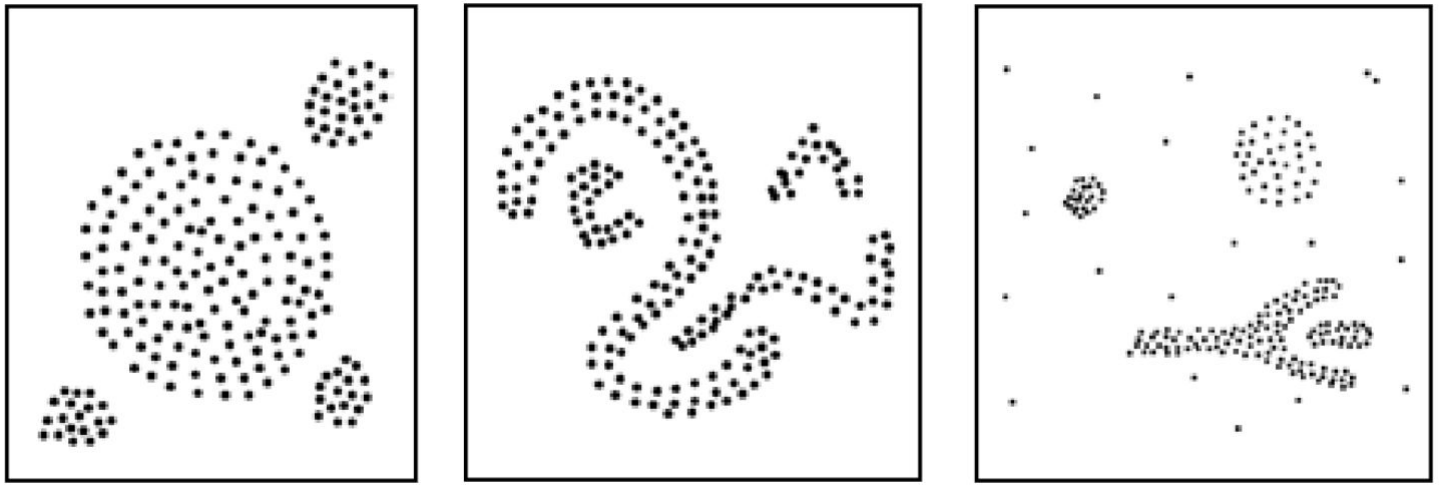
Published in Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)

A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise

Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu

Institute for Computer Science, University of Munich
Oettingenstr. 67, D-80538 München, Germany
{ester | kriegel | sander | xwxu}@informatik.uni-muenchen.de

Clusterización por densidad

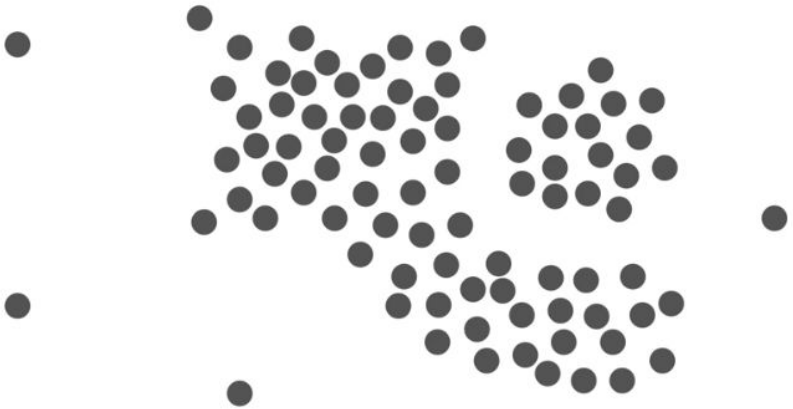


DBScan

DBScan es un método de agrupamiento por densidad

1. Definir los parámetros: El algoritmo DBSCAN requiere de dos parámetros principales: epsilon (ε), que determina la distancia máxima entre dos puntos para que sean considerados vecinos, y min_samples, que es el número mínimo de puntos dentro de un radio ε para formar un clúster.
2. Etiquetar los puntos: Se etiqueta cada punto como "núcleo", "borde" o "ruido" basándose en su vecindario y los parámetros definidos.
3. Expandir los clústeres: Se expanden los clústeres conectando puntos vecinos y creando clusters adicionales si se cumplen las condiciones establecidas.
4. Asignar puntos no etiquetados: Se asignan los puntos no etiquetados a los clústeres existentes o se los considera como ruido.

DBScan



DBScan

Desventajas de DBScan:

- Paramétrico. Su resultado depende ε y min_samples

Ventajas de DBScan:

- No requiere especificar previamente el número de clusters.
- Se comporta de forma robusta frente a outliers
- Es capaz de detectar clusters de forma arbitraria y de diferentes formas y tamaños.

Reducción de dimensionalidad

- En muchos contextos de análisis de datos, trabajar con un gran número de variables puede ser un problema.
- La maldición de la dimensionalidad, en la medida que aumenta la dimensionalidad la información va perdiendo relevancia.
- La reducción de dimensionalidad es una técnica que permite representar los datos en un espacio de menor dimensión, manteniendo la mayor cantidad posible de información relevante.

PCA (Principal Component Analysis):

- PCA es una técnica popular de reducción de dimensionalidad que busca transformar los datos originales en un nuevo conjunto de variables llamadas componentes principales.
- Los componentes principales se obtienen de manera que la primera componente capture la mayor varianza de los datos, la segunda componente capture la segunda mayor varianza, y así sucesivamente.

Pasos para aplicar PCA

1. Normalización de datos: Es importante estandarizar los datos para asegurar que todas las variables tengan la misma escala.
2. Cálculo de la matriz de covarianza: Se calcula la matriz de covarianza para entender las relaciones lineales entre las variables originales.
3. Cálculo de los autovalores y autovectores: Se obtienen los autovalores y autovectores de la matriz de covarianza.
4. Selección de componentes principales: Se seleccionan los componentes principales basándose en los autovalores más altos.
5. Proyección de los datos: Los datos originales se proyectan en el nuevo espacio de componentes principales.

Cómo lo interpretamos?

- Cada componente principal es una combinación de las variables originales.
- Los componentes principales se pueden interpretar como nuevas características que capturan la mayor parte de la varianza en los datos.
- Los componentes principales se ordenan en función de su contribución a la varianza total, por lo que los primeros componentes son los más informativos.

Ventajas

- Reduce la dimensionalidad de los datos, lo que puede mejorar la eficiencia computacional y reducir la complejidad del modelo.
- Ayuda a identificar patrones y relaciones ocultas en los datos al enfocarse en las variables más informativas.
- Permite visualizar los datos en un espacio de menor dimensión, facilitando la interpretación y la exploración.

T-SNE

- Método de reducción de dimensional principalmente pensado en la visualización de los datos
- PCA es un método que busca la máxima varianza en los datos, pero podría no mantener el vecindario y la configuración espacial dada.

T-SNE (t-Distributed Stochastic Neighbor Embedding)

- t-SNE es un algoritmo que preserva las relaciones locales entre los puntos en el espacio de alta dimensión al mapearlos en un espacio de menor dimensión.
- Utiliza técnicas de probabilidad para calcular similitudes entre pares de puntos y modelar la distribución conjunta en el espacio de alta dimensión y en el espacio reducido.
- El algoritmo busca minimizar la entropía de los datos en el nuevo espacio minimizando la KL-divergence.

T-SNE (t-Distributed Stochastic Neighbor Embedding)

- El perplexity es un parámetro clave en el algoritmo t-SNE (t-Distributed Stochastic Neighbor Embedding). Es un valor que controla el equilibrio entre la conservación de las estructuras locales y globales durante la reducción de dimensionalidad.
- **El perplexity se refiere a la idea de la incertidumbre o la perplejidad de una distribución de probabilidad. En el contexto de t-SNE, se utiliza para determinar cuántos vecinos cercanos se consideran al calcular las similitudes entre puntos en el espacio de alta dimensión.**

T-SNE (t-Distributed Stochastic Neighbor Embedding)

- Cuando se establece el valor del perplexity en t-SNE, el algoritmo busca seleccionar una distribución de vecinos cercanos para cada punto de manera que la entropía de la distribución objetivo sea aproximadamente igual al perplexity especificado.
- Un valor bajo de perplexity dará lugar a una estructura más focalizada y local en la visualización, enfatizando las similitudes locales entre puntos vecinos. Por otro lado, un valor alto de perplexity permitirá capturar estructuras más globales y suaves en la visualización.