Short Course. Análisis No Supervisado

Víctor Aceña Gil

Data Science Lab. URJC

Enero 2023

Índice

1. Entender el problema. Objetivos de negocio
2. Técnicas
3. Paso a paso funcional
4. Para recordar

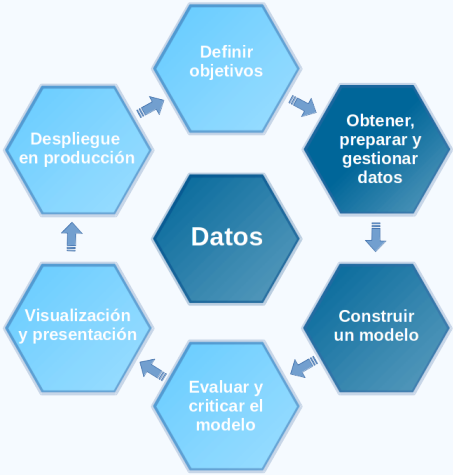
Entender el problema. Objetivos de negocio

¿Qué es el Análisis No Supervisado?

El aprendizaje automático que utiliza datos no etiquetados se denomina aprendizaje no supervisado.

- A veces, la tarea consiste en encontrar cualquier relación que exista el conjunto de datos (**reglas de asociación**).
- Los algoritmos de agrupación o **Clustering** examinan datos para encontrar grupos de observaciones que sean más similares entre sí que los objetos de otros grupos.

Ciclo de la Ciencia de Datos



El análisis clustering es una herramienta potente generalmente utilizada para agrupar y segmentar observaciones o clientes en función de sus similitudes.

- En un entorno empresarial, este análisis puede utilizarse para obtener información valiosa sobre el comportamiento y las preferencias de los clientes, que luego puede utilizarse para mejorar las estrategias de marketing, optimizar las ventas y la distribución, y aumentar la satisfacción general de los clientes. Al identificar grupos de clientes similares, las empresas pueden dirigirse a grupos específicos con promociones y ofertas a medida, lo que se traduce en esfuerzos de marketing más eficaces y eficientes. Además, el análisis clustering puede utilizarse para identificar nuevos segmentos de mercado potenciales y oportunidades de expansión.

- **Segmentación del mercado**: La agrupación puede utilizarse para segmentar un mercado en diferentes grupos de clientes con características o comportamientos similares.
- **Perfiles de clientes**: El clustering puede utilizarse para crear perfiles de clientes basados en sus datos demográficos, historial de compras y otra información relevante.
- **Gestión de inventarios**: El clustering puede utilizarse para agrupar productos similares y optimizar las decisiones de gestión de inventario.

- **Evaluación de riesgos**: El clustering puede utilizarse para identificar y agrupar clientes o transacciones con mayor riesgo de impago o fraude.
  - **Detección de fraudes**: El clustering puede utilizarse para identificar patrones de comportamiento fraudulento agrupando transacciones o clientes con características sospechosas.
- **Detección de anomalías**: La agrupación puede utilizarse para detectar valores atípicos o anomalías en grandes conjuntos de datos, como tráfico de red, datos de sensores, etc.
- **Sistemas de recomendación**: La agrupación puede utilizarse para crear sistemas de recomendación que sugieran productos o servicios a los clientes en función de su historial de compras u otras características.

Técnicas

K-means: Se trata de un algoritmo muy utilizado que divide un conjunto de datos en k grupos, cada uno de ellos definido por la media (o "centroide") de los puntos del grupo.

- Fácil de entender y aplicar
- Sensible a la ubicación inicial de los centroides y puede no funcionar bien con conglomerados no esféricos o circulares.

Agrupaciones jerárquicas: Se construye un árbol jerárquico de conglomerados, donde cada nodo del árbol representa un conglomerado y las hojas representan puntos de datos individuales.

- Puede utilizarse para identificar clusters anidados y puede manejar bien clusters no esféricos.

DBSCAN (Agrupación espacial de aplicaciones con ruido basada en la densidad): Agrupa las observaciones que están muy cercanas y separa aquellas que están más alejadas.

- No requiere que se especifique de antemano el número de clusters y puede manejar clusters de diferentes formas y densidades.

Gaussian Mixture Model (GMM): Se trata de un modelo probabilístico que asume que los datos se generan a partir de una mezcla de distribuciones gaussianas.

- Puede utilizarse para identificar conglomerados de diferentes formas y tamaños, así como para estimar la densidad.

Algoritmo de Expectation-Maximization (EM) Se trata de un método que funciona con datos incompletos o inciertos, en el que la información que falta se infiere a partir de los datos.

- Se puede utilizar para estimar los parámetros de un modelo GMM.

Éstas son algunas de las técnicas de agrupación más populares y utilizadas, pero existen muchos otros algoritmos y enfoques. La elección de la técnica dependerá de las características específicas de los datos y de los objetivos del análisis.

Prácticas

Enlace al material: aquí.

Paso a paso funcional

- **Preparación de los datos**: Recoger y preprocesar los datos que se utilizarán para el análisis. Esto incluye la limpieza de los datos, el tratamiento de los valores faltantes y la transformación de los datos si es necesario. **EDA**.
- **Selección de características**: Seleccionar las características o variables que se utilizarán para el análisis. Este paso es importante para garantizar que las variables utilizadas son pertinentes e informativas.
- **Distancia o métrica de similitud**: Seleccionar una distancia o métrica de similitud que se utilizará para medir la similitud o disimilitud entre las observaciones. Las métricas más utilizadas son la distancia Euclídea, la distancia Manhattan y la similitud coseno.
- **Seleccionar el número de grupos**: Decidir el número de grupos que se desea formar. Pueden emplearse técnicas como el método del codo, la puntuación de silueta, etc.

- **Entrenamiento del modelo**: Entrenar el modelo de clustering utilizando los datos seleccionados, las características y la métrica de distancia. Se pueden utilizar diferentes algoritmos como K-means, Hierarchical, DBSCAN, etc.
- **Evaluación**: Evaluar el rendimiento del modelo de clustering comparando los clusters predichos con las etiquetas verdaderas o utilizando métricas como la puntuación de silueta o el índice de Davies-Bouldin.
- **Interpretación**: Interpretar los resultados del análisis de clustering analizando las características de los clusters e identificando patrones o tendencias en los datos.
- **Visualización**: Visualizar los resultados del análisis no supervisado utilizando diagramas y gráficos para ayudar a comunicar las lecciones aprendidas.

Los pasos pueden variar dependiendo del algoritmo de clustering específico utilizado, el conjunto de datos y el problema que se esté intentando resolver.

Para recordar

Clustering Cheat Sheet: aquí.