

# Introducción a la Minería de Datos

Abdelmalik Moujahid, Iñaki Inza y Pedro Larrañaga

Departamento de Ciencias de la Computación e Inteligencia Artificial  
Universidad del País Vasco

# Índice

**1 Minería de datos**

**2 Proceso de extracción de conocimiento**

# Índice

**1 Minería de datos**

2 Proceso de extracción de conocimiento

# Algunas definiciones

- **Data mining**. Minería de datos como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos (Witten y Frank, 2000)
- **Knowledge discovery in databases**. Descubrimiento de conocimiento en bases como proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles, y en última instancia, comprensibles a partir de los datos (Fayyad y col. 1996)

# Tipos de modelos

- De **datos** a **conocimiento** a través de **modelos computacionales**
- **Modelos descriptivos**: identifican patrones que explican o resumen los datos
  - **Reglas de asociación**: expresan patrones de comportamiento en los datos
  - **Clustering**: agrupación de casos homogéneos
- **Modelos predictivos**: estiman valores de variables de interés (a predecir) a partir de valores de otras variables (predictoras)
  - **Regresión**: Variable a predecir continua
  - **Clasificación supervisada**: Variable a predecir discreta (nominal u ordinal)

# Tipos de datos

- **Bases de datos relacionales**
  - Colección de relaciones (tablas). Tabla como conjunto de atributos (variables, columnas, campos) conteniendo tuplas (casos, filas, registros)
  - Presentación tabular: atributo-valor (vista minable)
- **Bases de datos espaciales**: datos geográficos, imágenes médicas, redes de transporte o tráfico, ....
- **Bases de datos temporales**: distintos instantes o intervalos temporales
- **Bases de datos documentales**: objetos son documentos de texto, variables desde palabras hasta resúmenes
- **Bases de datos multimedia**: imágenes, audio, video
- **La World Wide Web**: repositorio de información mas grande y diverso en la actualidad
  - Minería del contenido: encontrar patrones en las páginas web
  - Minería de la estructura: estudia los hipervínculos y URLs
  - Minería del uso: análisis de la navegación

# Relación con otras disciplinas

- **Estadística.** "Madre" de la minería de datos
- **Aprendizaje automático.** El ordenador aprende a partir de ejemplos
- **Reconocimiento de patrones.** Clustering y clasificación supervisada
- **Sistemas para la toma de decisión.** Herramientas y sistemas que asisten al directivo
- **Visualización de datos.** Descubrir, intuir o entender
- **Bases de datos.** Almacenes de datos. Acceso eficiente a los datos
- **Recuperación de la información.** Datos textuales. Bibliotecas digitales. Búsqueda por Internet

# Minería de datos versus estadística

- **Estadística** (Análisis de datos)
  - **Asunciones**: la población sigue una distribución probabilística previamente especificada
  - **Score**: verosimilitud de los datos dado el modelo
  - **Búsqueda**: modelización basada en un algoritmo *hill-climbing en combinación con un test de hipótesis basado en la razón de verosimilitud*
  - **No funcionan bien en**: *bases de datos de gran tamaño y alta dimensionalidad*
- **Minería de datos**
  - **Mayor libertad** en la construcción de modelos. Interpretabilidad y comprensión
  - **Score**: a veces más directo (porcentaje de bien clasificados)
  - **Búsqueda**: metaheurísticos

# Aplicaciones

## Financieras

- Detección de uso fraudulento de tarjetas de crédito
- Predicción del gasto en tarjeta de crédito por grupos
- Análisis de riesgos en concesión de créditos
- Identificación de reglas de mercado a partir de datos históricos
- Reconocimiento de clientes "infieles"

# Aplicaciones

## Comercio

- Análisis de la cesta de la compra
- Evaluación de campañas publicitarias
- Segmentación de clientes
- Estimación de *stocks*, de costes, de ventas

# Aplicaciones

## Seguros

- Determinación de clientes potencialmente caros
- Predicción de que tipo de clientes contratan nuevas pólizas
- Identificación de patrones de comportamiento para clientes con riesgo
- Identificación de comportamiento fraudulento

# Aplicaciones

## Educación

- Selección o captación de estudiantes
- Detección de abandonos o fracasos
- Estimación del tiempo de estancia en la institución

# Aplicaciones

## Medicina

- Diagnóstico de enfermedades
- Detección de pacientes con riesgo de sufrir una patología concreta
- Gestión hospitalaria y asistencial. Predicciones temporales de los centros sanitarios para el mejor uso de recursos
- Tratamiento de imágenes médicas

# Aplicaciones

## Bioinformática

- Búsqueda de genes (regiones codificantes del genoma)
- Predicción de la estructura secundaria de las proteínas
- Búsqueda de biomarcadores a partir de datos de microarrays o de datos de espectrometría de masas

# Aplicaciones

## Otras áreas

- Telecomunicaciones: detección del fraude
- Correo electrónico y agendas personales: clasificación y distribución automática de correo, detección de correo *spam*
- Hacienda: detección de fraude fiscal
- Web: análisis del comportamiento de los usuarios, análisis de los log de un servidor web
- Deportes: detección riesgo de lesiones a partir de datos médicos

# Índice

1 Minería de datos

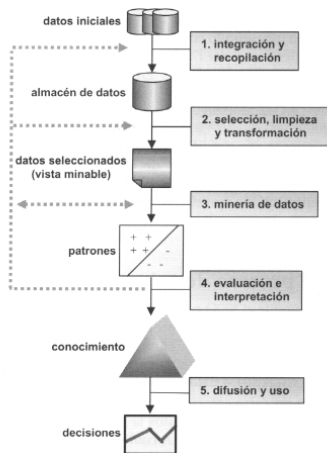
2 Proceso de extracción de conocimiento

# Knowledge Discovery from Databases (KDD)

## Fases del proceso iterativo e interactivo

- 1 Integración y recopilación de datos
- 2 Selección, limpieza y transformación
- 3 Minería de datos
- 4 Evaluación e interpretación
- 5 Difusión y uso

# Knowledge Discovery from Databases (KDD)



**Figura:** Proceso de extracción de conocimiento

# Knowledge Discovery from Databases (KDD)

## 1. Integración y recopilación de datos

- **Procesamiento transaccional en línea** (On-Line Transaction Processing, OLTP): suficiente para necesidades diarias (facturación, control de inventario, ...)
- **Decisiones estratégicas** basadas en el análisis, la planificación y la predicción: datos en varios departamentos
- **Cada fuente de datos** distintos formatos de registro, diferentes grados de agregación, diferentes claves primarias, ....
- Integración de múltiples bases de datos: **almacenes de datos** (data warehousing)
- Almacén de datos aconsejable cuando el volumen de información es grande. No estrictamente necesario (**archivos de texto, hojas de cálculo, ...**)

# Knowledge Discovery from Databases (KDD)

## 2. Selección, limpieza y transformación

- Calidad del conocimiento descubierto depende (además del algoritmo de minería) de la **calidad de los datos analizados**
- Presencia de datos que no se ajustan al comportamiento general de los datos (**outliers**)
- Presencia de datos perdidos (**missing values**)
- Selección de variables relevantes (**feature subset selection**)
- **Selección de casos aleatoria** en bases de datos de tamaño ingente. Muestreo aleatorio simple, por conglomerados, estratificado, polietápico
- Construcción automática de **nuevas variables** que faciliten el proceso de minería de datos
- **Discretización** de variables continuas

# Knowledge Discovery from Databases (KDD)

## 3. Minería de datos

- **Modelos descriptivos**
  - **Reglas de asociación**
  - **Clustering**: particional, probabilístico, jerárquico, conceptual
- **Modelos predictivos**:
  - **Regresión**: regresión lineal, regression tree, model tree, additive regression
  - **Clasificación supervisada**: clasificadores Bayesianos, regresión logística, redes neuronales, árboles de clasificación, inducción de reglas, K-NN, combinación de clasificadores

# Knowledge Discovery from Databases (KDD)

## 4. Evaluación e interpretación

- Técnicas de evaluación: **validación simple** (training + test), **validación cruzada con  $k$ -rodajas**, **bootstrapping**
- Reglas de asociación: **cobertura** (soporte), **confianza**
- Clustering: **variabilidad intra y entre**
- Regresión: **error cuadrático medio**
- Clasificación supervisada: **porcentaje de bien clasificados**, **matriz de confusión**, **análisis ROC**
- Modelos **precisos**, **comprensibles** (inteligibles) e **interesantes** (útiles y novedosos)

# Knowledge Discovery from Databases (KDD)

## 5. Difusión y uso

- **Difusión**: necesario distribuir, comunicar a los posibles usuarios, integrarlo en el *know-how* de la organización
- Medir la **evolución del modelo** a lo largo del tiempo (patrones tipo pueden cambiar)
- Modelo debe **cada cierto tiempo** de ser:
  - Reevaluado
  - Reentrenado
  - Reconstruido

# Introducción a la Minería de Datos

Abdelmalik Moujahid, Iñaki Inza y Pedro Larrañaga

Departamento de Ciencias de la Computación e Inteligencia Artificial  
Universidad del País Vasco