# Capítulo 2

# El proceso de extracción de conocimiento

En el capítulo anterior hemos visto que la minería de datos no es más que un paso esencial de un proceso más amplio cuyo objetivo es el descubrimiento de conocimiento en bases de datos (del inglés *Knowledge Discovery from Databases*, KDD). Este proceso consta de una secuencia iterativa de etapas o fases. En este capítulo se presentan las fases del proceso de extracción de conocimiento: Preparación de Datos, Minería de Datos, Evaluación, Difusión y Uso de Modelos. Se dan las nociones más básicas de cada una de ellas, se presenta una tipología de tareas de minería de datos (clasificación, estimación/regresión, agrupamiento, reglas de asociación...) y de técnicas para resolverlos (funciones lineales y no lineales, árboles de decisión, redes neuronales artificiales, aprendizaje basado en instancias o casos, métodos basados en núcleos, etc.). Se introducen las medidas básicas de evaluación (precisión, soporte, confianza, error cuadrático medio, distancias...) y el concepto de evaluación mediante los conjuntos de entrenamiento y de prueba.

## 2.1 Las fases del proceso de extracción de conocimiento

En la Figura 2.1 se muestra que el KDD es un proceso iterativo e interactivo. Es iterativo ya que la salida de alguna de las fases puede hacer volver a pasos anteriores y porque a menudo son necesarias varias iteraciones para extraer conocimiento de alta calidad. Es interactivo porque el usuario, o más generalmente un experto en el dominio del problema, debe ayudar en la preparación de los datos, validación del conocimiento extraído, etc.

El proceso de KDD se organiza entorno a cinco fases como se ilustra en la Figura 2.1. En la **fase de integración y recopilación de datos** se determinan las fuentes de información que pueden ser útiles y dónde conseguirlas. A continuación, se transforman todos los datos a un formato común, frecuentemente mediante un almacén de datos que consiga unificar

de manera operativa toda la información recogida, detectando y resolviendo las inconsistencias. Este almacén de datos facilita enormemente la "navegación" y visualización previa de sus datos, para discernir qué aspectos puede interesar que sean estudiados. Dado que los datos provienen de diferentes fuentes, pueden contener valores erróneos o faltantes. Estas situaciones se tratan en la **fase de selección, limpieza y transformación**, en la que se eliminan o corrigen los datos incorrectos y se decide la estrategia a seguir con los datos incompletos. Además, se proyectan los datos para considerar únicamente aquellas variables o atributos que van a ser relevantes, con el objetivo de hacer más fácil la tarea propia de minería y para que los resultados de la misma sean más útiles. La selección incluye tanto una criba o fusión horizontal (filas / registros) como vertical (columnas / atributos). Las dos primeras fases se suelen englobar bajo el nombre de "preparación de datos". En la **fase de minería de datos**, se decide cuál es la tarea a realizar (clasificar, agrupar, etc.) y se elige el método que se va a utilizar. En la **fase de evaluación e interpretación** se evalúan los patrones y se analizan por los expertos, y si es necesario se vuelve a las fases anteriores para una nueva iteración. Esto incluye resolver posibles conflictos con el conocimiento que se disponía anteriormente. Finalmente, en la **fase de difusión** se hace uso del nuevo conocimiento y se hace partícipe de él a todos los posbiles usuarios. Para cada una de estas fases se emplean distintas técnicas de las diferentes disciplinas relacionadas que vimos en la Sección 1.6.

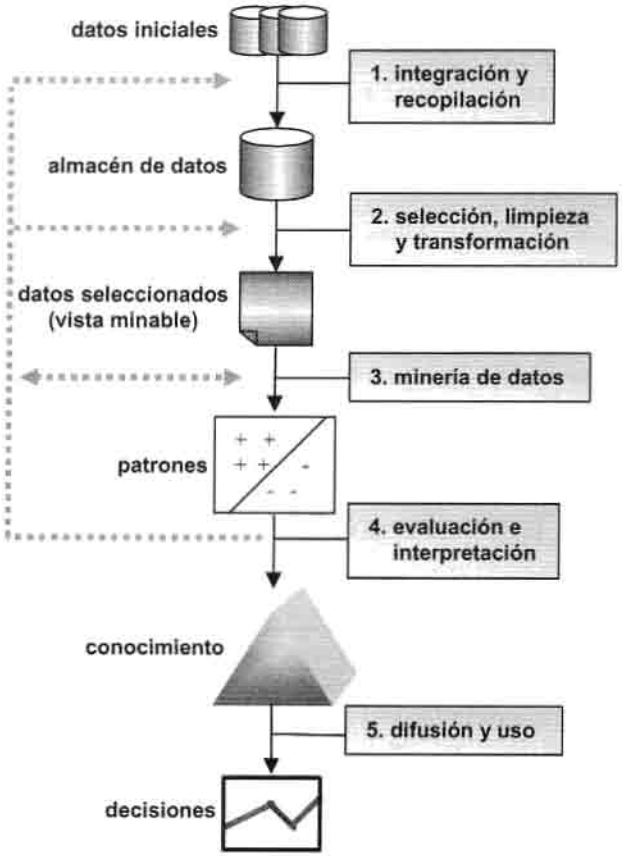


*Figura 2.1. Fases del proceso de descubrimiento de conocimiento en bases de datos, KDD.*

Además de las fases descritas, frecuentemente se incluye una fase previa de análisis de las necesidades de la organización y definición del problema [Two Crows Corporation 1999], en la que se establecen los objetivos de minería de datos. Por ejemplo, un objetivo de negocio de una entidad bancaria sería encontrar patrones en los datos que le ayuden a conservar los buenos clientes; para ello, podríamos tener varios objetivos de minería de datos: construir un modelo para predecir clientes rentables y un segundo modelo para identificar los clientes que probablemente dejarán de serlo.

A continuación pasamos a ver brevemente las fases de la Figura 2.1. A lo largo del libro iremos ampliando y profundizando sobre ellas.

## 2.2 Fase de integración y recopilación

Tal y como mencionamos en el capítulo anterior (Sección 1.1), las bases de datos y las aplicaciones basadas en el procesamiento tradicional de datos, que se conoce como **procesamiento transaccional en línea** (**OLTP**, *On-Line Transaction Processing*) son suficientes para cubrir las necesidades diarias de una organización (tales como la facturación, control de inventario, nóminas...). Sin embargo, resultan insuficientes para otras funciones más complejas como el análisis, la planificación y la predicción, es decir, para tomar decisiones estratégicas a largo plazo. En estos casos, y dependiendo de la aplicación, lo normal es que los datos necesarios para poder llevar a cabo un proceso de KDD pertenezcan a diferentes organizaciones, a distintos departamentos de una misma entidad. Incluso puede ocurrir que algunos datos necesarios para el análisis nunca hayan sido recolectados en el ámbito de la organización por no ser necesarios para sus aplicaciones. En muchos casos tendremos que adquirir además datos externos desde bases de datos públicas (como el censo, datos demográficos o climatológicos) o desde bases de datos privadas (como los datos de compañías de pagos, bancarias, eléctricas, etc., siempre que sea a un nivel agregado para no infringir la legalidad). Esto representa un reto, ya que cada fuente de datos usa diferentes formatos de registro, diferentes grados de agregación de los datos, diferentes claves primarias, diferentes tipos de error, etc. Lo primero, por lo tanto, es integrar todos estos datos. La idea de la integración de múltiples bases de datos ha dado lugar a la tecnología de **almacenes de datos** (*data warehousing*). Este término, tan popular actualmente, hace referencia a la tendencia actual en las empresas e instituciones de coleccionar datos de las bases de datos transaccionales y otras fuentes diversas para hacerlos accesibles para el análisis y la toma de decisiones.


*Figura 2.2. Integración en un almacén de datos.*

Un almacén de datos es un repositorio de información coleccionada desde varias fuentes, almacenada bajo un esquema unificado que normalmente reside en un único emplazamien-

to. Existen varias formas de mezclar las distintas bases de datos para crear el repositorio. Una posibilidad es simplemente hacer una copia de las bases de datos integrantes (probablemente eliminando inconsistencias y redundancias). Obviamente, esta aproximación limita las ventajas para acceder a bases de datos heterogéneas. Por ello, generalmente los almacenes de datos se construyen vía un proceso de integración y almacenamiento en un nuevo esquema integrado. En la Figura 2.2 se muestra este proceso de integración de un almacén de datos para tres fuentes de datos originales (A, B y C).

Esencialmente, los almacenes de datos se utilizan para poder agregar y cruzar eficientemente la información de maneras sofisticadas. Por ello, los datos se modelan con una estructura de base de datos multidimensional, donde cada dimensión corresponde a un atributo o conjunto de atributos en el esquema en torno a unos "hechos" que almacenan el valor de alguna medida agregada, como por ejemplo la cantidad vendida de un producto en un día concreto en una tienda. Esta visión multidimensional hace a los almacenes de datos adecuados para el **procesamiento analítico en línea** (*on-line analytical processing*, **OLAP**). Las operaciones OLAP permiten un análisis multidimensional de los datos, que es superior al SQL para computar resúmenes y desgloses en muchas dimensiones, pudiendo utilizar conocimiento previo sobre el dominio de los datos para permitir su presentación a diferentes niveles de abstracción, acomodando así diferentes puntos de vista del usuario.

Una cuestión importante para los profesionales del procesamiento de datos es la diferencia entre minería de datos y OLAP. El usuario de una herramienta OLAP utiliza la herramienta para obtener información agregada a partir de información detallada, combinando la información de manera flexible. Esto permite obtener informes y vistas sofisticadas en tiempo real. Además, las herramientas OLAP pueden utilizarse para comprobar rápidamente patrones y pautas hipotéticas sugeridas por el usuario con el objetivo de verificarlas o rechazarlas. Se trata, por lo tanto, de un proceso esencialmente deductivo. Por el contrario, la minería de datos, más que verificar patrones hipotéticos, usa los datos para encontrar estos patrones. Por lo tanto, es un proceso inductivo. Ambos tipos de herramientas se complementan: podemos usar OLAP al principio del proceso de KDD para explorar los datos (por ejemplo, para centrar nuestra atención en las variables importantes, identificar excepciones o encontrar interacciones), ya que cuanto más comprendamos los datos más efectivo será el proceso de descubrir conocimiento.

Como hemos dicho, un almacén de datos es muy aconsejable para la minería de datos, aunque no imprescindible. En algunos casos, en especial cuando el volumen no es muy grande, se puede trabajar con los datos originales o en formatos heterogéneos (archivos de texto, hojas de cálculo...).

## 2.3 Fase de selección, limpieza y transformación

La calidad del conocimiento descubierto no sólo depende del algoritmo de minería utilizado, sino también de la calidad de los datos minados. Por ello, después de la recopilación, el siguiente paso en el proceso de KDD es seleccionar y preparar el subconjunto de datos que se va a minar, los cuales constituyen lo que se conoce como *vista minable*. Este paso es necesario ya que algunos datos coleccionados en la etapa anterior son irrelevantes o innecesarios para la tarea de minería que se desea realizar.

Pero además de la irrelevancia, existen otros problemas que afectan a la calidad de los datos. Uno de estos problemas es la presencia de valores que no se ajustan al comportamiento general de los datos (*outliers*). Estos datos anómalos pueden representar errores en los datos o pueden ser valores correctos que son simplemente diferentes a los demás. Algunos algoritmos de minería de datos ignoran estos datos, otros los descartan considerándolos ruido o excepciones, pero otros son my sensibles y el resultado se ve claramente perjudicado por ello. Sin embargo, no siempre es conveniente eliminarlos, ya que, en algunas aplicaciones como la detección de compras fraudulentas efectuadas con tarjetas de crédito o la predicción de inundaciones, los eventos raros pueden ser más interesantes que los regulares (por ejemplo, compras por un importe mucho más elevado que el de las compras efectuadas habitualmente con la tarjeta, o días en los que la cantidad de lluvia recogida es muy superior a la media).

La presencia de datos faltantes o perdidos (*missing values*) puede ser también un problema pernicioso que puede conducir a resultados poco precisos. No obstante, es necesario reflexionar primero sobre el significado de los valores faltantes antes de tomar ninguna decisión sobre cómo tratarlos ya que éstos pueden deberse a causas muy diversas, como a un mal funcionamiento del dispositivo que hizo la lectura del valor, a cambios efectuados en los procedimientos usados durante la colección de los datos o al hecho de que los datos se recopilen desde fuentes diversas. Por ello, existen muchas aproximaciones para manejar los datos faltantes, como veremos en el Capítulo 4.

Estos dos problemas son sólo dos ejemplos que muestran la necesidad de la limpieza de datos, es decir, de mejorar su calidad. Como hemos dicho, no es sólo suficiente con tener una buena calidad de datos, sino además poder proporcionar a los métodos de minería de datos el subconjunto de datos más adecuado para resolver el problema. Para ello es necesario seleccionar los datos apropiados.

La selección de atributos relevantes es uno de los preprocesamientos más importantes, ya que es crucial que los atributos utilizados sean relevantes para la tarea de minería de datos. Por ejemplo, supongamos que los jueces del torneo de Wimbledon desean determinar a partir de las condiciones climatológicas (nubosidad, humedad, temperatura, etc.) si se puede jugar o no al tenis. Para ello se cuenta con los datos recogidos de experiencias anteriores. Probablemente, la base de datos contenga un atributo que identifica cada uno de los días considerados (por ejemplo, la fecha). Si consideramos este atributo en el proceso de minería, un algoritmo de generación de reglas podría obtener reglas como

**SI** (fecha=10/06/2003) **ENTONCES** (jugar_tenis=sí)

que, aunque correcta, es inútil para realizar predicciones futuras.

Idealmente, uno podría usar todas las variables y dejar que la herramienta de minería de datos fuera probando hasta elegir las mejores variables predictoras. Obviamente, esta forma de trabajar no funciona bien, entre otras cosas porque el tiempo requerido para construir un modelo crece con el número de variables. Aunque en principio algunos algoritmos de minería de datos automáticamente ignoran las variables irrelevantes, en la práctica nuestro conocimiento sobre el dominio del problema puede permitirnos hacer correctamente muchas de esas selecciones.

Como en el caso de las variables, también podríamos construir el modelo usando todos los datos. Pero si tenemos muchos, tardaríamos mucho tiempo y probablemente también necesitaríamos una máquina más potente. Consecuentemente, una buena idea es usar una muestra (*sample*) a partir de algunos datos (o filas). La selección de la muestra debe ser hecha cuidadosamente para asegurar que es verdaderamente aleatoria.

Otra tarea de preparación de los datos es la construcción de atributos, la cual consiste en construir automáticamente nuevos atributos aplicando alguna operación o función a los atributos originales con objeto de que estos nuevos atributos hagan más fácil el proceso de minería. La motivación principal para esta tarea es fuerte cuando los atributos originales no tienen mucho poder predictivo por sí solos o los patrones dependen de variaciones lineales de las variables originales. Por ejemplo, el precio de las viviendas de una zona se puede estimar mucho mejor a partir de la densidad de población de la zona que de la población absoluta y de su superficie. Por tanto, es razonable derivar el atributo densidad de población de los otros dos.

El tipo de los datos puede también modificarse para facilitar el uso de técnicas que requieren tipos de datos específicos. Así, algunos atributos se pueden numerizar, lo que reduce el espacio y permite usar técnicas numéricas. Por ejemplo, podemos reemplazar los valores del atributo "tipo de vivienda" por enteros.

El proceso inverso consiste en discretizar los atributos continuos, es decir, transformar valores numéricos en atributos discretos o nominales. Los atributos discretizados pueden tratarse como atributos categóricos con un número más pequeño de valores. La idea básica es partir los valores de un atributo continuo en una pequeña lista de intervalos, tal que cada intervalo es visto como un valor discreto del atributo.

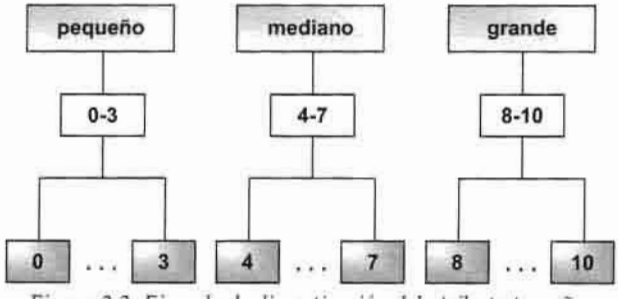


*Figura 2.3. Ejemplo de discretización del atributo tamaño.*

La Figura 2.3 ilustra una posible discretización para el atributo *tamaño*, con valores de 0 a 10. La parte inferior de la figura muestra la lista ordenada de los valores continuos, los cuales se han discretizado en tres intervalos a los que se les ha asignado los valores discretos *pequeño*, *mediano* y *grande*, como puede verse en la parte superior de la figura.

## 2.4 Fase de minería de datos

La fase de minería de datos es la más característica del KDD y, por esta razón, muchas veces se utiliza esta fase para nombrar todo el proceso. El objetivo de esta fase es producir nuevo conocimiento que pueda utilizar el usuario. Esto se realiza construyendo un modelo basado en los datos recopilados para este efecto. El modelo es una descripción de los patrones y relaciones entre los datos que pueden usarse para hacer predicciones, para

entender mejor los datos o para explicar situaciones pasadas. Para ello es necesario tomar una serie de decisiones antes de empezar el proceso:

- Determinar qué tipo de tarea de minería es el más apropiado. Por ejemplo, podríamos usar la clasificación para predecir en una entidad bancaria los clientes que dejarán de serlo.
- Elegir el tipo de modelo. Por ejemplo, para una tarea de clasificación podríamos usar un árbol de decisión, porque queremos obtener un modelo en forma de reglas.
- Elegir el algoritmo de minería que resuelva la tarea y obtenga el tipo de modelo que estamos buscando. Esta elección es pertinente porque existen muchos métodos para construir los modelos. Por ejemplo, para crear árboles de decisión para clasificación podríamos usar CART o C5.0, entre otros. En los capítulos siguientes se presentarán los métodos más importantes para cada tipo de modelo.

En lo que resta de esta sección, describimos las tareas y modelos más utilizados, así como algunos conceptos relacionados con la construcción del modelo. Una descripción más completa se dará en el Capítulo 6.

## 2.4.1 Tareas de la minería de datos

Dentro de la minería de datos hemos de distinguir tipos de tareas, cada una de las cuales puede considerarse como un tipo de problema a ser resuelto por un algoritmo de minería de datos. Esto significa que cada tarea tiene sus propios requisitos, y que el tipo de información obtenida con una tarea puede diferir mucho de la obtenida con otra.

Tal y como comentamos en el capítulo anterior, las distintas tareas pueden ser predictivas o descriptivas. Entre las tareas predictivas encontramos la clasificación y la regresión, mientras que el agrupamiento (*clustering*), las reglas de asociación, las reglas de asociación secuenciales y las correlaciones son tareas descriptivas. Veamos en mayor detalle todas ellas.

La **clasificación** es quizá la tarea más utilizada. En ella, cada instancia (o registro de la base de datos) pertenece a una clase, la cual se indica mediante el valor de un atributo que llamamos la clase de la instancia. Este atributo puede tomar diferentes valores discretos, cada uno de los cuales corresponde a una clase. El resto de los atributos de la instancia (los relevantes a la clase) se utilizan para predecir la clase. El objetivo es predecir la clase de nuevas instancias de las que se desconoce la clase. Más concretamente, el objetivo del algoritmo es maximizar la razón de precisión de la clasificación de las nuevas instancias, la cual se calcula como el cociente entre las predicciones correctas y el número total de predicciones (correctas e incorrectas).

**Ejemplo**. *Consideremos un oftalmólogo que desea disponer de un sistema que le sirva para determinar la conveniencia o no de recomendar la cirugía ocular a sus pacientes. Para ello dispone de una base de datos de sus antiguos pacientes clasificados en operados satisfactoriamente o no en función del tipo de problema que padecían (miopía y su grado, o astigmatismo) y de su edad. El modelo encontrado se utiliza para clasificar nuevos pacientes, es decir, para decidir si es conveniente operarlos o no.*

Existen variantes de la tarea de la clasificación, como son el aprendizaje de "*rankings*", el aprendizaje de preferencias, el aprendizaje de estimadores de probabilidad, etc.

La **regresión** es también una tarea predictiva que consiste en aprender una función real que asigna a cada instancia un valor real. Ésta es la principal diferencia respecto a la clasificación; el valor a predecir es numérico. El objetivo en este caso es minimizar el error (generalmente el error cuadrático medio) entre el valor predicho y el valor real.

**Ejemplo**. *Un empresario quiere conocer cuál es el costo de un nuevo contrato basándose en los datos correspondientes a contratos anteriores. Para ello usa una fórmula de regresión lineal, ajustando con los datos pasados la función lineal y usándola para predecir el costo en el futuro.*

El **agrupamiento** (*clustering*) es la tarea descriptiva por excelencia y consiste en obtener grupos "naturales" a partir de los datos. Hablamos de grupos y no de clases, porque, a diferencia de la clasificación, en lugar de analizar datos etiquetados con una clase, los analiza para generar esta etiqueta. Los datos son agrupados basándose en el principio de maximizar la similitud entre los elementos de un grupo minimizando la similitud entre los distintos grupos. Es decir, se forman grupos tales que los objetos de un mismo grupo son muy similares entre sí y, al mismo tiempo, son muy diferentes a los objetos de otro grupo. Al agrupamiento también se le suele llamar segmentación, ya que parte o segmenta los datos en grupos que pueden ser o no disjuntos. El agrupamiento está muy relacionado con la sumarización, que algunos autores consideran una tarea en sí misma, en la que cada grupo formado se considera como un resumen de los elementos que lo forman para así describir de una manera concisa los datos.

**Ejemplo.** *Una librería que ofrece sus servicios a través de la red usa el agrupamiento para identificar grupos de clientes en base a sus preferencias de compras que le permita dar un servicio más personalizado. Así, cada vez que un cliente se interesa por un libro, el sistema identifica a qué grupo pertenece y le recomienda otros libros comprados por clientes de su mismo grupo.*

Las **correlaciones** son una tarea descriptiva que se usa para examinar el grado de similitud de los valores de dos variables numéricas. Una fórmula estándar para medir la correlación lineal es el coeficiente de correlación $r$, el cual es un valor real comprendido entre –1 y 1. Si $r$ es 1 (respectivamente, –1) las variables están perfectamente correlacionadas (perfectamente correlacionadas negativamente), mientras que si es 0 no hay correlación. Esto quiere decir que cuando $r$ es positivo, las variables tienen un comportamiento similar (ambas crecen o decrecen al mismo tiempo) y cuando $r$ es negativo si una variable crece la otra decrece. El análisis de correlaciones, sobre todo las negativas, puede ser muy útil para establecer reglas de items correlacionados, como se muestra en el siguiente ejemplo.

**Ejemplo.** *Un inspector de incendios que desea obtener información útil para la prevención de incendios probablemente esté interesado en conocer correlaciones negativas entre el empleo de distintos grosores de protección del material eléctrico y la frecuencia de ocurrencia de incendios.*

Las **reglas de asociación** son también una tarea descriptiva, muy similar a las correlaciones, que tiene como objetivo identificar relaciones no explícitas entre atributos *categóricos*. Pueden ser de muchas formas, aunque la formulación más común es del estilo "si el atributo $X$ toma el valor $d$ entonces el atributo $Y$ toma el valor $b$". Las reglas de asociación no implican una relación causa-efecto, es decir, puede no existir una causa para que los datos estén asociados. Este tipo de tarea se utiliza frecuentemente en el análisis de la cesta

de la compra, para identificar productos que son frecuentemente comprados juntos, información esta que puede usarse para ajustar los inventarios, para la organización física del almacén o en campañas publicitarias. Las reglas se evalúan usando dos parámetros: precisión y soporte (cobertura)

**Ejemplo.** *Una compañía de asistencia sanitaria desea analizar las peticiones de servicios médicos solicitados por sus asegurados. Cada petición contiene información sobre las pruebas médicas que fueron realizadas al paciente durante una visita. Toda esta información se almacena en una base de datos en la que cada petición es un registro cuyos atributos expresan si se realiza o no cada una de las posibles pruebas médicas que pueden ser realizadas a un paciente. Mediante reglas de asociación, un sistema encontraría aquellas pruebas médicas que frecuentemente se realizan juntas, por ejemplo que un 70 por ciento de las veces que se pide un análisis de orina también se solicita uno de sangre, y esto ocurre en dos de cada diez pacientes. La precisión de esta regla es del 70 por ciento y el soporte del 20 por ciento.*

Un caso especial de reglas de asociación, que recibe el nombre de **reglas de asociación secuenciales**, se usa para determinar patrones secuenciales en los datos. Estos patrones se basan en secuencias temporales de acciones y difieren de las reglas de asociación en que las relaciones entre los datos se basan en el tiempo.

**Ejemplo.** *Una tienda de venta de electrodomésticos y equipos de audio analiza las ventas que ha efectuado usando análisis secuencial y descubre que el 30 por ciento de los clientes que compraron un televisor hace seis meses compraron un DVD en los siguientes dos meses.*

## 2.4.2 Técnicas de minería de datos

Dado que la minería de datos es un campo muy interdisciplinar, como vimos en la Sección 1.6, existen diferentes paradigmas detrás de las técnicas utilizadas para esta fase: técnicas de inferencia estadística, árboles de decisión, redes neuronales, inducción de reglas, aprendizaje basado en instancias, algoritmos genéticos, aprendizaje bayesiano, programación lógica inductiva y varios tipos de métodos basados en núcleos, entre otros. Cada uno de estos paradigmas incluye diferentes algoritmos y variaciones de los mismos, así como otro tipo de restricciones que hacen que la efectividad del algoritmo dependa del dominio de aplicación, no existiendo lo que podríamos llamar el método universal aplicable a todo tipo de aplicación.

A continuación revisamos los aspectos principales de algunas de las técnicas mencionadas. Para un estudio más detallado y profundo de paradigmas y algoritmos, remitimos al lector a la Parte III de este libro.

Existen muchos conceptos **estadísticos** que son la base de muchas técnicas de minería de datos. Ya hemos mencionado la regresión lineal, como un método simple pero frecuentemente utilizado para la tarea de regresión, como se muestra en la Figura 2.4.

En general, la fórmula para una regresión lineal es $y=c_0 + c_1x_1 + ... + c_nx_n$, donde $x_i$ son los atributos predictores e $y$ la salida (la variable dependiente). Si los atributos son modificados en la función de regresión por alguna otra función (cuadrados, inversa, logaritmos, combinaciones de variables...), es decir $y=c_0 + f_1(x_1) + ... + f_n(x_n)$, la regresión se dice no lineal. Se pueden incorporar variantes locales o transformaciones en las variables