**Que es data mining**

Data Mining como un conjunto de técnicas estadísticas No existe una única definición del término Data Mining (DM). Se puede decir que DM se refiere a un conjunto de métodos estadísticos que proporcionan información (correlaciones o patrones) cuando se dispone de muchos datos (de aquí viene el nombre Minería de Datos). Esta idea de DM lleva a la siguiente estructura de conocimiento:

Datos + Estadística → Información

Generalmente, el Data Mining es el proceso de analizar datos desde diferentes perspectivas con el objetivo de resumir los datos en segmentos de información útiles. Esta información que puede ser usada para incrementar réditos o beneficios, reducir costos, etc. El DM permite a los usuarios analizar datos desde diferentes dimensiones o ángulos, categorizándolos y resumiendo las relaciones identificadas.

Además el DM permite trabajar con grandes cantidades de observaciones (varios millones) sin ningún inconveniente. También permite tratar una gran cantidad de variables predictivas (hasta varios millares). Esto último es de gran utilidad para seleccionar variables (determinar las más útiles dentro de una gran masa)

**Algunas cosas que se puede hacer con el DM**

El usuario del DM usualmente busca los siguientes cuatro tipos de relaciones:

**Clases**: las observaciones se asignan a grupos predeterminados. El proceso de clasificación consiste en asignar un conjunto de datos a grupos fijados de manera que se minimice la probabilidad de una clasificación errónea. Por ejemplo, un problema típico de clasificación es el de dividir una base de datos de bancos en grupos que sean lo más homogéneos posibles con respecto a variables como posibilidades de crédito en términos de valores tales como bueno o malo

**Clusters**: se construyen grupos de observaciones similares según un criterio prefijado. El proceso de clustering (agrupamiento) consiste en subdividir un conjunto de datos en grupos mutuamente excluyentes de tal manera que cada miembro de un grupo esté lo más cercano posible a otro elemento, y grupos diferentes estén lo más lejos posible entre sí, de modo que la distancia está medida respecto a todas las variables disponibles. Un típico ejemplo de aplicación de clustering es la clasificación de segmentos de mercado. Por ejemplo, una empresa quiere introducirse en el mercado de bebidas alcohólicas, pero antes hace una encuesta de mercado para averiguar si existen grupos de clientes con costumbres particulares en el consumo de bebidas. La empresa quiere introducirse en el grupo (si existe) que esté menos servido por la competencia. En este ejemplo no existen grupos de clientes predeterminados.

**Asociaciones**: las observaciones son usadas para identificar asociaciones entre variables. La búsqueda de asociaciones es diferente a la búsqueda de relaciones causales. Las relaciones causales son mucho más difíciles de encontrar que las asociaciones, debido a la presencia de variables no observadas. Las relaciones causales y asociaciones no son equivalentes: si hay asociaciones no tiene por qué haber causalidad.

**Patrones secuenciales**: se trata de identificar patrones de comportamiento y tendencias. Un ejemplo sería intensidades de expresión en microarrays que permiten distinguir entre diferentes expresiones de genes para individuos con cáncer o sin él.

**Aplicación del Data Mining**

Algunas áreas de aplicación del DM son

Toma de Decisiones. Ejemplos: banca, finanzas, seguros, marketing, políticas sanitarias o demográficas.

Procesos Industriales.

Investigación Científica Ejemplos: medicina, epidemiología, bioinformática, psicología.

Soporte al Diseño de Bases de Datos.

Mejora de Calidad de Datos.

Mejora en el área de empresas de Consulting

**A continuación se indican algunos ejemplos de aplicación del DM.**

1. Comercio/Marketing

a) Identificación de patrones de compra de los clientes.

b) Búsqueda de asociaciones entre clientes y características demográficas.

c) Predicción de respuesta a campañas de correo.

d) Análisis de cestas de la compra.

2. Banca

a) Detección de patrones de uso fraudulento de tarjetas de crédito

b) Identificación de clientes leales.

c) Predicción de clientes con probabilidad de cambiar su afiliación

d) Determinación del gasto de tarjeta de crédito por grupos.

e) Búsqueda de correlaciones entre indicadores financieros

f ) Identificación de reglas de mercado de valores a partir de históricos.

4. Transportes

a) Determinación de la planificación de la distribución entre tiendas.

b) Análisis de patrones de carga.

**Criterios para la toma de decisión clasificación**

Es posible utilizar diversos criterios para la toma de decisiones. El termino decisiones se puede bien referir a la toma de decisiones de clasificación, así que los criterios presentados a continuación podrán ser utilizados para resolver problemas de clasificación. Entre distintos criterios destacamos los siguientes.

— Criterio de Laplace

— Criterio de Wald

— Criterio de Hurwicz

— Criterio de Savage

— Criterio de Bayes

**Criterio de Laplace**

Este criterio, supone que los distintos estados de la naturaleza, tienen toda la misma probabilidad de ocurrencia unos que otros. Este criterio, es pues un criterio de racionalidad, y se basa en el principio de razón insuficiente: no tengo insuficiente razón de pensar que un estado de la naturaleza tenga mas probabilidades que otros. Dicho en otras palabras: si somos totalmente ignorantes (estamos en una situación de incertidumbre total sobre las posibilidades de los distintos estados de la naturaleza), debemos pensar que todos ellos tienen la misma probabilidad de producirse.

**Criterio de Wald**

sugiere que los responsables de tomar decisiones, son siempre pesimistas o conservadores puesto que siempre deben de esperar lo peor (la naturaleza actuara contra ellos) y, por lo tanto, deben elegir aquella estrategia que maximize el pago mínimo. Esta definición, es la que hace que este criterio, reciba el nombre de maximín. El criterio de Wald, supone pensar que la naturaleza actuará siempre de forma malévola, produciendo siempre el estado de la naturaleza que más nos perjudique. En estas circunstancias, continuamente adversas, se debe seleccionar la estrategia que ofrezca un pago mínimo tan grande como sea posible.

**Criterio de Hurwicz**

El criterio de Hurwicz es un criterio de optimismo apoyado en la idea de que los humanos tenemos golpes de suerte favorables. Como la naturaleza nos suele ser propicia, los encargados de tomar decisiones, deberán seleccionar aquel estado de la naturaleza que ofrezca el máximo pago para la estrategia seleccionada. Este criterio, es el típico de los jugadores puros, que no abandonan la mesa de juego mientras exista esperanza por mínima que ésta sea de obtener ganancias. Es el criterio de ganancia máxima, o perdida o riesgo máximo. Se le conoce así mismo con el nombre de maximax, ya que trata de seleccionar aquella estrategia que maximice el pago máximo. Hurwicz no sugiere que los responsables de la toma de decisiones sean absolutamente optimistas en todos los caso

**Criterio de Savage**

Considera que los decisores podrían lamentarse después de haber tomado una decisión y que el estado de la naturaleza ocurra. Podría llegar a desear haber seleccionado una estrategia completamente diferente. El criterio de Savage, trata de minimizar el arrepentimiento antes de seleccionar realmente una estrategia en particular. Savage sugiere que la magnitud del arrepentimiento se puede medir con la diferencia que existe entre el pago que realmente puede recibirse y el que podría haberse recibido al haber seleccionado la estrategia más adecuada al estado de la naturaleza que se ha producido.

**Criterio de Bayes**

En la Teoría de la Información existe actualmente un concepto, tomado de la electrónica feed-back (retroacción o retroalimentación). Un sencillo ejemplo en electrónica seria, por ejemplo, el de un regulador de un motor eléctrico. Este regulador informa al motor de las revoluciones que debe de llevar el eje en función de la carga soportada. Un simple termostato de un aparato de aire acondicionado, o de una nevera, sería otro sencillo símil de esta retroalimentación.

**Por qué se hace necesario el Data Mining**

Porque existen extensos volúmenes de datos almacenados en Fuentes de Información, los cuales se acumulan bajo la creencia que alguien, en algún momento los utilizará. Sin embargo, crece progresivamente la diferencia entre Generación de datos y Entendimiento de éstos: como el volumen de datos aumenta, el número de personas que entienden estos datos desafortunadamente disminuye. - Porque la información oculta en los datos es útil y generalmente no se encuentra en forma explícita para tomar ventaja de ésta. - Porque en algunos casos, los datos no se pueden analizar por métodos estadísticos estándar, porque pueden existir valores perdidos o bien, los datos pueden estar en forma cualitativa y no cuantitativa. - Porque en ciertas situaciones, el acceso a los datos no es sencillo. - Las técnicas de Data Mining también se hacen necesarias por el desarrollo actual de Almacenes de Datos (Data Warehouse) a gran escala, que son los sistemas utilizados para el almacenamiento y distribución de cantidades masivas de datos.