APUNTES LIBRO

Introducción a la Minería de Datos

Alvarado Becerra Ludwig

Libro de José Hernández Orallo

24 de mayo de 2024

Índice

1. ¿Qué es la minería de datos?

La minería de datos tiene su fundemantación en la estadística y en el *Machine Learning*. Utiliza grandes cantidades de datos para generar conocimiento a partir de los datos, mientras que en otras ramas de estadística la idea era organizar los datos y obtener información a partir de ella a partir de consultas como lo era en el caso de SQL. En la minería de datos se buscan patrones que puedan ayudarnos a generar conocimiento a partir de esos datos, es ampliamente útil en todo tipo de industria, ingeniería, campo de la salud, todo lado!

Según el autor, el objetivo de la minería de datos es convertir datos en conocimiento. Para eso mismo, expone unos ejemplos, de los cuales voy a detallar sólo en uno que fue el que más me gustó ;)

1.1. Determinar grupos diferenciados de empleados

El departamento de recursos humanos de una gran empresa desea categorizar a sus empleados en distintos grupos con el objetivo de entender mejor su comportamiento y tratarlos de manera adecuada. Para ello dispone en sus bases de datos de información sobre los mismos (sueldo, estado civil, sii tiene coche, número de hijos, si su casa es propia o de alquier, si está sindicado, número de bajas al año, antigüedad y sexo). La tabla 1 muestra algunos de los registros de su base de datos.

id	**Sueldo**	**Casado**	**Coche**	**Hijos**	**Alq/-prop**	**Sindicado**	**Bajas/Año**	**Antigüedad**	**Sexo**
1	1.000	Sí	No	0	Alquiler	No	7	15	H
2	2.000	No	Sí	1	Alquiler	Sí	3	3	M
3	1.500	Sí	Sí	2	Prop	Sí	5	10	H
4	3.000	Sí	Sí	1	Alquiler	No	15	7	M
5	1.000	Sí	Sí	0	Prop	Sí	1	6	H
6	4.000	No	Sí	0	Alquiler	Sí	3	16	M
7	2.500	No	No	0	Alquiler	Sí	0	8	H
8	2.000	No	Sí	0	Prop	Sí	2	6	M
9	2.000	Sí	Sí	3	Prop	No	7	5	H
10	3.000	Sí	Sí	2	Prop	No	1	20	H
11	5.000	No	No	0	Alquiler	No	2	12	M
12	800	Sí	Sí	2	Prop	No	3	1	H
13	2.000	No	No	0	Alquiler	No	27	5	M
14	1.000	No	Sí	0	Alquiler	Sí	0	7	H
15	800	No	Sí	0	Alquiler	No	3	2	H

⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Cuadro 1: Datos de los empleados

Un sistema de minería de datos podría obtener tres grupos con la siguiente descripción:

Grupo 1:

- Sueldo: 1.535,2 €.
- Casado: No -> 0,777 — Sí -> 0,223
- Coche: No -> 0,82 — Sí -> 0,18
- Hijos: 0,05
- Alq/Prop: Alquiler -> 0,99 — Propia -> 0,01
- Sindic.: No -> 0,8 — Sí -> 0,2
- Bajas/Año: 8,3
- Antigüedad: 8,7
- Sexo: H -> 0,61 — M -> 0,39

Grupo 2:

- Sueldo: 1.428,7 €.
- Casado: No -> 0,98 — Sí -> 0,02
- Coche: No -> 0,01 — Sí -> 0,99
- Hijos: 0,3
- Alq/Prop: Alquiler -> 0,75 — Propia -> 0,25
- Sindic.: Sí -> 1,0
- Bajas/Año: 2,3
- Antigüedad: 8
- Sexo: H -> 0,25 — M -> 0,75

Grupo 3:

- Sueldo: 1.233,8 €.
- Casado: Sí -> 1,0
- Coche: No -> 0,05 — Sí -> 0,95
- Hijos: 2,3
- Alq/Prop: Alquiler -> 0,17 — Propia -> 0,83
- Sindic.: No -> 0,67 — Sí -> 0,33
- Bajas/Año: 5,1
- Antigüedad: 8,1
- Sexo: H -> 0,83 — M -> 0,17

Estos grupos podrían ser interpretados por el departamento de recursos humanos de la siguiente manera:

- Grupo 1: sin hijos y con vivienda de alquiler. Poco sindicados. Muchas bajas.
- Grupo 2: sin hijos y con coche. Muy sindicados. Pocas bajas. Normalmente son mujeres y viven en casas de alquiler.
- Grupo 3: con hijos, casados y con coche. Mayoritariamente hombres propietarios de su vivienda. Poco sindicados.

1.2. Tipos de datos

La minería de datos se puede aplicar a cualquier tipo de información, lo que cambia son las técnicas que se aplican a los datos. Se diferencian entre datos relacionados y no estructurados.

1.2.1. Bases de datos relacionales

Estas son las que ya conozco de toda la vida para SQL, utilizadas en todo lado y según el libro, los tipos de datos estructurados y por tablas son los que más se utilizan para aplicar técnicas de minería de datos. La integridad de los datos se expresa a través de las restricciones de integridad (las *constraints* de toda la vida). Primero, los datos se sacan de una consulta SQL y después pasan a un modelo de minería de datos. A aquella información sacada de varias tablas que se requiere para hacer una tarea de minería de datos, se le conoce como *vista minable* y sale de una consulta SQL.

En las presentaciones tabulares se expresa la importancia de tener dos tipos de atributos; numéricos y categóricos.

- **Numéricos**: contienen valores enteros o reales. Por ejemplo, atributos como el salario o la edad.

- **Categóricos o nominales**: Toman valores en un conjunto finito y preestablecido de categorías. Por ejemplo, sexo (H, M), nombre del departament, etc.

1.2.2. Otros tipos de bases de datos

Las relacionales son las más utilizadas hoy en día, pero existen otras, como:

- **Bases de datos espaciales:** Contienen información relacionada cno el espacio físico. Cosas geográficas, imágenes médicas, redes de transporte o información de tráfico, etc. **relaciones espaciales**. La minería permite encontrar patrones entre estos datos, características de una casa en zona montañosa, planificación de nuevas líneas de metro en función de la distancia de las distintas áreas a las líneas existentes, etc. A mí se me ocurre hacer cosas con *Open Street map*, pueden estar interesantes, ya que, los datos ya se tienen y puede estar todo muy chevere con eso.
- **Temporales:** Almacenan datos que incluyen atributos relacionados con el tiempo. La minería se utiliza para encontrar características de evolución o tendencia en el cambio de valores en la base de datos.
- **Documentales:** Descripciones para los objetos (documentos de texto). Incluyen documentos no estructurados, semi-estructurados o estructurados. La minería se utiliza para obtener asociaciones entre contenidos.
- **Multimedia:** Almacenan imágenes, audio y vídeo. Soportan gran cantidad de almacenamiento (giga bytes y todo eso). La minería se utiliza para integrar métodos de búsqueda y almacenamiento.

1.2.3. *World Wide Web*

Este es el mayor repositorio de información que se puede tener, sin embargo, presenta muchos problemas para ser utilizada y empleada. Se categorizan en 3 campos:

- **Minería del contenido:** Encontrar patrones de los datos de las páginas web.
- **Minería de la estructura:** Estructura se define como hipervínculos y URLs.
- **Minería del uso:** Navegación que hace el usuario por la web.

1.3. Tipos de modelos

Existen dos tipos:

- **Predictivos:** Pretenden predecir o estimar valores futuros para variables de interés, denominadas *variables objetivo*, usando otras variables llamadas *variables independientes*. Por ejemplo, estimar la demanda de un grupo en función del gasto que se hizo en publicidad.
- **Descriptivo:** Identifican patrones que explican o resumen los datos, examinar las propiedades de los datos examinados, no para predecir nuevos datos. Por ejemplo, agencia de viajes que agrupa a sus usuarios para así poder enviar información a esos grupos.

1.4. La minería de datos y el proceso de descubrimiento de conocimiento en bases de datos

El término que se utiliza frecuentemente es el "Análisis (inteligente) de datos" que hace hincapié en las técnicas de análisis estadístico. Otro muy utilizado es el "Descubrimiento de conocimiento en bases de datos" (*Knowledge Discovery in Databases, KDD*). Este último se define como "el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos". Entonces, las propiedades que debe tener el conocimiento extraído deben ser:

- **Valido**: Los patrones debe ser precisos para datos nuevos (se contempla un grado de incertidumbre).
- **Novedoso:** Aporte algo desconocido al sistema y preferiblemente para el usuario.
- **Útiles:** El conocimiento debe conducir a acciones que reporten algún tipo de beneficio para el usuario.
- **Comprensible:** Si se tiene lo contrario se puede dificultar la interpretación, revisión, validación y uso en la toma de decisiones. La información incomprensible no proporciona conocimiento.

El KDD entonces permite la obtención de modelos o patrones, también la evaluación y posible interpretación de los mismos. El proceso de KDD sigue la siguiente estructura.

1. Sistema de información.
2. Preparación de los datos.
3. Patrones.
4. Evaluación/Interpretación/Visualización.
5. Conocimiento.