

Minería de Datos

I. Introducción

Pedro Larrañaga, Iñaki Inza, Abdelmalik Moujahid

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

MMCC, 2006-2007

Índice

- 1 **Introducción a la minería de datos**
- 2 **Paradigmas de minería de datos**
- 3 **Campos de aplicación de la minería de datos**
- 4 **Conclusiones**

Índice

- 1** **Introducción a la minería de datos**
- 2 Paradigmas de minería de datos
- 3 Campos de aplicación de la minería de datos
- 4 Conclusiones

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)

Reconocimiento de patrones

Ideas básicas

- **Clasificación supervisada**
 - Dadas N **instancias** (objetos, ejemplos ...) caracterizadas por sus **variables predictoras** (atributos) y una **etiqueta** variable clase
 - El objetivo es **transformar estos datos en un modelo de clasificación** capaz de predecir con una alta fiabilidad la clase de un nuevo ejemplo caracterizado por sus variables predictoras
- **Clasificación no supervisada**
 - Dadas N **instancias** (objetos, ejemplos, ...) caracterizados por sus **atributos**
 - El objetivo es **obtener grupos** (clusters) de ejemplos con una **alta variabilidad entre los clusters y una baja variabilidad dentro de cada cluster**
 - Tres tipos de métodos: clustering **particional**, clustering **jerárquico** y clustering **probabilístico**

Reconocimiento de patrones

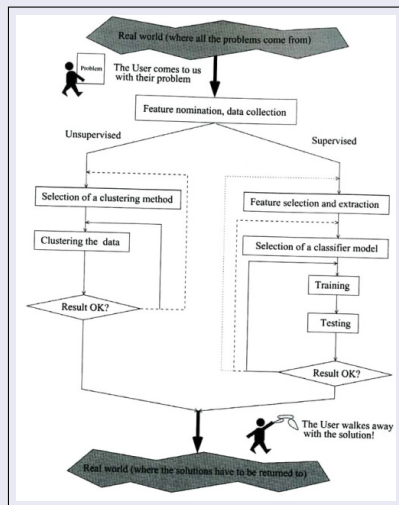


Figura: Ciclo de un sistema de reconocimiento. Tomado del libro de Kuncheva (2004)

De datos etiquetados a modelos clasificatorios

	X_1	\dots	X_n	C
$(\mathbf{x}^{(1)}, c^{(1)})$	$x_1^{(1)}$	\dots	$x_n^{(1)}$	$c^{(1)}$
$(\mathbf{x}^{(2)}, c^{(2)})$	$x_1^{(2)}$	\dots	$x_n^{(2)}$	$c^{(2)}$
\dots	\dots	\dots	\dots	\dots
$(\mathbf{x}^{(N)}, c^{(N)})$	$x_1^{(N)}$	\dots	$x_n^{(N)}$	$c^{(N)}$

Figura: Un fichero de N casos conteniendo ejemplos caracterizados por n variables predictoras y una variable clase

Ejemplo: Reconocimiento óptico de caracteres

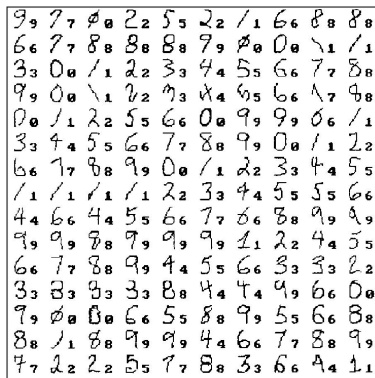


Figura: Reconocimiento de caracteres escritos a mano

Ejemplo: Predicción de tiempo atmosférico

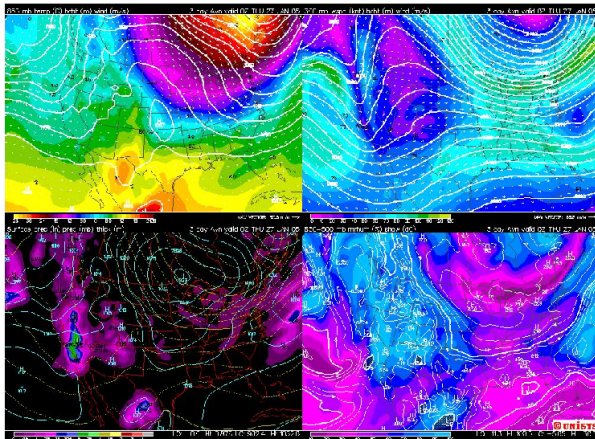


Figura: Tiempo atmosférico

Ejemplo: Biología computacional

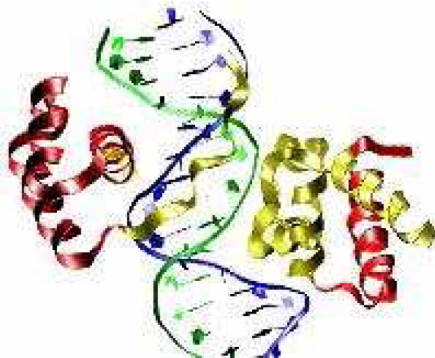


Figura: Predicción de la estructura secundaria de las proteínas

Reconocimiento supervisado de patrones



Figura: Humano como reconocedor de objetos

Clasificación semisupervisada

- Hay problemas en los que algunos ejemplos están etiquetados pero **la gran mayoría** de los ejemplos no están etiquetados (**unlabelled**)
- En **web mining**, podemos estar interesados en algunas páginas webs (**ejemplos positivos**), no interesados en otras páginas webs (**ejemplos negativos**), y no tener opinión sobre un gran porcentaje de páginas webs (**no etiquetados**)

Clasificación semisupervisada

	X_1	...	X_n	C
$(\mathbf{x}^{(1)}, c^{(1)})$	$x_1^{(1)}$...	$x_n^{(1)}$	1
$(\mathbf{x}^{(2)}, c^{(2)})$	$x_1^{(2)}$...	$x_n^{(2)}$	0
$(\mathbf{x}^{(3)}, c^{(3)})$	$x_1^{(3)}$...	$x_n^{(3)}$?
$(\mathbf{x}^{(4)}, c^{(4)})$	$x_1^{(4)}$...	$x_n^{(4)}$?
$(\mathbf{x}^{(5)}, c^{(5)})$	$x_1^{(5)}$...	$x_n^{(5)}$	1
$(\mathbf{x}^{(6)}, c^{(6)})$	$x_1^{(6)}$...	$x_n^{(6)}$?
...	
$(\mathbf{x}^{(N)}, c^{(N)})$	$x_1^{(N)}$...	$x_n^{(N)}$?

Figura: Fichero de casos conteniendo ejemplos etiquetados y no etiquetados

Clasificación parcialmente supervisada

- Descubrimiento de genes (instancias) asociados con un enfermedad dada (variable clase)
- Conocemos que algunos de los genes están asociados con la enfermedad (instancias positivas)
- Para el resto de los genes no es posible decir que no estén asociados con la enfermedad (no tenemos instancias negativas)

Clasificación parcialmente supervisada

	X_1	...	X_n	C
$(\mathbf{x}^{(1)}, c^{(1)})$	$x_1^{(1)}$...	$x_n^{(1)}$	1
$(\mathbf{x}^{(2)}, c^{(2)})$	$x_1^{(2)}$...	$x_n^{(2)}$	1
$(\mathbf{x}^{(3)}, c^{(3)})$	$x_1^{(3)}$...	$x_n^{(3)}$	1
$(\mathbf{x}^{(4)}, c^{(4)})$	$x_1^{(4)}$...	$x_n^{(4)}$?
$(\mathbf{x}^{(5)}, c^{(5)})$	$x_1^{(5)}$...	$x_n^{(5)}$?
$(\mathbf{x}^{(6)}, c^{(6)})$	$x_1^{(6)}$...	$x_n^{(6)}$?
...	
$(\mathbf{x}^{(N)}, c^{(N)})$	$x_1^{(N)}$...	$x_n^{(N)}$?

Figura: Fichero de casos con ejemplos etiquetados como positivos y ejemplos no etiquetados

Clasificación no supervisada. Clustering

	X_1	...	X_i	...	X_n
O_1	x_1^1	...	x_i^1	...	x_n^1
...
O_j	x_1^j	...	x_i^j	...	x_n^j
...
O_N	x_1^N	...	x_i^N	...	x_n^N

Figura: N instancias caracterizadas por n variables

Clustering de datos de microarrays

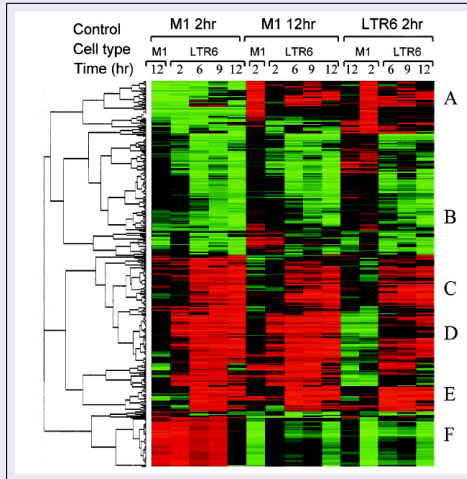


Figura: Clustering jerárquico de datos de microarrays

Clustering de datos de microarrays

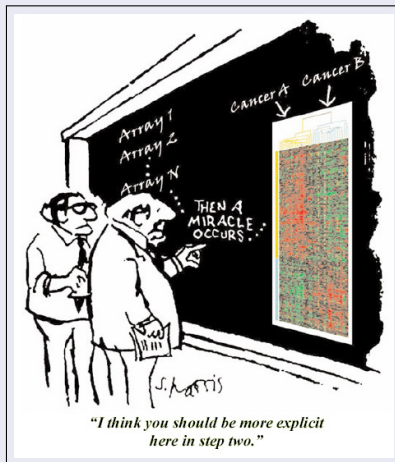


Figura: De datos de microarrays a clustering de genes

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
- **Computación paralela y distribuida.** Procesamiento paralelo, distribuido o computación en *grid*

Minería de datos versus estadística

- **Estadística** (Análisis de datos)
 - **Encorsetamiento**: premisas, teoremas, independencia de muestras, modelos a veces crípticos
 - **Score**: verosimilitud de los datos dado el modelo
 - **Búsqueda**: modelización basada en el test de la razón de verosimilitud (hacia adelante, hacia atrás, paso a paso)
 - **No funcionan bien en**: bases de datos de gran tamaño y alta dimensionalidad o con datos textuales, multimedia, variables nominales con gran número de valores distintos, no se integran bien en sistemas de información
- **Minería de datos**
 - **Mayor libertad** en la construcción de modelos. Interpretabilidad y comprensión
 - **Score**: a veces más directo
 - **Búsqueda**: metaheurísticos

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)

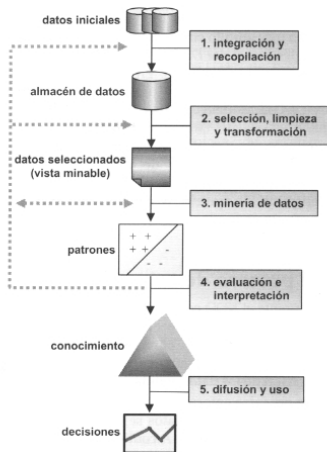


Figura: Proceso de extracción de conocimiento

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, jerárquico, probabilístico,
- **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

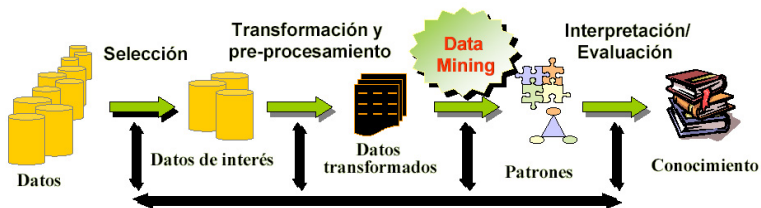
- Técnicas de evaluación: **validación simple** (training + test), **validación cruzada con k -rodajas**, **bootstrapping**
- Reglas de asociación: **cobertura** (soporte), **confianza**
- Clustering: **variabilidad intra y entre**
- Regresión: **error cuadrático medio**
- Clasificación supervisada: **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

Knowledge Discovery from Databases (KDD)



Índice

- 1 Introducción a la minería de datos
- 2 Paradigmas de minería de datos**
- 3 Campos de aplicación de la minería de datos
- 4 Conclusiones

Paradigmas de clasificación supervisada

Estadísticos y de aprendizaje automático

- Árboles de clasificación (Quinlan, 1986; Breiman y col. 1984)
- Clasificadores k -NN (Covert y Hart, 1967; Dasarathy, 1991)
- Regresión logística (Hosmer y Lemeshow, 1989)
- Redes Bayesianas (Pearl, 1988)
- Sistemas clasificadores (Holland, 1975)
- Redes neuronales (McCulloch y Pitts, 1943)
- Inducción de reglas (Clark y Nibblet, 1989; Cohen, 1995; Holte, 1993)
- Máquinas de soporte vectorial (Cristianini y Shawe-Taylor, 2000)
- Análisis discriminante (Fisher, 1936)

Árboles de clasificación

- **Árbol de decisión**
- **Particionamiento recursivo**
- Puede expresarse en forma de reglas:
IF ... THEN ...
- **Cercano** a la manera en que los **humanos** estructuramos un dominio
- **Alta transparencia**

Árboles de clasificación

Ejemplo con dos variables predictoras

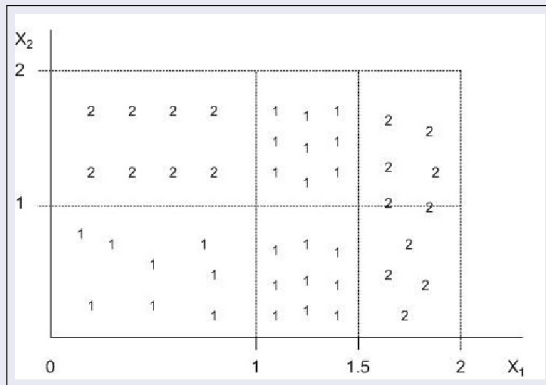


Figura: X_1 y X_2 denotan las variables predictoras. La variable clase C tiene dos posibles valores

Árboles de clasificación

De un árbol de clasificación a un conjunto de reglas

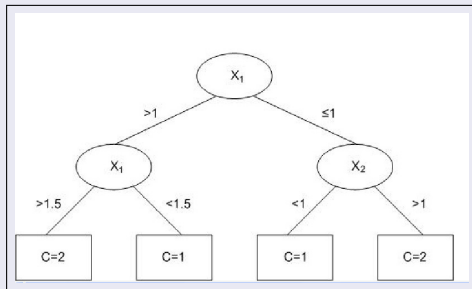


Figura: Árbol de clasificación del ejemplo anterior

Conjunto de reglas equivalentes al árbol de clasificación:

R_1 : If $X_1 > 1,5$ then $C = 2$

R_2 : If $1 < X_1 < 1,5$ then $C = 1$

R_3 : If $X_1 < 1$ and $X_2 < 1$ then $C = 1$

R_4 : If $X_1 < 1$ and $X_2 > 1$ then $C = 2$

K-NN

K-NN \equiv IBL, CBR, lazy learning

- Un nuevo ejemplo se clasifica como **la clase mas frecuente en sus K vecinos mas cercanos**
- Idea muy **simple e intuitiva**
- **Fácil de implementar**
- No hay modelo explícito: **transducción**
- **K-NN** \equiv instance based learning (**IBL**), case based reasoning (**CBR**), **lazy learning**

K-NN

Ejemplo

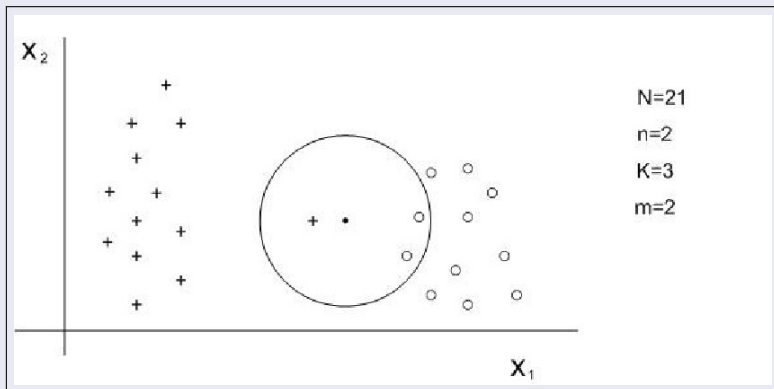


Figura: Ejemplo de un 3-NN. m denota el número de clases, n el número de variables predictoras, y N el número de casos etiquetados

Regresión logística

Variable clase binaria

- Modelo discriminativo:

$$P(C = 1|\mathbf{x}) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^n \beta_i x_i)}}$$

- $\beta_0, \beta_1, \dots, \beta_n$ parámetros estimados por máxima verosimilitud a partir de los datos
- Si C es binaria:

$$\begin{aligned} P(C = 0|\mathbf{x}) &= 1 - \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^n \beta_i x_i)}} \\ &= \frac{e^{-(\beta_0 + \sum_{i=1}^n \beta_i x_i)}}{1 + e^{-(\beta_0 + \sum_{i=1}^n \beta_i x_i)}} \end{aligned}$$

Redes Bayesianas

Naïve Bayes

Variables predictoras condicionalmente independientes dada C

$$c^* = \arg \max_c P(C = c) \prod_{i=1}^n P(X_i = x_i | C = c)$$

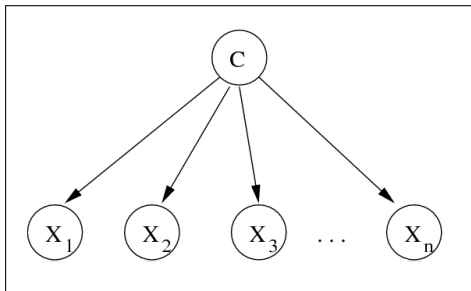


Figura: Estructura de un modelo naïve Bayes

Redes Bayesianas

k-DB

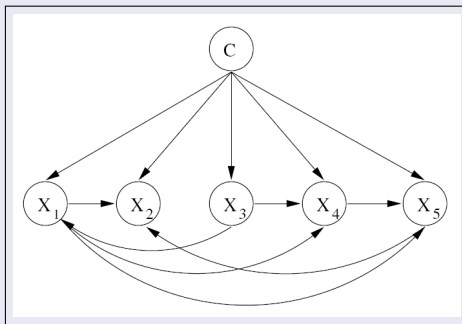


Figura: Ejemplo de *k*-DB para $k = 2$

$$P(c|x_1, x_2, x_3, x_4, x_5) \propto$$

$$P(c)P(x_1|x_3, c)P(x_2|x_1, x_5, c)P(x_3|c)P(x_4|x_1, x_3, c)P(x_5|x_1, x_4, c)$$

Red neuronal

Perceptron multicapa

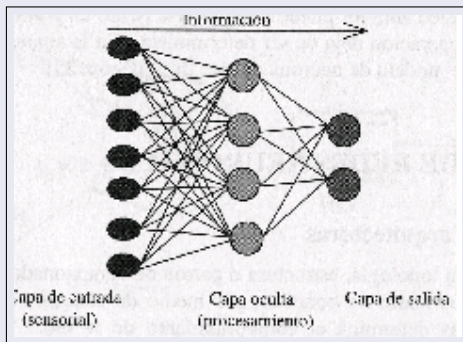


Figura: Arquitectura unidireccional con tres capas de neuronas: una capa de entrada, una capa oculta y una capa de salida

Combinación de clasificadores

- *Non free-lunch theorem* (Wolpert and MacReady, 1996)
 - **Hibridación**: el clasificador final se induce teniendo en cuenta dos o mas paradigmas
 - Combinación de dos o mas clasificadores que tengan el **mismo clasificador de base**: bagging, boosting
 - Combinación de la decisión de dos o mas clasificadores con **diferente clasificador de base**: fusión de etiquetas, fusión de rankings, fusión de salidas continuas, combinación en cascada, *stacked*

Midiendo las prestaciones de un modelo de clasificación supervisada

Criterios de comparación

- Medidas cuantitativas
 - Accuracy
 - Área bajo la curva ROC
- Medidas cualitativas
 - Complejidad del inductor
 - Transparencia del modelo
 - Simplicidad del modelo
 - Comprensión del modelo
 - ...

Midiendo las prestaciones de un modelo de clasificación supervisada

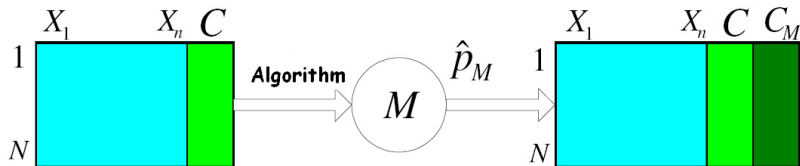
Matriz de confusión

		C Clase verdadera	
		+	-
C_M Clase predicha	+	a	b
	-	c	d

Criterios

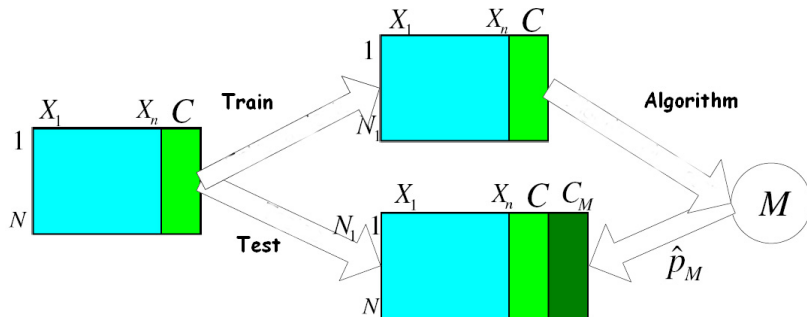
- **Accuracy:** $\frac{a+d}{a+b+c+d}$
- **Porcentaje de error:** $\frac{c+b}{a+b+c+d}$
- Ratio de verdadero positivos (**sensitivity**): $\frac{a}{a+c}$
- Ratio de verdaderos negativos (**specificity**): $\frac{d}{b+d}$

Métodos de estimación. No honesto



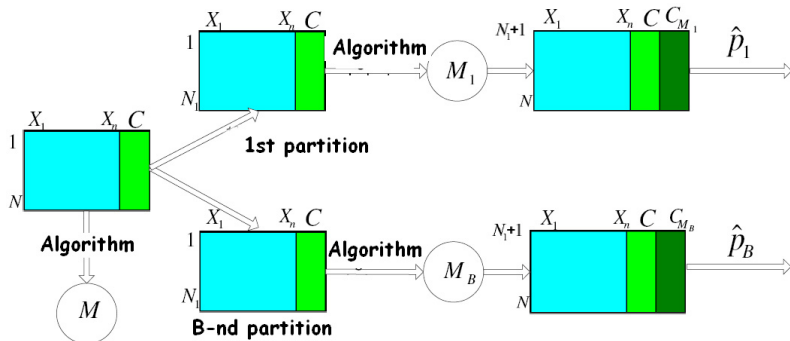
$$\hat{p}_M = \frac{1}{N} \sum_{i=1}^N \delta(c^{(i)} = c_M^{(i)})$$

Métodos de estimación. Entrenamiento y testeo



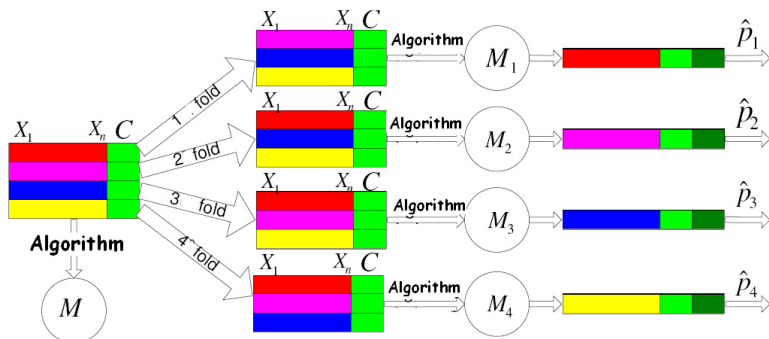
$$\hat{p}_M = \frac{1}{N - N_1} \sum_{i=1}^{N - N_1} \delta(c^{(N_1+i)} = c_M^{(N_1+i)})$$

Métodos de estimación. Entrenamiento y testeo repetidas veces



$$\hat{p}_M = \frac{1}{B} \sum_{i=1}^B \hat{p}_i$$

Métodos de estimación. k -fold cross validation



$$\hat{p}_M = \frac{1}{k} \sum_{i=1}^k \hat{p}_i$$

Clasificación sensible al costo

Coste total vs accuracy

- En general el **coste** de los falsos positivos y los falsos negativos **no es el mismo**
- Interesa obtener el clasificador proporcionando el **menor coste total**
- Este clasificador puede ser diferente del que tenga un mayor accuracy (una mayor probabilidad de clasificar correctamente casos nuevos)

Clasificación sensible al costo

Confusion matrices

		C	
		0	1
CM	0	300	500
	1	200	99000

		C	
		0	1
CM	0	0	0
	1	500	99500

		C	
		0	1
CM	0	400	5400
	1	100	94100

		C	
		0	1
CM	0	0 €	100 €
	1	2000 €	0 €

Cost matrix

		C	
		0	1
CM	0	0 €	50000 €
	1	400000 €	0 €

		C	
		0	1
CM	0	0 €	0 €
	1	1000000	0 €

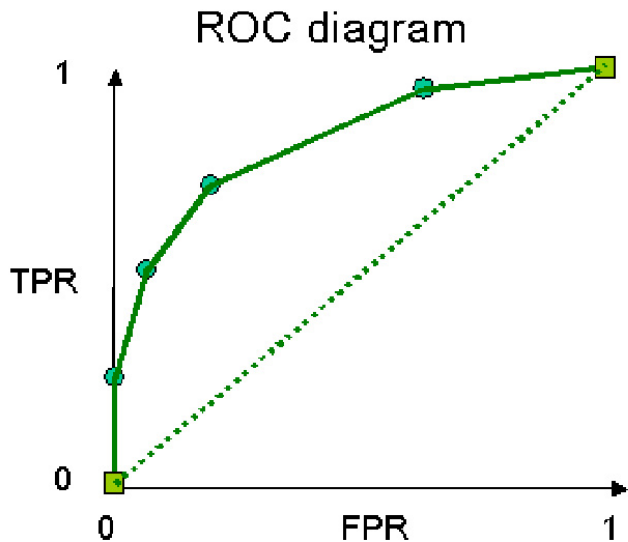
		C	
		0	1
CM	0	0 €	540000 €
	1	200000 €	0 €

Total cost: 450000 €

Total cost: 1000000 €

Total cost: 740000 €

Curva ROC



Paradigmas de clasificación no supervisada

Clustering particional. Evolución de K -medias

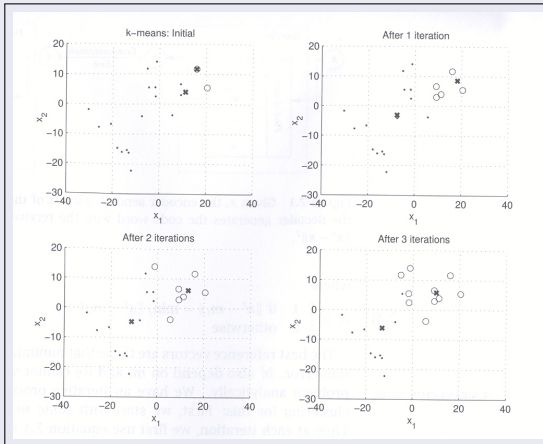


Figure: Cruces indican posiciones de los centroides. Los puntos representando los datos están marcados dependiendo de su centroide mas cercano

Índice

- 1 Introducción a la minería de datos
- 2 Paradigmas de minería de datos
- 3 Campos de aplicación de la minería de datos**
- 4 Conclusiones

Financieras

Tarjetas de crédito



- Detección de **uso fraudulento** de tarjetas de crédito
- **Predicción del gasto** en tarjeta de crédito por grupos

Financieras

Concesión de crédito



- Análisis de riesgos en concesión de créditos

Financieras

Fidelización de clientes



- Detección de **clientes con riesgo de abandonar** la entidad financiera

Comercio

Análisis de la cesta de la compra



- Optimizar la disposición de los productos a partir de reglas obtenidas al **analizar la cesta de la compra**

Comercio

Segmentación de clientes



- Campañas de **marketing** dirigidas a cada **segmento** poblacional

Compañías de seguros

Clientes

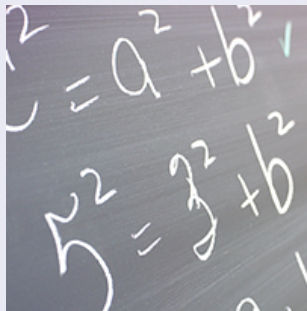
LOS SEGUROS
MAS COMPLETOS
PARA TI Y
TU FAMILIA



- Determinación de clientes **potencialmente caros**
- Predicción de que tipo de clientes **contratan nuevas pólizas**
- Identificación de **comportamiento fraudulento**

Educación

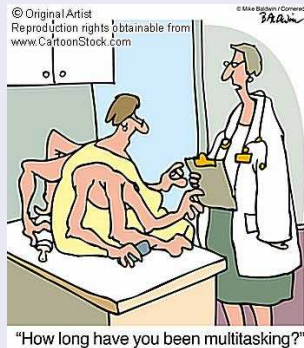
Educación



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

Medicina

Diagnóstico de enfermedades



- **Diagnóstico de enfermedades** a partir del historial, datos clínicos del paciente y test diagnósticos

Medicina

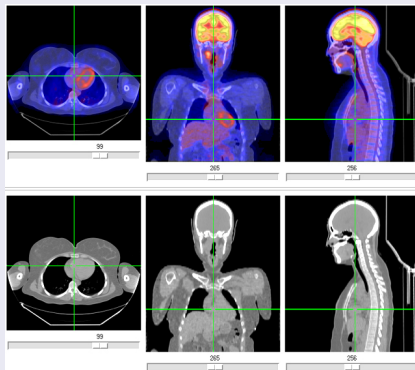
Optimización de recursos



- **Gestión** hospitalaria y asistencial
- Predicciones temporales de los centros sanitarios para el **mejor uso de recursos**

Medicina

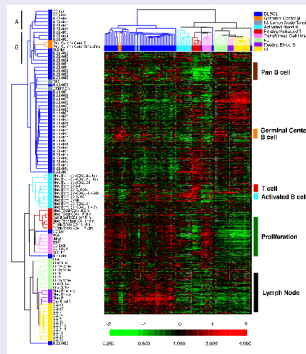
Imágenes médicas



- Diagnóstico automático (ayuda al diagnóstico) basado en **imagen**

Bioinformática

Microarrays de ADN



- Búsqueda de **biomarcadores** a partir de datos de microarrays

Bioinformática

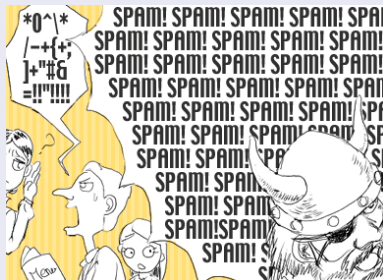
Predicción de la estructura secundaria de las proteínas



- La **estructura secundaria** de las proteínas a partir de la cadena de aminoácidos

Computación

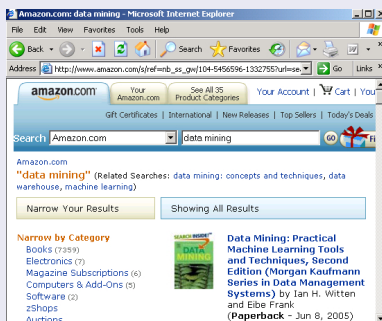
Basura en correo electrónico



- **Clasificar los e-mails** en función de si son de interés o no

Web mining

Filtrado colaborativo



- Si usted está interesado en estos libros, **probablemente también le interesen** estos otros

Música

Exito de canciones



- **Predecir el éxito** de una determinada canción (o un grupo) en un determinado mercado (nación)

Deporte

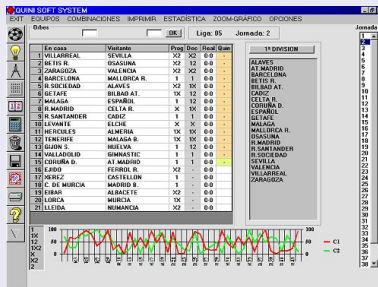
Predicción de lesiones deportivas



- **Predicir lesiones** deportivas en base a parámetros del jugador

Deporte

Quinielas



- **Predecir** los resultados de un partido de fútbol (1-X-2) en base a variables de cada uno de los equipos

Índice

- 1 Introducción a la minería de datos
- 2 Paradigmas de minería de datos
- 3 Campos de aplicación de la minería de datos
- 4 Conclusiones**

Conclusiones

Minería de Datos

- Accesibilidad creciente a **grandes volúmenes de datos**
- Transformar **datos en conocimiento** de manera automática
- **Decisiones mas racionales** basadas en modelos computacionales

Minería de Datos

I. Introducción

Pedro Larrañaga, Iñaki Inza, Abdelmalik Moujahid

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

MMCC, 2006-2007