Introducción a 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

El símbolo → tiene el siguiente sentido: los datos están *bien* recogidos y la estadística *bien* aplicada.

Según algunos autores, el Data Mining es aquella parte de la estadística (principalmente estadística no paramétrica) que se usa para problemas que se presentan actualmente en Análisis de Datos. Los problemas actuales se diferencian de los clásicos en que el número de datos a analizar es mucho mayor y, como consecuencia, las técnicas estadísticas clásicas no pueden ser aplicadas.

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.

Con estas técnicas es posible, a veces, hacer evidente las relaciones ocultas entre sucesos. Un ejemplo simple sería averiguar la relación entre la compra de pañales y de cerveza el sábado por la tarde en los supermercados. Este ejemplo ilustra muy bien la necesidad de

conocer el campo de trabajo para aplicar el Data Mining: sólo un especialista que conozca a su clientela es capaz de interpretar una correlación bruta que permita realizar el retrato típico de una pareja haciendo sus compras. Encontrar las relaciones causales que llevan a correlaciones como la anterior puede ser más rápido y sencillo con el Data Mining.

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:

(*i*) 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*.

(*ii*) 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

(*iii*) **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.

(*iv*) **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 cancer o sin él.

Ejemplos de 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.

3. **Seguros y Salud Privada**

a) Análisis de procedimientos médicos solicitados.

b) Predicción de qué clientes compran nuevas pólizas.

c) Identificación patrones de comportamiento para clientes con riesgo.

d) Identificación de comportamiento fraudulento.

4. **Transportes**

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

b) Análisis de patrones de carga.

5. **Medicina**

a) Identificación de terapias médicas adecuadas para diferentes enfermedades.

b) Asociación de síntomas y clasificación diferencial de patologías.

c) Estudio de factores (genéticos, precedentes, hábitos, alimenticios, etc.) de riesgo en distintas patologías.

d) Segmentación de pacientes para una atención más adecuada según su grupo.

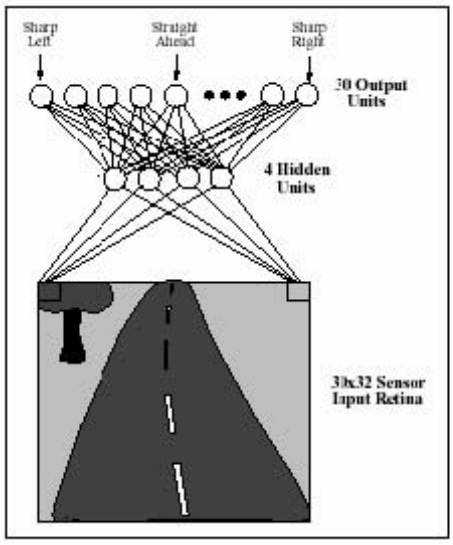
e) Predicciones temporales de los centros asistenciales para el mejor uso de recursos, consultas, salas y habitaciones.

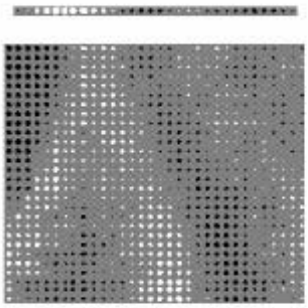
f) Estudios epidemiológicos, análisis de rendimientos de campañas de información, prevención, sustitución de fármacos, etc.

Ejemplos de éxito en la aplicación del Data Mining

Este coche conduce automáticamente en la autovía.

Utilizas una cámara digital y las imágenes se elaboran para reconocer las líneas blancas.





Algunas Técnicas Estadísticas utilizadas en Data Mining

Como en todo procedimiento automático, las predicciones estadísticas de Data Mining deben ser inspeccionadas por personas familiarizadas con la materia de trabajo, de manera que comprendan y verifiquen lo que se ha producido.

Hay que encontrar un término medio entre la capacidad explicativa del modelo (claridad) y su poder de predicción. En general, conforme aumenta el poder de predicción del modelo baja su capacidad de interpretar el fenómeno objeto de estudio. Mientras más sencilla sea la forma del modelo, más fácil será su comprensión, pero tendrá menor capacidad para tener en cuenta dependencias sutiles o demasiado variadas (no lineales). Por ejemplo, los árboles de decisión conducen a modelos de fácil interpretación, pero tienen un bajo poder predictivo porque las decisiones son tomadas al contestar a preguntas de tipo binario Si–No. Al contrario, las redes neuronales tienen una gran poder predictivo (y tienen también la posibilidad de adaptarse a valores bastante indefinidos e incluso ausentes), pero resulta muy difícil asignar una interpretación a su funcionamiento: seria, un poco, como si quisiéramos examinar el cerebro de alguien para saber lo que piensa.

Sin embargo, una buena herramienta de visualización le da la posibilidad al usuario de reconstruir el *razonamiento* de la red neuronal. Según cuál sea el precio a pagar, y una vez que se haya establecido la confianza en la herramienta establecida, el usuario notará, la mayoría de las veces, que la pérdida parcial de comprensión será más que compensada por la calidad de las predicciones.

Ninguno de los modelos estadísticos presentados es nuevo. Los árboles de decisión y de regresión han sido utilizados en ciencias sociales en los años 60; las bases de reglas fueron popularizadas durante el auge de los *sistemas expertos* en los 80 y la evaluación por puntuación apreciada por los banqueros durante largos decenios. Incluso las redes neuronales aparecieron ya en los años 40, pero ha sido preciso el desarrollo del poder de cálculo de estos últimos años para que, por fin, fueran utilizables de manera sencilla.

La mayoría de estos previsores se fabrican, no por cálculo directo partiendo de los datos como antes, sino a través de métodos tomados del campo de la *inteligencia artificial*. Las dos técnicas principales son el aprendizaje (a partir de un modelo *cualquiera* que se ajusta progresivamente a la realidad) y la evolución (o *vida artificial*, un conglomerado de varios miles de modelos *cualquiera* son susceptibles de *evolucionar* de manera competitiva o *darwiniana*). Además, todas las herramientas permiten que se determine la importancia de cada variable para la decisión (*distintividad* o carácter pertinente). Esto resulta de extrema utilidad para proceder a la selección de variables. Al haber determinado con precisión las variables más pertinentes, se podrá optar por retomar el problema con técnicas más convencionales si ciertas restricciones de explotación lo imponen.

Las técnicas hasta ahora descritas sólo tratan datos numéricos o cualitativos. El *mining* surge ante el problema cada vez más apremiante de extraer información automáticamente a partir de masas de textos. La enorme cantidad de referencias recogidas durante una búsqueda en Internet ilustra muy bien este problema. La investigación literal simple se ha mostrado limitada desde hace ya mucho tiempo; hay muchos problemas como los errores al teclear, la sinonimia, las acepciones múltiples, etc. En definitiva, es necesario inyectarle

al ordenador un cierto sentido común o *conocimiento del mundo*. Aún en ese caso, la memoria y el poder de cálculo disponibles en nuestra época permiten ciertas soluciones que no siempre son las más elegantes pero sí potentes y rápidas.

Las técnicas de DM suelen dividirse en dos grupos según los objetivos de los análisis:
– técnicas supervisadas: donde hay una variable que debe ser explicada por las otras;
– técnicas no supervisadas: donde no hay una variable preferente que debe ser explicada por las otras.

A continuación se indican algunas de las técnicas estadísticas más utilizadas en DM. Algunas de estas técnicas serán estudiadas a lo largo del curso.

Redes neuronales

Se trata de una herramienta de análisis estadístico que permite la construcción de un modelo de comportamiento a partir de una determinada cantidad de ejemplos (constituidos por una determinada cantidad de *variables descriptivas* de dicho comportamiento. La red neuronal, completamente *ignorante* al principio, efectúa un *aprendizaje* partiendo de los ejemplos, para luego transformarse, a través de modificaciones sucesivas, en un modelo susceptible de rendir cuenta del comportamiento observado en función de las variables descriptivas.

Por ejemplo, al impartir a una red neuronal un aprendizaje relacionado con descripciones de personas que piden préstamos (estado civil, profesión, etc.), junto a su comportamiento adoptado frente al reintegro del dinero, nos encontramos con capacidad de construir un modelo del riesgo asociado con la descripción de los clientes. Si luego le pedimos a ese modelo predicciones sobre nuevos expedientes, podemos constatar que la red neuronal predice con buena precisión si el cliente pagará bien o no. En todo caso, la red neuronal, una vez construida, constituye une verdadero modelo *a la medida* que actúa en función de lo que percibe. Tampoco se trata de ir a buscar dentro de una biblioteca un modelo más o menos adaptado. Si en realidad existe una relación de causa–efecto en medio de las descripciones introducidas (perfil del prestatario, cotizaciones anteriores de

una acción, relaciones de medidas, punto de funcionamiento deseado) y los valores a prever (riesgo de *ruptura* del préstamo, curso de la acción 10 días más tarde, naturaleza de la avería, variables de mando), la red intentará descubrirla.

La red neuronal es robusta respecto a valores aberrantes en la muestra de aprendizaje. No se queda invalidada con algunos ejemplos enredados o falsos: estos serán descartados del resto por su incoherencia. Los valores ausentes son también hábilmente manejados y no perturban la construcción del modelo. En un ámbito completamente diferente, se puede aprender a asociar en una máquina–herramienta relaciones de medidas y sus averías: el previsor despejado realiza una manutención preventiva indicando la posibilidad de avería desde el momento en que las medidas tomarán valores que él estimará como sospechosas (o realizando un diagnóstico a partir de las últimas relaciones si es demasiado tarde). Así mismo, en función automática y en función mando es posible modelizar el comportamiento de un reactor químico o de un robot. La red neuronal indica según el modo de funcionamiento que se desee cuáles son los valores necesarios de las variables de mando.

Esta capacidad para aprender todo aquello que tenga un sentido ha sido establecida de manera rigurosa (Teorema de Kolmogorov). Este método no se ha popularizado sino en la época actual por la simple razón de que ahora es cuando se ha llegado a un cierto poder de cálculo necesario para su puesta en aplicación.

Árboles de decisión

Son modelos que tienen estructuras de forma de árbol que representan conjuntos de decisiones. Estas decisiones generan reglas para la clasificación de un conjunto de datos.

Métodos específicos de árboles de decisión incluyen Árboles de Clasificación y Regresión (*CART*: Classification And Regression Tree). Básicamente, los árboles de decisión, son representaciones gráficas de la lógica de las probabilidades aplicada a las alternativas de decisión. El tronco del árbol es el punto de partida de la decisión. Las ramas de éste comienzan con la probabilidad del primer acontecimiento. La probabilidad de cada acontecimiento produce dos o más efectos posibles, algunos de los cuales conducen a otros

acontecimientos de probabilidad y a puntos de decisión subconsecuentes. Los valores en los que se cifran las ramas del árbol, provienen de un análisis muy cuidadoso que se basa en el establecimiento de un criterio para la toma de decisión.

Algoritmos Genéticos

Este conjunto de técnicas no serán considerada en el seguimiento del curso, pero se trata de técnicas típicas de la literatura de DM. El algoritmo genético permite obtener soluciones a un problema que no tiene ningún método de resolución descrito de forma precisa, o cuya solución exacta, si es conocida, es demasiado complicada para ser calculada en un tiempo aceptable. Es el caso particular de cuando se encuentran restricciones múltiples y complejas, e incluso contradictorias, que deben ser satisfechas simultáneamente como, por ejemplo, para formar equipos de trabajo, planificar rondas de entregas, implantar puntos de venta de manera óptima, construir modelos estadísticos.

Según el algoritmo genético, numerosas soluciones más o menos correctas inherentes a dicho problema son creadas al azar, según una forma ya definida: itinerario, horarios, base de reglas de decisión, evaluación por puntuación, red neuronal, etc. Cada solución será representada a través de una cadena de 0 y de 1 en *cromosomas* que se verán entonces sometidos a una imitación de la evolución de las especies: mutaciones y reproducción por hibridación. Al favorecer la supervivencia de los más *aptos* (las soluciones más correctas), se provoca la aparición de híbridos cada vez mejores que sus padres. La población inicial da paso de esta manera a generaciones sucesivas mutadas y procreadas por hibridación a partir de sus *padres*. Al despejar los elementos más aptos *presión de la evolución* se garantiza que las generaciones sucesivas serán cada vez más adaptadas a la resolución del problema. Este mecanismo sorprendente de clasificación ha sido validado matemáticamente con el rigor que le corresponde.

El mecanismo de evolución y de selección es independiente del problema por resolver: sólo varían la función que descodifica el genotipo en una solución posible (cualquier tipo de descodificación tiene la posibilidad de ser utilizado de la manera más sencilla posible)

y la función que evalúa la justeza de la solución (en el caso de los previsores probándolos en unas cuantas centenas de casos). Esta técnica es de aplicación general. El algoritmo genético puede aplicarse a la producción de una variedad de objetos mientras sea posible obtener una calificación que represente la justeza de la solución. En particular, es posible fabricar previsores estadísticos no a través de cálculos de datos como en la estadística clásica sino haciendo evolucionar los datos por algoritmo genético (*inducción*. Por problemas de clasificación o de segmentación, la justeza significa simple y llanamente la tasa de reordenación del previsor con respecto a un conjunto dado de ejemplos. El mecanismo de estimulación de lo más apto permite entonces la aparición del previsor que reordenará los datos lo mejor posible. Este tipo de construcción de previsor forma parte de las técnicas de algoritmo genético utilizadas en DM.

La técnica del algoritmo genético da enfoque un poco brutal que necesita un gran poder de cálculo pero que posee la inmensa ventaja de proporcionar soluciones no muy lejos de lo óptimo incluso sin conocer métodos de soluciones. El algoritmo genético no exige ningún conocimiento acerca de la manera más idónea de resolver el problema; sólo es necesario la capacidad de evaluar la calidad de una solución. También es muy ligero para ponerlo en práctica (el motor es común, no hay mucha programación específica que hacer). En la resolución de un mismo problema el enfoque algorítmico es específico, muy rápido, mientras el algoritmo genético se caracteriza por ser general pero muy lento.

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 todos 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.

Este criterio, propone seleccionar aquella estrategia cuyo pago esperado sea máximo. Para calcular el pago esperado, se suman los pagos de cada estrategia, y se divide esta suma por él numero de estados de la naturaleza. Este criterio, conduce a una asignación de probabilidades a los distintos estados de la naturaleza. Si estamos en incertidumbre total o ignorancia completa, debemos de asignar probabilidades iguales y elegir aquella cuyo pago esperado sea máximo.

Criterio de Wald

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 casos. Esto equivaldría a vivir en un estado utópico y no en un mundo real. Para vencer este optimismo total, Hurwicz introdujo el concepto de coeficiente de optimismo. Este coeficiente de optimismo, implica que los decisores deben considerar tanto el pago más alto, como el más bajo y deben considerar la importancia de ambos, atendiendo a ciertos factores de probabilidad. Las probabilidades asignadas a los pagos más altos y más bajos deben sumar 1 en total, basándose en la posición del responsable de la toma de decisiones respecto a las condiciones optimistas.

Criterio de Savage

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.

Los costes condicionales de oportunidad se deben a la falta de información perfecta. Una vez obtenida la matriz de costes condicionales de oportunidad, Savage propone un criterio similar al de Wald. Se muestra también pesimista. Considera que la naturaleza va a obrar en contra nuestra y, en consecuencia, debe minimizar el coste condicional de oportunidad máximo. De este modo se calcula el coste condicional de oportunidad máximo de cada estrategia y se elige el mínimo de estos máximos. Por esta razón, se conoce al criterio de Savage, como el criterio del *minimax*.

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.

En nuestro caso, estas nociones pueden aprovecharse con éxito, puesto que planteada una situación a priori, los resultados pueden llevar a conclusiones diferentes de las que sé pensó. En este sentido el problema se concibe, por así decirlo, en forma pasiva y se modifica la estimación inicial en aras de los resultados que se van obteniendo.

El teorema de Bayes, viene en auxilio de esta situación planteada, sugiriendo una fórmula de arreglo que, independientemente de su rigor matemático, viene avalada por el sentido común. Bayes postula la obtención de una probabilidad combinada a base de ejecutar una ponderación entre ambas fuentes: credencial a priori y datos observados.

Sean, por ejemplo, π_1, π_2 y π_3 las probabilidades de que sucedan {{1},{2},{3}}, bajo el punto de vista subjetivo, y supongamos que sean f_1, f_2, y f_3 las verosimilitudes de los estados de naturaleza. La probabilidad combinada resultante que reúne la información

aportada (probabilidad a posteriori), seria:

$$\begin{aligned} p_1 &= \frac{\pi_1 f_1}{\pi_1 f_1 + \pi_2 f_2 + \pi_3 f_3}, \\ p_2 &= \frac{\pi_2 f_2}{\pi_1 f_1 + \pi_2 f_2 + \pi_3 f_3}, \\ p_3 &= \frac{\pi_3 f_3}{\pi_1 f_1 + \pi_2 f_2 + \pi_3 f_3} \end{aligned}$$

Se aprecia cómo las formulas posteriores, no son mas que una ponderación combinada de ambas estructuras probabilísticas. Los p_i (para $i = 1, 2, 3$) son expresiones coherentes con el sentido que tiene la probabilidad cuando ésta abarca todos los casos posibles que pueden darse. Así, la suma de sus probabilidades será la unidad.

Se entiende que sólo en el limite, cuando el numero de experiencias fuera muy grande, la influencia de la probabilidad objetiva en las fórmulas de probabilidad finales, anularía la de la probabilidad subjetiva original la cual depende del sujeto y que podemos considerar como probabilidad de partida. Así pues, dos decisores u observadores, partiendo de posiciones diferentes, pero recibiendo ambos el mismo conjunto de informaciones, corregirán progresivamente sus estimaciones iniciales, descubriendo finalmente que sus sistemas de probabilidad se van aproximando. Pero mientras que no se disponga de una información infinita, permanecerá al menos vestigios de su subjetividad inicial.

En esta misma idea inicial de la probabilidad subjetiva, está implícito el conocimiento por parte del decisor, de tal forma que le sitúa con cierta operatividad frente al problema. Sin embargo, pudiera suceder que, por ignorancia o por un criterio conservador a ultranza, no se decida a asignar diferentes pesos específicos, es decir probabilidades a los diversos casos posibles que se presentan. Para paliar esta situación, habrá que utilizar un sistema de reglas formales para asignar probabilidades a priori.

Introduction to Data Mining

Slides based on documentation from *SAS* training courses and *Data Mining: Practical Machine Learning Tools and Techniques* (2^{th} Ed.) by I.H. Witten and E. Frank (2005).

- We are overwhelmed with data. The amount of data in the world, in our lives, seems to go on and on increasing. As the volume of data increases, inexorably, the proportion of it that people understand decreases, alarmingly.

- Lying hidden in all this data is **information**, potentially useful information, that is rarely made explicit or taken advantage of.

- People have been seeking patterns in data since human life began: Hunters seek patterns in animal migration behavior, farmers seek patterns in crop growth, politicians seek patterns in voter opinion, and lovers seek patterns in their partners' responses...

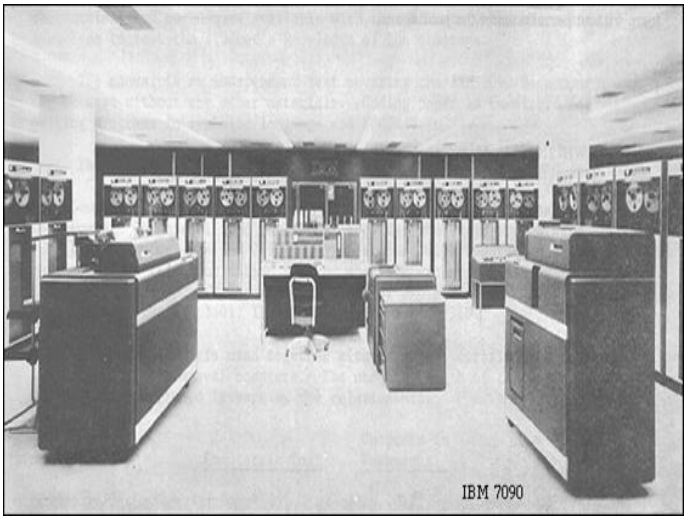
- A scientist's job is to make sense of data, to discover the patterns that govern how the physical world works and encapsulate them in theories that can be used for predicting what will happen in new situations.

- **A tentative definition of Data Mining**: *Advanced methods for exploring and modeling relationships in large amounts of data*.

- There are other similar definitions. However, exploring and modeling relationships in data has a much longer history than the term data mining.

- Data mining analysis was limited by the computing power of the time. The IBM 7090 is a transistorized mainframe introduced in 1959. It had a processor speed of approximately 0.5 MHz and roughly 0.2 MB of RAM using ferrite magnetic cores.

Data sets were stored on cards and then transferred to magnetic tape using separate equipment. A data set with 600 rows and 4 columns would have used approximately 3000 cards. Tape storage was limited by the size of the room. The room pictured below contains the tape drives and controllers for the IBM 7090. The computer itself would need a larger room.



- In data mining, the data is stored electronically and the search is automated by computer.
- Computer performance has been doubling every 18 to 24 months. This has led to technological advances in storage structures and a corresponding increase in MB of storage space per dollar.
- Parkinson's law of data: *Data expands to fill the space available for storage*.
- The amount of data in the world has been doubling every 18 to 24 months. Multi-gigabyte commercial databases are now commonplace.
- Economists, statisticians, forecasters, and communication engineers have long worked with the idea that patterns in data can be sought automatically, identified, validated, and used for prediction.

- The unbridled growth of databases in recent years, databases on such everyday activities as customer choices, brings data mining to the forefront of new business technologies.

- It has been estimated that the amount of data stored in the world's databases doubles every 20 months.

- Data mining is about solving problems by analyzing data already present in databases.

- The data deluge is the result of the prevalence of automatic data collection, electronic instrumentation, and online transactional processing (*OLTP*).

- There is a growing recognition of the untapped value in these databases. This recognition is driving the development of data mining and data warehousing.

- Historically, most data was generated or collected for research purposes. But, today, businesses have massive amounts of operational data and were not generated with data analysis in mind. It is aptly characterized as *opportunistic*. This is in contrast to experimental data where factors are controlled and varied in order to answer specific questions.

- The owners of the data and sponsors of the analyses are typically not researchers. The objectives are usually to support crucial business decisions.

- Database marketing makes use of customer and transaction databases to improve product introduction, cross-sell, trade-up, and customer loyalty promotions.

- One of the facets of customer relationship management is concerned with identifying and profiling customers who are likely to switch brands or cancel services (*churn*). These customers can then be targeted for loyalty promotions.

- **Example:** *Credit scoring* is chiefly concerned with whether to extend credit to an applicant. The aim is to anticipate and reduce defaults and serious delinquencies. Other credit risk management concerns are the maintenance of existing credit lines (should the credit limit be raised?) and determining the best action to be taken on delinquent accounts.

- The aim of fraud detection is to uncover the patterns that characterize deliberate deception. These patterns are used by banks to prevent fraudulent credit card transactions and bad checks, by telecommunication companies to prevent fraudulent calling card transactions, and by insurance companies to identify fictitious or abusive claims.

- Healthcare informatics is concerned with decision-support systems that relate clinical information to patient outcomes. Practitioners and healthcare administrators use the information to improve the quality and cost effectiveness of different therapies and practices.

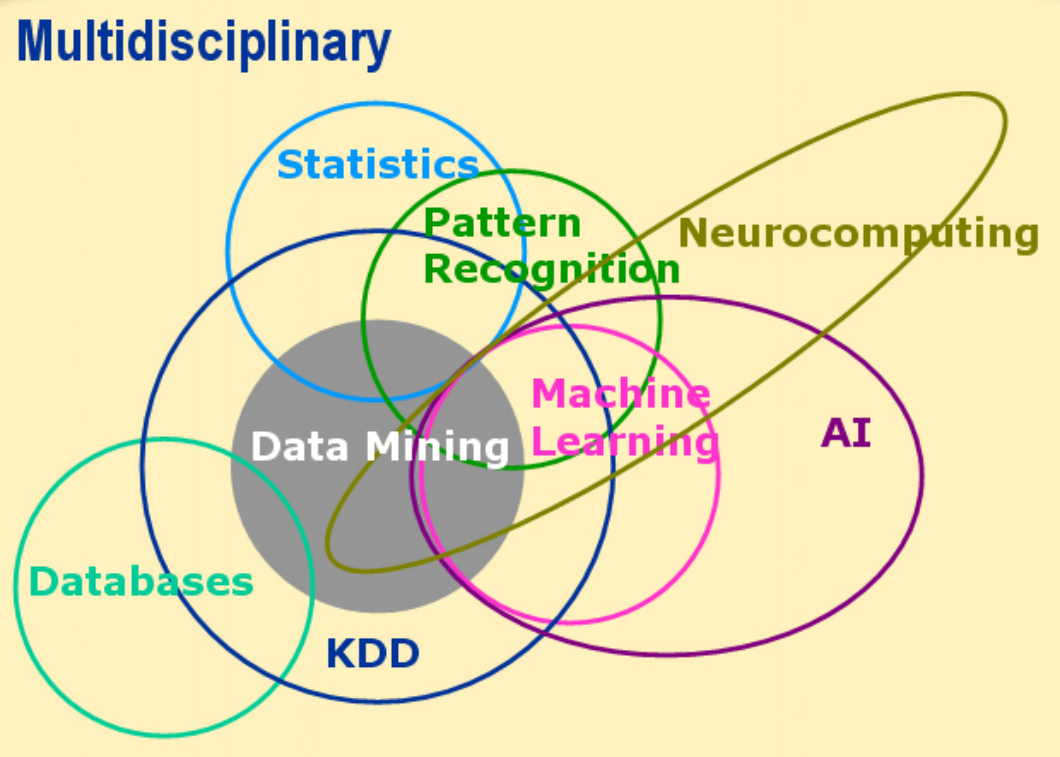
- The analytical tools used in data mining were developed mainly by statisticians, artificial intelligence (AI) researchers, and database system researchers.

- KDD (knowledge discovery in databases) is a multidisciplinary research area concerned with the extraction of patterns from large databases. KDD is often used synonymously with data mining. More precisely, data mining is considered a single step in the overall discovery process.

- Machine learning is a branch of AI concerned with creating and understanding semiautomatic learning methods.

- Pattern recognition has its roots in engineering and is typically concerned with image classification.

- Neurocomputing is, itself, a multidisciplinary field concerned with neural networks.



- One consequence of the multidisciplinary lineage of data mining methods is confusing terminology. The same terms are often used in different senses, and synonyms abound.

- A related pitfall is to specify the objectives in terms of analytical methods:

  - Implement neural networks.

  - Apply visualization tools.

  - Cluster the database.

- The same analytical tools may be applied (or misapplied) to many different problems. The choice of the most appropriate analytical tool often depends on subtle differences in the objectives. The objectives eventually must be translated in terms of analytical methods. This should occur only after they are specified in ordinary language.

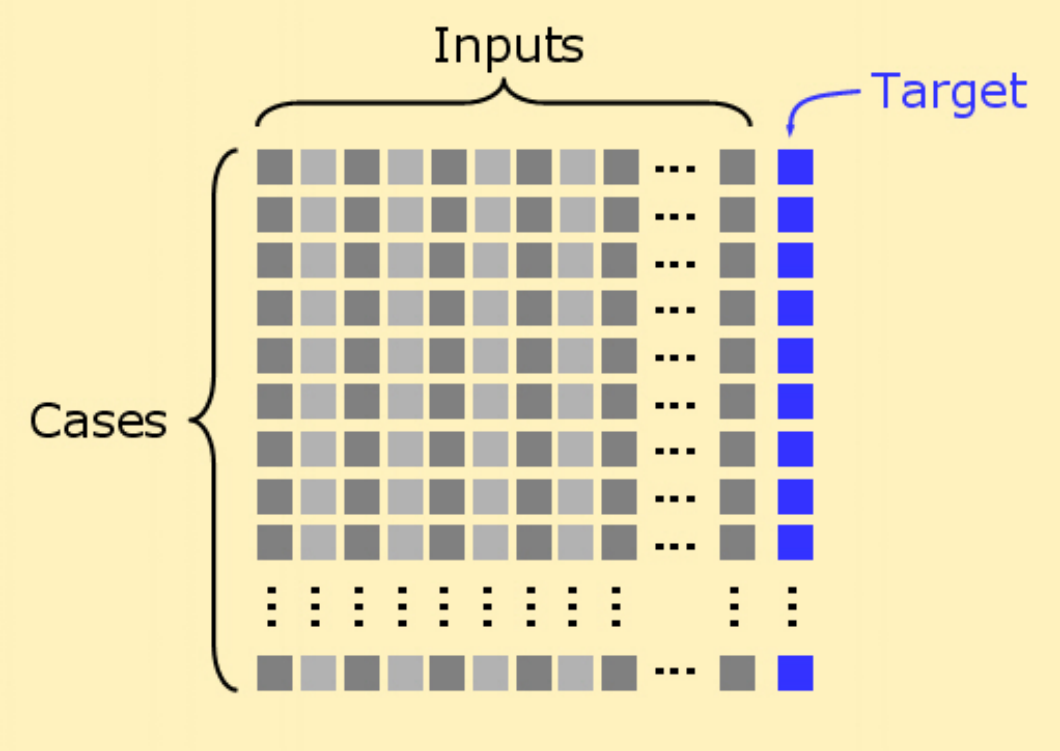
- Many people think data mining means magically discovering hidden nuggets of information without having to formulate the problem and without regard to the structure or content of the data. This is an unfortunate misconception.

- The database community has a tendency to view data mining methods as more complicated types of database queries.

- For example, standard query tools can answer questions such as *how many surgeries resulted in hospital stays longer than 10 days?* But data mining is needed for more complicated queries such as *what are the important preoperative predictors of excessive length of stay?*

- This view has led many to confuse data mining with query tools.

  The problem translation step involves determining what analytical methods are relevant to the objectives.



- Predictive modeling (supervised prediction, supervised learning) is the fundamental data mining task. The training data set consists of cases (observations, examples, instances, records).

- Associated with each case is a vector of input variables (predictors, features, explanatory variables, independent variables) and a target variable (response, outcome, dependent variable). The training data is used to construct a model (rule) that can predict the values of the target from the inputs.

- The task is referred to as supervised because the prediction model is constructed from data where the target is known. It allows you to predict new cases when the target is unknown. Typically, the target is unknown because it refers to a future event. In addition, the target may be difficult, expensive, or destructive to measure.

- The measurement scale of the inputs can be varied. The inputs may be numeric variables such as income. They may be nominal variables such as occupation. They are often binary variables such as home ownership.

- The main differences among the analytical methods for predictive modeling depend on the type of target variable.

- In supervised classification, the target is a class label (categorical). The training data consists of labeled cases. The aim is to construct a model (classifier) that can allocate cases to the classes using only the values of the inputs.

- Regression analysis is supervised prediction where the target is a continuous variable (it can also be used more generally; for example, logistic regression). The aim is to construct a model that can predict the values of the target from the inputs.

- In survival analysis, the target is the time until some event occurs. The outcome for some cases is censored; all that is known is that the event has not yet occurred.

EXAMPLE: The weather problem

Consider a fictitious weather problem with 14 examples in the training set and four attributes: *outlook*, *temperature*, *humidity*, and *windy*. The outcome is whether to play or not.

This creates 36 possible combinations ($3 \times 3 \times 2 \times 2 = 36$).

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

A set of rules learned from this information might look as follows:

```
If outlook=sunny and humidity=high then play=no

If outlook=rainy and windy=true then play=no
```

```
If outlook=overcast then play=yes
If humidity=normal then play=yes
If none of the above then play=yes
```

- These rules are interpreted in **order**: the first one, then if it doesn't apply the second, and so on.
- A set of rules that are intended to be interpreted in sequence is called a *decision list*.
- Interpreted as a decision list, the rules correctly classify all of the examples in the table, whereas taken individually, out of context, some of the rules are incorrect.
- The previous rules are *classification rules*: they predict the classification of the example in terms of whether to play or not.
- It is also possible to just look for any rules that strongly associate different attribute values. These are called *association rules*.
- Many association rules can be derived from the weather data. Some good ones are as follows:

```
If temperature=cool then humidity=normal
If humidity=normal and windy=false then play=yes
```

```
If outlook=sunny and play=no then humidity=high

If windy=false and play=no then outlook=sunny and humidity=high
```

- All these rules are 100% correct on the given data; they make no false predictions.

- There are many more rules that are less than 100% correct because unlike classification rules, association rules can *predict* any of the attributes, not just a specified class, and can even predict more than one thing.

- For example, the fourth rule predicts both that outlook will be sunny and that humidity will be high.

- The search space, although finite, is extremely big, and it is generally quite impractical to enumerate all possible descriptions and then see which ones fit.

- In the weather problem there are $4 \times 4 \times 3 \times 3 \times 2 = 288$ possibilities for each rule.

- If we restrict the rule set to contain no more than 14 rules (because there are 14 examples in the training set), there are around $288^{14} = 2.7 \times 10^{34}$ possible different rule sets!!!

- Another way of looking at generalization as search is to imagine it as a kind of *hill-climbing* in description space to find the description that best matches the set of examples according to some pre-specified matching criterion.

- This is the way that most practical machine learning methods work. However, except in the most trivial cases, it is impractical to search the whole space exhaustively; most practical algorithms involve heuristic search and cannot guarantee to find the optimal description.

EXAMPLE: Iris, a classic numeric dataset

The iris dataset, which dates back to 1935 and is arguably the most famous dataset used in data mining, contains 50 examples each of three types of plant: *Iris setosa*, *Iris versicolor*, and *Iris virginica*. There are four attributes: sepal length, sepal width, petal length, and petal width (all measured in centimeters).

	Sepal length (cm)	Sepal width (cm)	Petal length (cm)	Petal width (cm)	Type
1	5.1	3.5	1.4	0.2	*Iris setosa*
2	4.9	3.0	1.4	0.2	*Iris setosa*
3	4.7	3.2	1.3	0.2	*Iris setosa*
4	4.6	3.1	1.5	0.2	*Iris setosa*
5	5.0	3.6	1.4	0.2	*Iris setosa*
...					
51	7.0	3.2	4.7	1.4	*Iris versicolor*
52	6.4	3.2	4.5	1.5	*Iris versicolor*
53	6.9	3.1	4.9	1.5	*Iris versicolor*
54	5.5	2.3	4.0	1.3	*Iris versicolor*
55	6.5	2.8	4.6	1.5	*Iris versicolor*
...					
101	6.3	3.3	6.0	2.5	*Iris virginica*
102	5.8	2.7	5.1	1.9	*Iris virginica*
103	7.1	3.0	5.9	2.1	*Iris virginica*
104	6.3	2.9	5.6	1.8	*Iris virginica*
105	6.5	3.0	5.8	2.2	*Iris virginica*
...					

All attributes have values that are numeric. The following set of rules might be learned from this dataset:

```
If petal length < 2.45 then Iris setosa

If sepal width < 2.10 then Iris versicolor

If sepal width < 2.45 and petal length < 4.55 then Iris
versicolor

If sepal width < 2.95 and petal width < 1.35 then Iris
versicolor

If petal length >= 2.45 and petal length < 4.45 then
Iris versicolor

If sepal length >= 5.85 and petal length < 4.75 then
Iris versicolor

If sepal width < 2.55 and petal length < 4.95 and
petal width < 1.55 then Iris versicolor

If petal length >= 2.45 and petal length < 4.95 and
petal width < 1.55 then Iris versicolor

If sepal length >= 6.55 and petal length < 5.05 then
Iris versicolor
```

```
If sepal width < 2.75 and petal width < 1.65 and
sepal length < 6.05 then Iris versicolor
If sepal length >= 5.85 and sepal length < 5.95 and
petal length < 4.85 then Iris versicolor
If petal length >= 5.15 then Iris virginica
If petal width >= 1.85 then Iris virginica
If petal width >= 1.75 and sepal width < 3.05 then Iris
virginica
If petal length >= 4.95 and petal width < 1.55 then Iris
virginica
```

Concepts, Instances, and Attributes

- We begin by looking at the different forms the input might take.
- The input takes the form of *concepts*, *instances*, and *attributes*. We call which is to be learned as a *concept description*.
- The information that the learner is given takes the form of a set of instances. In the previous examples, each instance was an individual, independent example of the concept to be learned.
- Each instance is characterized by the values of attributes that measure different aspects of the instance.

- There are many different types of attributes, although typical data mining methods deal only with numeric and nominal (categorical) ones.

What is a concept?

- In *classification learning*, the learning scheme is presented with a set of classified examples from which it is expected to learn a way of classifying *unseen* examples.
- In *association learning*, any association among features is sought, not just ones that predict a particular class value.
- In *clustering*, groups of examples that belong together are sought.
- In *numeric prediction*, the outcome to be predicted is not a discrete class but a numeric quantity.
- Regardless of the type of learning involved, we call the thing to be learned the **concept** and the output produced by a learning scheme the **concept description**.
- The weather example is a classification problem. It presents a set of days together with a decision for each as to whether to play the game or not.

- Classification learning is sometimes called *supervised* because the method operates under supervision by being provided with the actual outcome for each of the training examples.
- The outcome is called the *class* of the example.
- The success of classification learning can be judged by trying out the concept description that is learned, on an independent set of test data for which the true classifications are known but not made available to the machine.
- The success rate on test data gives an objective measure of how well the concept has been learned.
- In *association learning* there is no specified class. The problem is to discover any structure in the data that is *interesting*.
- Association rules differ from classification rules in two ways: they can *predict* **any attribute**, not just the class, and they can predict **more than one** attribute's value at a time.
- Association rules usually involve only nonnumeric attributes: thus you would not normally look for association rules in the iris dataset.
- When there is no specified class, **clustering** is used to group items that seem to fall naturally together.

- Imagine a version of the iris data in which the type of iris is omitted. Then it is likely that the 150 instances fall into natural clusters corresponding to the three iris types.

- The challenge is to find these clusters and assign the instances to them and to be able to assign new instances to the clusters as well.

- It may be that one or more of the iris types splits naturally into subtypes, in which case the data will exhibit more than three natural clusters.

- Clustering may be followed by a second step of classification learning in which rules are learned that give an intelligible description of how new instances should be placed into the clusters.

- Numeric prediction is a variant of classification learning in which the outcome is a numeric value rather than a category.

- With numeric prediction problems, the predicted value for new instances is often of less interest than the structure of the description that is learned, expressed in terms of what the important attributes are and how they relate to the numeric outcome.