Clustering

También conocido como agrupamiento, es una de las técnicas de minería de datos, el proceso consiste en la división de los datos en grupos de objetos similares.
Las técnicas de Clustering son las que utilizando algoritmos matemáticos se encargan de agrupar objetos. Usando la información que brindan las variables que pertenecen a cada objeto se mide la similitud entre los mismos, y una vez hecho esto se colocan en clases que son muy similares internamente y a la vez diferente entre los miembros de las diferentes clases.
Un clúster es una colección de objetos de datos. Similares entre sí dentro del mismo grupo. Disimilar a los objetos en otros grupos.
Análisis de clúster: dado un conjunto de puntos de datos tratar de entender su estructura. Encuentra similitudes entre los datos de acuerdo con las características encontradas en los datos. Es un aprendizaje no supervisado ya que no hay clases predefinidas.

Regla de asociación

Las reglas de asociación se utilizan para descubrir hechos que ocurren en común dentro de un determinado conjunto de datos. Se han investigado ampliamente diversos métodos para aprendizaje de reglas de asociación que han resultado ser muy interesantes para descubrir relaciones entre variables en grandes conjuntos de datos.

Se describe el análisis y la presentación de reglas 'fuertes' descubiertas en bases de datos utilizando diferentes medidas de interés. Por ejemplo

$$\{cebollas, vegetales\} \Rightarrow \{carne\}$$

Encontrada en los datos de ventas de un supermercado, indicaría que un consumidor que compra cebollas y verdura a la vez, es probable que compre también carne. Esta información se puede utilizar como base para tomar decisiones sobre marketing como precios promocionales para ciertos productos o dónde ubicar estos dentro del supermercado. Además del ejemplo anterior aplicado al análisis de *la cesta de la compra*, hoy en día, las reglas de asociación también son de aplicación en otras muchas áreas como el web mining, la detección de intrusos o la bioinformática.

Detección de outliers

También llamados datos atípicos, un valor atípico (en inglés outliers) es una observación que es numéricamente distante del resto de los datos. Las estadísticas derivadas de los conjuntos de datos que incluyen valores atípicos serán frecuentemente engañosas. Porque pueden afectar considerablemente a los resultados que pueda obtener un modelo de Machine Learning. Para mal o para bien. Por eso hay que detectarlos, y tenerlos en cuenta.
Por ejemplo en regresión lineal o algoritmos de ensamble puede tener un impacto negativo en sus predicciones. También a la hora de analizar algún resultado en un examen médico, sea una radiografía, análisis de sangre etcétera. Estos pueden significar un claro problema de salud que debe ser atendido.

Visualización

Las visualizaciones son una herramienta fundamental para entender y compartir ideas sobre los datos. La visualización correcta puede ayudar a expresar una idea central, o abrir un espacio para una más profunda investigación; con ella se puede conseguir que todo el mundo hable sobre un conjunto de datos, o compartir una visión sobre lo que los datos nos quieren decir.

Una buena visualización puede dar a quien la observa un sentido rico y amplio de un conjunto de datos. Puede comunicar los datos de manera precisa a la vez que expone los lugares en dónde se necesita más información o dónde una hipótesis no se sostiene. Por otra parte, la visualización nos proporciona un lienzo para aplicar nuestras propias ideas, experiencias y conocimientos cuando observamos y analizamos datos, permitiendo realizar múltiples interpretaciones. Si como dice el dicho *"una imagen vale más que mil palabras"*, un gráfico interactivo bien elegido entonces podría valer cientos de pruebas estadísticas.

Regresión

Es un modelo estadístico que permite establecer un patrón que describe la relación lineal entre variables. Su función principal es examinar y predecir una variable a partir de otra(s). Es por ello que en las tareas de aprendizaje automático cuyo objetivo principal es la estimación del valor, se pueden denominar *tareas de la regresión*.
Los métodos basados en regresión, están entrenados en muestras de datos de entrada que tienen respuestas de salida como valores numéricos continuos; a diferencia de la clasificación, donde tenemos categorías o clases discretas. Los modelos de regresión se valen de atributos o características de datos de entrada (también llamadas variables explicativas o independientes), y sus correspondientes valores de salida numéricos continuos (también llamados como respuesta, dependiente o variable de resultado), para aprender relaciones y asociaciones específicas entre las entradas y sus salidas correspondientes.

Clasificación

También conocidos como arboles de decisión. Los arboles de decisión son representaciones gráficas de posibles soluciones a una decisión basadas en ciertas condiciones, es uno de los algoritmos de aprendizaje supervisado más utilizados en machine Learning y pueden realizar tareas de clasificación o regresión (acrónimo del inglés CART). La comprensión de su funcionamiento suele ser simple y a la vez muy potente.
Utilizamos mentalmente estructuras de árbol de decisión constantemente en nuestra vida diaria sin darnos cuenta:
¿Llueve? => lleva paraguas. ¿Soleado? => lleva gafas de sol. ¿estoy cansado? => toma café. (decisiones del tipo IF THIS THEN THAT)
Los árboles de decisión *tienen un primer nodo llamado raíz* (root) y luego se descomponen el resto de atributos de entrada en dos ramas (podrían ser más, pero no nos meteremos en eso ahora) planteando una condición que puede ser cierta o falsa. Se bifurca cada nodo en 2 y vuelven a subdividirse hasta llegar a las hojas que son los nodos finales y que equivalen a respuestas a la solución: Si/No, Comprar/Vender, o lo que sea que estemos clasificando.

Patrones secuenciales

En el campo de la minería de datos y extracción de conocimiento, la minería de secuencias es un caso particular de la minería de datos estructurados. Consiste en encontrar patrones estadísticamente relevantes en colecciones de datos que están representados de forma secuencial.[1] Debido a la frecuencia con que aparecen este tipo de datos en escenarios de aplicaciones reales, esta técnica constituye uno de los métodos más populares de descubrimiento de patrones.

Los patrones frecuentes obtenidos durante el minado de secuencias, se usan en tareas de detección de dependencias funcionales, predicción de tendencias, interpretación de fenómenos y como soporte de decisiones en estrategias de producción
Las secuencias son un importante tipo de datos, que representa una clase especial de estructura donde importa el orden que ocupan los elementos (conjunto de ítems del inglés itemset). Este orden, puede o no estar relacionado con el factor tiempo. El tamaño de una secuencia va a estar dado por la cantidad de itemsets que contenga, mientras que su longitud por la cantidad de ítems.

Ejemplos típicos de secuencias, son las secuencias genéticas como las de ADN y ARN. Un caso menos trivial es la representación de los registros de compras a partir de secuencias. Este patrón secuencial relacionaría las transacciones de un grupo de clientes a lo largo del tiempo y las asociaciones que se obtendrían al minar estos datos, mostrarían las dependencias entre los productos en el intervalo de tiempo analizado.

Predicción

La predicción en minería de datos la podemos definir como: analizar valores históricos reales para darnos información sobre acontecimientos no conocidos o futuros.
La predicción más que una técnica es una clasificación de las técnicas dentro de la minería de datos.
La minería de datos y los modelos predictivos son la base del conocimiento empresarial. Su fin es buscar patrones en grandes volúmenes de datos que aporten valor a la organización y a su estrategia. Ahora bien, ¿qué aspectos debemos tener en cuenta?
Hoy en día, la minería de datos se sirve de la inteligencia artificial y del machine learning o aprendizaje automático, lo que potencia su alcance y el impacto que puedan tener los modelos que resultan del entrenamiento de los algoritmos con datos y más datos. Es por ello que siempre partimos de una correcta gestión de datos, para que estos puedan llevarnos al siguiente nivel.