Cuestiones y terminología de la Ciencia de Datos

Christopher González Delgado

1. Cuestiones y Términos III

1.1. ¿Qué es el clustering?

Es un tipo de algoritmo que entra en el aprendizaje no supervisado, es decir, simplemente tenemos datos de entrada sin etiquetar. Clustering es buscar aspectos comunes entre los datos y particionar las entradas en diferentes grupos individuales con características comunes.

1.2. ¿Qué diferencias hay entre clustering y clasificación?

La principal diferencia que encontramos es que la clasificación es un tipo de aprendizaje supervisado y el clustering es no supervisado. En concreto, la clasificación estudia las diferentes características de los datos de entrada y tratamos de encajar los datos en una categoría única y predefinida (¡en el aprendizaje supervisado todo está etiquetado!), mientras que en clustering buscamos hacer grupos de datos con características similares.

1.3. ¿Qué es un Cluster?

Un cluster es un grupo de objetos que son similares a los objetos que se encuentran en un mismo cluster, pero que son diferentes a los objetos en el resto de clusters. El clustering consiste en partir los datos y agruparlos en diferentes clusters.

1.4. Tres tipos de algoritmos de clustering conocidos

- K-Means
- Clustering jerárquico
- Clustering basado en densidad

1.5. K-means

Es uno de los algortimos que se pueden utilizar en clustering. El clustering usando K-means consiste en agrupar datos basándose en la similaridad de los datos de entrada entre sí. Los datos se dividen en K clusters que no se

solapen entre sí y sin etiquetar.
Las distancias entre los objetos de cada cluster es minimizada, mientras que la distancia entre distintos clusters es maximizada.

1.6. Funcionamiento del K-means

El primer paso en el desarrollo del algoritmo es elegir el número K de clusters. Se inicializan aleatoriamente K **centroides**, que son los puntos centrales de cada cluster. Una vez hecho esto, se calculan las distancias y se asigna un cluster a cada punto del set de datos según la distancia a cada centroide. A continuación se computan los nuevos centroides para cada cluster. Para ello se calcula el punto medio de cada cluster y este pasará a ser el nuevo centroide.
Este proceso se repite hasta que no ocurran más cambios en la distribución de puntos y asignación de centroides.

1.7. Clustering jerárquico

El clustering jerárquico es otro tipo de algoritmo que se utiliza, evidentemente, en el clustering. El funcionamiento de este método es el siguiente: en primer lugar cada muestra o cada dato se toma como un cluster o grupo individual. A continuación, se unen los clusters más similares (los más cercanos) en un nuevo cluster de mayor tamaño, creando así un proceso iterativo que culmina en un gran grupo genérico del que, jerárquicamente, aparecen los clusters más pequeños hasta finalmente alcanzar los datos individuales.

1.8. Clustering basado en densidad

El clustering basado en densidad funciona de forma eficaz cuando la distribución de datos es aleatoria o cuando la forma de los clusters no es suficientemente esférica. Estos algoritmos detectan en qué areas se concentran los puntos y dónde hay áreas vacías. También detecta puntos de ruido o outliers.

1.9. ¿Qué son los datos?

Los datos son información desorganizada que se utilizan con el fin de darles significado. Los datos abarcan observaciones, percepciones, números,

caracteres, símbolos o incluso imágenes que pueden ser interpretadas para derivar un significado.

1.10. ¿Cómo se clasifican los datos según su estructura?

Estructurados: Datos que siguen un formato rígido y que pueden ser organizados limpiamente en filas y columnas. E.g. bases de datos, hojas de cálculo...

Semi-estructurados: Mezcla de datos que tienen unas características consistentes pero que además tiene datos que no conforman una estructura sólida. E.g. un email, tiene asunto o remitente (estructurado) pero el contenido del email es no estructurado.

No estructurados: Datos que son complejos y principalmente cualitativos que no pueden ser reducidos a columnas y filas. E.g fotos o videos.

1.11. Formatos de datos estándar

- Formatos de archivos de texto delimitado (.CSV): Archivos usados para almacenar datos como texto. Los valores se separan por algún delimitador.
- Open Microsoft Excel (XLS) y (XLSX):Es un tipo de archivo que se encuentra bajo el formato de una hoja de cálculo. Se basa en XML. Se compone de varias hojas de trabajo y cada hoja de trabajo está organizada en filas y columnas.
- Extensible Markup Language o Lenguaje de Marcado Extensible (.XML): Es un tipo de lenguaje de marcado con unas reglas establecidas para codificar los datos. En algunos aspectos es similar al HTML. Es un tipo de archivo auto-descriptible que se utiliza para enviar información por internet.
- Formato de documento portátill (.PDF): Es un formato de archivo ideado para presentar documentos independientemente del software, hardware y del sistema operativo.

1.12. ¿De qué fuentes se pueden obtener datos?

- Bases de datos relacionales
- Bases de datos no relacionales
- APIs
- Servicios Web
- Plataformas o redes sociales
- Sensores o dispositivos

1.13. ¿Qué es un repositorio de datos?

Un repositorio de datos es un término que incluye bases de datos, almacenes de datos, mercados de datos, lagos de datos y almacenes de Big Data.

1.14. Lenguajes utilizados en el ámbito del análisis de datos

SQL para **consultas (querys)**. Para hacer consultas y manipular datos se suele utilizar SQL.
Python para **desarrollar aplicaciones de datos**.
Shell y Scripting para **t areas operacionales repetitivas**

1.15. ¿Qué son APIs?

Una API (Application Programming Interface) es un servicio con el que los diferentes usuarios o aplicaciones pueden interactuar y obtener datos para su procesamiento o análisis. Estos servicios (APIS) obtienen peticiones o solicitudes de usuarios (solicitudes web) o de aplicaciones (solicitudes de red) y devuelven datos en diferentes tipos de formato.

1.16. ¿Qué es Web Scraping?

El web scraping es una técnica que se utiliza para extraer datos e información relevante de fuentes no-estructuradas. A través del web scraping podemos obtener datos específicos basándonos en parámetros definidos.