

# Aprendizaje

---

1. Preliminares
2. Algoritmos genéticos y redes neuronales
3. Inducción de árboles clasificadores
4. Inducción de reglas
5. **Minería de datos**

## Minería de datos

---

- Definiciones, aplicaciones y técnicas
- Clasificadores bayesianos en minería de textos
- Minería de web
- Herramientas

## Sobrecarga de información

---

- Vannevar Bush, «As we may think» (1945): «Memex» como ayuda a la recuperación de información
- Milford y Perry (1977), «Information overload»: Cuando el volumen de estímulos (datos de entrada) al procesador cognitivo supera su capacidad de procesamiento
- Blum (1980): Extracción de conocimiento de bases de datos médicas
- Date (1990). Publicaciones de BD: 100.000 págs./año
- Piatetsky–Shapiro y Frawley (1991): La cantidad de «información» en el mundo se duplica cada 20 meses.

## La «ley de Piatetsky» en la web

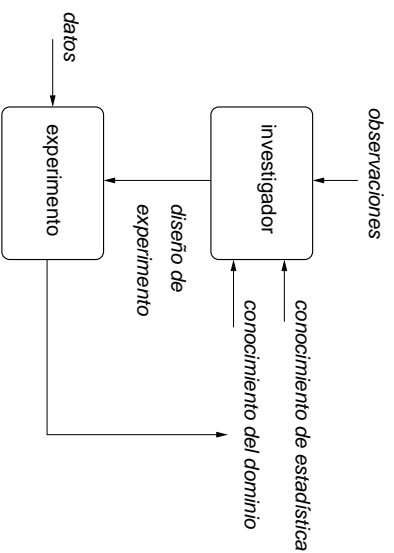
---

- En 1994 WWW tenía indexadas 110.000 páginas web
- En 1995 se estimaban 50 M ficheros disponibles  
Previsiones para el 2000: 150 M
- En 2000, [alltheweb.com](http://alltheweb.com): 300 M páginas web y 100 M ftp
- Google: 3.084 M documentos en 2002  
8.058 M en noviembre 2004  
1 B («trillion»:  $10^{12}$ ) (?) en julio 2008
- The Internet Archive Wayback Machine ([www.archive.org](http://www.archive.org)):  
100 TB en 2001  
2 PB (85.000 M págs.) en 2008,  
creciendo a 20 TB/mes

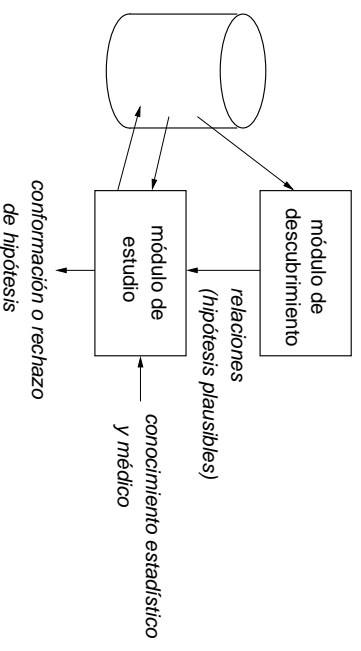
## Extracción de conocimiento en bases de datos

Proyecto RX: «programa que examina una base de datos clínica y genera un conjunto de posibles relaciones causales»  
(Blum, 1980)

**Modelo:**



**Implementación:**



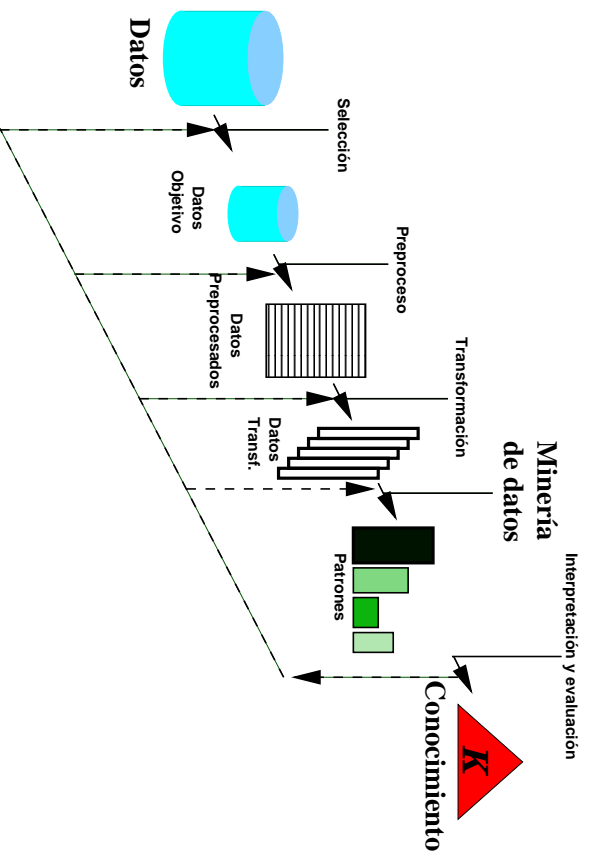
## KDD y DM: definiciones

**KDD:** descubrimiento de conocimiento en bases de datos

**DM:** minería de datos

- **KDD = proceso completo:**  
«extracción no trivial de conocimiento implícito, previamente desconocido y potencialmente útil, a partir de una base de datos»  
(Frawley *et al.*, 1991)

- **DM = etapa de descubrimiento** en el proceso de KDD:  
«paso consistente en el uso de algoritmos concretos que generan una enumeración de patrones a partir de los datos preprocesados»  
(Fayyad *et al.*, 1996)



## Selección, preproceso, transformación

### Selección previa:

- Si es un almacén (*data warehouse*), elección de vistas
- Si los datos están en una BDR, elección las tablas más adecuadas (dependiendo de los objetivos)
- Si las relaciones tienen muchas tuplas, muestreo

### Preproceso:

- Filtrado de valores imposibles, valores ausentes...
- Selección de atributos más relevantes (búsqueda en el espacio de características)
- Definición de nuevos atributos, función de otros

### Transformación para adaptar al algoritmo de DM:

- Redes neuronales, SVM... requieren valores numéricos  $\rightsquigarrow$  codificación de valores nominales (nombres, direcciones...), normalización...
- Otros requieren valores nominales  $\rightsquigarrow$  discretización de los numéricos

## Ejemplo típico de aplicación de la minería de datos (1)

---

- Bank of America
- Objetivo: captar clientes para potenciar créditos con garantía hipotecaria
- Premisas de campaña de marketing basadas en «sentido común»: personas con hijos en edad escolar, y personas con ingresos altos pero variables. Resultados discretos
- Inducción de un árbol de clasificación (solicita crédito o no) con un subconjunto de la base de datos de clientes según diversos atributos: tipo de cuenta, situación familiar...
- Se aplicó al resto de clientes y se etiquetaron como «probables» (11%) o «no probables» (89%)

## Ejemplo típico de aplicación de la minería de datos (2)

---

Por otra parte,

- Algoritmo de agrupamiento (*clustering*) para segmentar a todos los clientes
  - 14 grupos, uno interesante:
    - 39% tenían una cuenta personal y una cuenta de negocios
    - el grupo incluía al 27% de los clientes que el árbol había etiquetado como «probables»
  - Nueva premisa de campaña: los clientes utilizan los créditos para iniciar pequeñas empresas
  - Con la nueva campaña se dobló la tasa de éxitos
- (Berry y Linof: *Data Mining Techniques*, Wiley, 1997)

## MD: aplicaciones actuales

---

- Análisis de la cesta de la compra mediante reglas de asociación
- Modelos para análisis de riesgos (seguros, créditos...)
- Evaluación de campañas publicitarias
- Análisis de la fidelidad de clientes (*churning*)

Otras:

- Análisis de valores de bolsa
- Detección y prevención de fraude en comercio electrónico
- Modelos de tráfico a partir de datos GPS
- Perfiles de usuarios de redes
- Detección de intrusos en redes

*KDD-2006 Conference on Knowledge Discovery and Data Mining*

## Minería de datos: varios enfoques

---

Comunidades (y enfoques) involucradas en la MD:

- Estadística: regresión, agrupamiento numérico...
- **Inteligencia artificial, aprendizaje**
- Visualización

## Aprendizaje vs. Minería de datos

---

- **Objetivos:**
  - **Aprendizaje:** mejorar el funcionamiento de un agente
    - Conductista: modificando su estructura
    - Cognitivo: induciendo previamente el conocimiento
  - **Minería de datos:** adquisición de conocimiento para su uso por personas u organizaciones
- **Métodos:**
  - **Aprendizaje:** pocos ejemplos (eficiencia y escalabilidad de algoritmos no importa)
  - **Minería de datos:** gran número de datos, incompletos, ruido...

---

## Técnicas aplicadas en MD

- **Sin algoritmos de aprendizaje**
  - Consultas (SQL)
  - OLAP (OnLine Analytical Processing) en BDM
  - Estadísticas: correlación, regresión, agrupamiento (*clustering*) numérico
- **Con algoritmos «clásicos» de aprendizaje**
  - Redes neuronales y algoritmos genéticos
  - Inducción de árboles y reglas de clasificación
  - Agrupamiento conceptual
- **«Nuevos» algoritmos**
  - Inducción de reglas de asociación
  - Inducción de clasificadores bayesianos

## Clasificadores bayesianos

---

Clasificador basado en el teorema de Bayes:

Si  $h$  es una hipótesis

(conocida  $h$  se conoce la clase)

y  $D$  son los datos

(valores de los atributos para un ejemplar nuevo),

$$P(h|D) = \frac{P(h) \cdot P(D|h)}{P(D)} \quad \text{con } P(D) = \sum_j P(h_j) \cdot P(D|h_j)$$

- Hipótesis máxima a posteriori:

$$h_{\text{MAP}} = \operatorname{argmax}_{h \in H} P(h|D) = \operatorname{argmax}_{h \in H} \frac{P(h) \cdot P(D|h)}{P(D)} = \operatorname{argmax}_{h \in H} (P(h) \cdot P(D|h))$$

- Hipótesis de máxima verosimilitud: Si  $h_i$  equiprobables,

$$h_{\text{MV}} = \operatorname{argmax}_{h \in H} P(D|h)$$

---

## Clasificador bayesiano óptimo

Función de clasificación:  $f_c : H \rightarrow C$

Si  $H \neq C$ ,  $h_{\text{MAP}}$  no da siempre la mejor clasificación

Ejemplo:

- $H = \{h_1, h_2, h_3\}$ ;  $C = \{\oplus, \ominus\}$
- $f_c(h_1) = \oplus, f_c(h_2) = \ominus, f_c(h_3) = \ominus$
- Si  $P(h_1|D) = 0,4, P(h_2|D) = 0,3, P(h_3|D) = 0,3$ ,  
 $h_{\text{MAP}} = h_1$ , pero el resultado debe ser  $\ominus$   
( $P(\oplus) = 0,4, P(\ominus) = 0,6$ )

En general (con  $f_c$  aleatorio):

$$c_{\text{MAP}} = \operatorname{argmax}_{c_j \in C} P(c_j|D) = \operatorname{argmax}_{c_j \in C} \sum_{h_i \in H} P(c_j|h_i) \cdot P(h_i|D)$$

## El clasificador bayesiano óptimo en la práctica

---

- Supongamos  $C = H$  y
- $D = x_1 \text{ Y } x_2 \text{ Y } \dots \text{ Y } x_n$

$$\begin{aligned} C_{\text{MAP}} &= \operatorname{argmax}_{c_j \in C} P(c_j | x_1, x_2 \dots x_n) = \\ &= \operatorname{argmax}_{c_j \in C} \frac{P(c_j) \cdot P(x_1, x_2 \dots x_n | c_j)}{P(x_1, x_2 \dots x_n)} \\ &= \operatorname{argmax}_{c_j \in C} (P(c_j) \cdot P(x_1, x_2 \dots x_n | c_j)) \end{aligned}$$

- $P(c_j)$ : simple recuento (frecuencias)
- $P(x_1, x_2 \dots x_n | c_j)$ : **inviable**

---

## Clasificador bayesiano ingenuo (naïve)

---

**Suposición:** los valores de los atributos, dado  $c_j$ , son condicionalmente independientes

$$P(x_1, x_2 \dots x_n | c_j) = P(x_1 | c_j) \cdot P(x_2 | c_j) \dots P(x_n | c_j)$$

$$C_{\text{NB}} = \operatorname{argmax}_{c_j \in C} P(c_j) \prod_i P(x_i | c_j)$$

- Sencillo
- Resultados sorprendentemente buenos (comparables a RN y árboles)
- Muy utilizado en las herramientas

## Clasificador bayesiano incremental

---

Si  $x_2$  «viene después» de  $x_1$ , probabilidad de  $h_j$  condicionada a  $x_2$  en el contexto de  $x_1$ :

$$P(h_j|x_1 \cap x_2) = \frac{P(h_j|x_1)P(x_2|h_j \cap x_1)}{P(x_1 \cap x_2)}$$

**Suposición:**  $P(x_i|h_j \cap x_k) = P(x_i|h_j)$  ( $h_j$  causa directa de  $x_i$ )

$$P(h_j|x_1 \cap x_2) = \frac{P(h_j|x_1)P(x_2|h_j)}{P(x_1 \cap x_2)}$$

Es decir, Bayes aplicado al resultado de tener en cuenta  $x_1$

⇒ **algoritmo de actualización sucesiva de  $c_{MAP}$**

---

## Minería de información

- «Proceso de extraer información previamente desconocida, integrada y útil, de cualquier fuente
- »– transacciones, documentos, mensajes, páginas web, etc. –
  - »y utilizarla para la toma de decisiones»

D.S. Tkach: *Information Mining*

with the *IBM Intelligent Miner Family*. IBM White Paper, 1998.

**Minería de textos:**

De datos estructurados a datos no estructurados...

## Minería de textos: ejemplo (1)

---

Clasificación de documentos recibidos en un grupo de *news* como «interesantes» o «no interesantes» para un usuario

- $C = \{\oplus, \ominus\}$ ,  $E = \{\text{documentos}\}$
- Extracción de características: un atributo por cada posición de palabra cuyo valor es la palabra en esa posición. Para este párrafo: 35 posiciones,  $a_1 = \text{«Extracción»}$ ... $a_{35} = \text{«más»}$  (En un documento real, muchos más)
- $c_{NB} = \operatorname{argmax}_{c_j \in \{\oplus, \ominus\}}$   
 $P(c_j) \cdot P(a_1 = \text{«Extracción»} | c_j) \cdot \dots \cdot P(a_{35} = \text{«más»} | c_j)]$
- ¿Cómo calcular las probabilidades?

## Minería de textos: ejemplo (2)

---

Estimaciones de probabilidades:

- Para  $P(c_j)$  fácil: proporciones de ejemplos  
Si hay 300  $\oplus$  y 700  $\ominus$ ,  $P(\oplus) = 0,3$  y  $P(\ominus) = 0,7$
- Suposición adicional:  $P(a_i = x_k | c_j) = P(x_k | c_j)$   
 $P(x_k | c_j) = \frac{n_k + 1}{n + |V|}$  («estimador  $m$ »), con
  - $n = n^{\ominus}$  total posiciones en ejemplos clasificados  $c_j$
  - $n_k = n^{\ominus}$  veces  $x_k$  está en una de esas  $n$
  - $|V| = n^{\ominus}$  total de palabras diferentes en  $E$

Resultados en casos reales con precisiones  $\approx 90\%$   
(Mitchell, 1997)

## Minería de web: perdidos en el hiperespacio

---

**Problema de la navegación:** el «navegante» no sabe qué dirección tomar para llegar al «destino» (información que busca)

⇒ «*getting lost in hyperspace*» (Conklin, 1987)

Intentos de solución:

- Directorios de recursos (yahoo.com, dmoz.org...)
- Motores de búsqueda (google.com, wisenut.com...)
- Sitios web adaptativos
- Agentes asistentes
- Web semántica

■ ...

## Aplicaciones de minería de datos en la web

---

Además de ayudas a la navegación,

- Detección de ataques y fraudes en comercio electrónico  
Ejemplos en <http://www.cs.columbia.edu/ids/>
- Mejoras del diseño de los sitios
- Caracterización de visitantes
- Personalización de páginas
- ...

## Minería de web: tipos

---

- **Minería de contenido**  
Tipo especial (o generalización) de minería de textos (hipertextos)
- **Minería de estructura**  
Descubrimiento de relaciones estructurales entre páginas
- **Minería de uso**  
Descubrimiento de patrones de acceso, perfiles de usuario...

## Minería de estructura de la web

---

Hace uso de la información contenida en los enlaces (*hyperlinks*)

- Búsqueda de páginas «autoritativas»: algoritmo PageRank de Larry Page y Sergey Brin (Google)
- Búsqueda de páginas «eruditas» (o «concentradores», *hubs*): algoritmo HITS (Hypertext Induced Topic Selection) de Kleinberg
- «La web como un grafo»: los nodos son las páginas y los arcos orientados los enlaces de unas páginas hacia otras
- Idea básica: la existencia de un enlace en una página hacia otra representa un «reconocimiento de autoridad» del autor de la primera hacia la segunda
- Definición recursiva de «autoridad»: una página es «importante» si otras páginas importantes apuntan a ella

En los ficheros de registros (*logs*) del servidor web:

- Fecha y hora
- Elemento servido (URL)
- Parámetros (para consultas)
- Bytes reales/transferidos
- Cookie
- ...
- Dirección IP de acceso
- Navegador y versión
- Errores
- Tiempo en transferir
- Página anterior

```
host196_20.inter.edu - - [20/Nov/2001:14:35:22 +0100] "GET /~gfer/ssii/intro-ia/T004.jpeg HTTP/1.0" 200 3393
host196_20.inter.edu - - [20/Nov/2001:14:35:24 +0100] "GET /~gfer/ssii/intro-ia/T008.jpeg HTTP/1.0" 200 3062
64.76.178.101 - - [20/Nov/2001:14:35:24 +0100] "GET /~nga/is/ejemplos/parking2.zip HTTP/1.1" 200 50619
64.76.178.101 - - [20/Nov/2001:14:35:24 +0100] "GET /~nga/is/ejemplos/acceso1.zip HTTP/1.1" 200 50619
host196_20.inter.edu - - [20/Nov/2001:14:35:25 +0100] "GET /~gfer/ssii/intro-ia/T009.jpeg HTTP/1.0" 200 3197
host196_20.inter.edu - - [20/Nov/2001:14:35:26 +0100] "GET /~gfer/ssii/intro-ia/T007.jpeg HTTP/1.0" 200 3142
host196_20.inter.edu - - [20/Nov/2001:14:35:27 +0100] "GET /~gfer/ssii/intro-ia/P001.html HTTP/1.0" 200 501
```

 © 2010 DIT-ETSIT-UPM

Minería de datos

transp. 27

---

## Herramientas para el proceso completo de KDD

Entornos de desarrollo integrado (IDE) que contienen implementaciones de redes neuronales, inducción de árboles, etc, junto con utilidades para preprocesamiento, clasificación, agrupamiento, visualización...

Piatetsky-Shapiro da enlaces a 63 privativas y 17 libres o shareware:  
<http://www.kdnuggets.com/software/suites.html>

Dos herramientas libres:

- **WEKA** (Waikato Environment for Knowledge Analysis)  
Universidad de Waikato, Nueva Zelanda  
<http://www.cs.waikato.ac.nz/ml/weka/>
- **RapidMiner** (antes, **YALE**: Yet Another Learning Environment)  
Universidad de Dortmund, Alemania, <http://rapid-i.com/>

Ambas GPL, implementadas en Java y muy bien documentadas

---

UCI (University of California, Irvine) Machine Learning Repository:  
<http://archive.ics.uci.edu/ml/>, 174 datasets

---

