INTRODUCCIÓN A LA MINERÍA DE DATOS

Dr. David F. Barrero, Universidad de Alcalá

Universidad de Alcalá

Índice

- MINERÍA DE DATOS
  - APLICACIONES
  - INTRODUCCIÓN A LA MINERÍA DE DATOS
  - TAREAS EN MINERÍA DE DATOS
    - PREDICCIÓN (CLASIFICACIÓN / REGRESIÓN)
    - AGRUPAMIENTO
    - CLUSTERING
  - FASES EN MINERÍA DE DATOS
    - INTEGRACIÓN Y RECOPILACIÓN DE DATOS
    - SELECCIÓN, LIMPIEZA Y TRANSFORMACIÓN
    - MINERÍA DE DATOS
    - EVALUACIÓN DEL CONOCIMIENTO MINADO
  - PRÁCTICAS

AGRADECIMIENTO AL DR. RICARDO ALER SIERRA Y AL DR. DANIEL RODRÍGUEZ POR FACILITAR LAS DIAPOSITIVAS ORIGINALES

DIAPOSITIVAS DISPONIBLES EN

http://atc1.aut.uah.es/~david/mineriaDatosSimple.pdf

Minería de Datos
Aplicaciones (I)

- Financieras y banca
  - Obtención de patrones de fraude de tarjetas de crédito
  - Predicción de devolución de créditos
- Análisis de mercado
  - Análisis de cesta de la compra
  - Análisis de fidelidad de clientes. Reducción de fuga
  - Segmentación de clientes
- Seguros y salud privada: Determinación de clientes potencialmente caros
- Educación: Detección de abandonos
- Industria: Predicción de la demanda eléctrica, de gas, etc.

Minería de Datos
Aplicaciones (II)

- Medicina
  - Diagnóstico de enfermedades
  - Predecir si un compuesto químico causa cáncer
  - Predecir si una persona puede tener potencialmente una enfermedad a partir de su ADN
- Ciencia
  - Interfaces Cerebro-Máquina
  - Análisis de secuencias de proteínas
  - Clasificación de cuerpos celestes (SKYCAT)
- Internet
  - Detección de spam (*SpamAssassin*, bayesiano)
  - Web: Asociar libros que compran usuarios

Introducción a la Minería de Datos

Justificación

- **Nuevas posibilidades**
  - **D**isponibilidad de grandes cantidades de datos (bancos, la web, tarjetas fidelización, DNA, ...) Y potencia de cómputo
- **Nuevas necesidades**
  - Es complicado analizar los datos de manera manual. Necesidad de técnicas automáticas
- **Objetivo**
  - Convertir datos en conocimiento

- MD = BBDD + estadística + aprendizaje automático

Tareas en Minería de Datos

Resumen

- Predicción
  - Clasificación
  - Regresión
- Asociación
- Agrupación (*clustering*)

Tareas en Minería de Datos
Predicción (I)

- También conocido como aprendizaje supervisado
  - El algoritmo se entrena con datos y la salida deseada
  - Tarea: Predecir la salida de nuevos datos
    - Categórica -> Clasificación
    - Numérica -> Regresión
  - Ejemplos de clases:
    - Persona sana o enferma
    - Crédito dudoso o no dudoso
    - Un patrón en el EEG
    - El alfabeto, cada letra es una clase

Tareas en Minería de Datos

Predicción (II)

- Formato de la tabla
  - Columnas, llamadas atributos
    - Pueden ser categóricos o numéricos
  - Filas, llamadas instancias

A_1	...	A_n	C
$a_{1,1}$	...	$a_{1,n}$	c_1
...	...	...	...
$a_{m,1}$	...	$a_{m,1}$	c_m

Tareas en Minería de Datos

Ejemplo de clasificación: Créditos bancarios (I)

- Un banco por Internet desea obtener reglas para predecir qué personas de las que solicitan un crédito no van a devolverlo
- La entidad bancaria cuenta con una gran base de datos correspondientes a los créditos concedidos (o no) a otros clientes con anterioridad
- Dos clases
  - Crédito seguro y crédito dudoso

Tareas en Minería de Datos

Ejemplo de clasificación: Créditos bancarios (II)

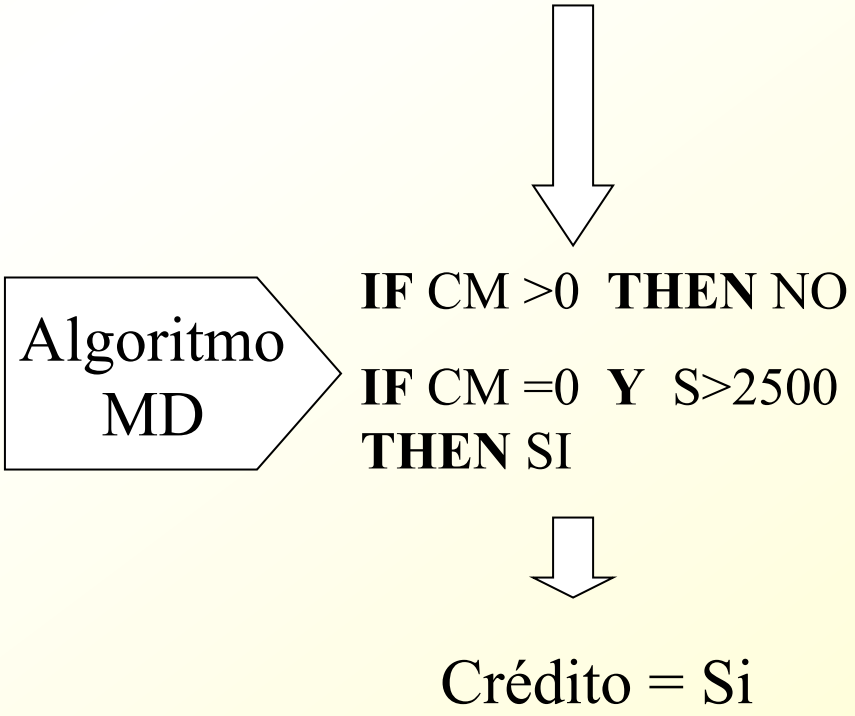
IDC	Años	Euros	Salario	Casa propia	Cuentas morosas	...	Devuelve el crédito
101	15	60000	2200	Si	2	...	No
102	2	30000	3500	Si	0	...	Si
103	9	9000	1700	Si	1	...	No
104	15	18000	1900	No	0	...	Si
105	10	24000	2100	No	0	...	No
...	...	...	...	...	...		...

Tareas en Minería de Datos

Ejemplo de clasificación: Créditos bancarios (III)

Años	Euros	Salario	Casa propia	Cuentas morosas	Crédito
10	50000	3000	Si	0	??

Años	Euros	Salario	Casa propia	Cuentas morosas	Crédito
15	60000	2200	Si	2	No
2	30000	3500	Si	0	Si
9	9000	1700	Si	1	No
15	18000	1900	No	0	Si
10	24000	2100	No	0	No
	…	…	…	…	…



Tareas en Minería de Datos

Ejemplo de clasificación: Créditos bancarios (IV)

- Conocimiento obtenido:
  - **SI** (cuentas-morosas > 0) **ENTONCES** Devuelve-crédito = no
  - **SI** (cuentas-morosas = 0) **Y** ((salario > 2500) **O** (años > 10)) **ENTONCES** devuelve-crédito = si
- Se obtiene un modelo de moroso
  - Analizable por expertos en banca

Tareas en Minería de Datos

Ejemplo de regresión: Ventas (I)

- Situación
  - Una cadena de tiendas desea optimizar el funcionamiento de su almacén manteniendo un stock de cada producto suficiente
- Problema
  - Predecir cuánto tiempo tarda en venderse un producto

Tareas en Minería de Datos

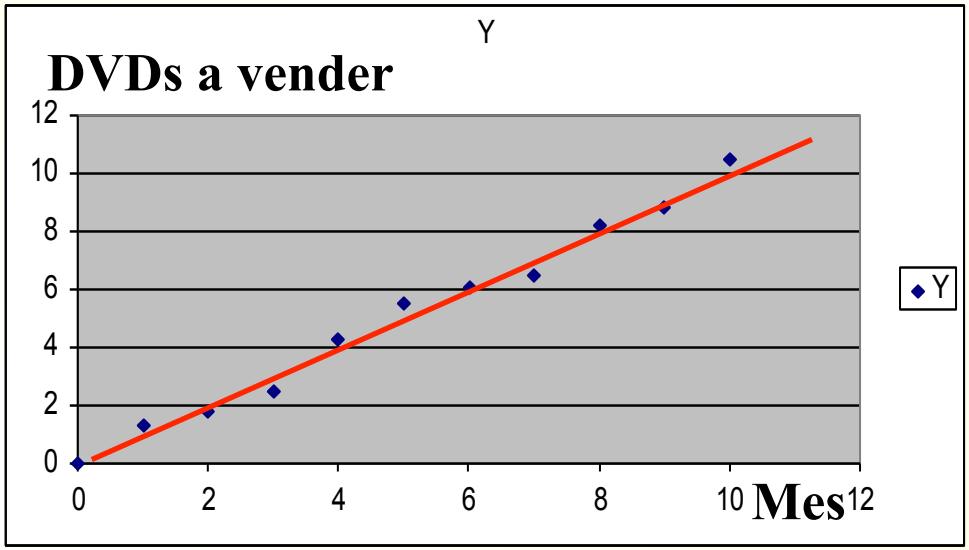
Ejemplo de regresión: Ventas (II)

Producto	Mes-12	...	Mes-4	Mes-3	Mes-2	Mes-1
Televisor plano	20	...	52	14	139	74
Video	11	...	43	32	26	59
Nevera	50	...	61	14	5	28
Microondas	3	...	21	27	1	49
Discman	14	...	27	2	25	12
...	...	...	...	...	...	...

Tareas en Minería de Datos

Ejemplo de regresión: Ventas (III)

- Conocimiento adquirido
  - Modelo que prediga lo que se va a vender cada mes a partir de lo que se vendió en los meses anteriores (serie temporal)



Tareas en Minería de Datos

- Predicción
  - Clasificación
  - Regresión
- **Asociación**
- Agrupación (*clustering*)

Tareas en Minería de Datos

Ejemplo de asociación: Cesta de la compra (I)

- Situación
  - Un supermercado quiere obtener información sobre el comportamiento de compra de sus clientes
- Objetivo
  - Identificar productos complementarios
  - Mejorar el servicio, colocando ciertos productos juntos, por ejemplo

Tareas en Minería de Datos

Ejemplo de asociación: Cesta de la compra (II)

Id	Huevos	Aceite	Pañales	Vino	Leche	Mantequilla	Salmón	Lechugas	...
1	Si	No	No	Si	No	Si	Si	Si	...
2	No	Si	No	No	Si	No	No	Si	...
3	No	No	Si	No	Si	No	No	No	...
4	No	Si	Si	No	Si	No	No	No	...
5	Si	Si	No	No	No	Si	No	Si	...
6	Si	No	No	Si	Si	Si	Si	No	...
7	No	No	No	No	No	No	No	No	...
8	Si	Si	Si	Si	Si	Si	Si	No	...
...	...	...	...	...	...	...	...	...	...

Tareas en Minería de Datos

Ejemplo de asociación: Cesta de la compra (III)

- Conocimiento adquirido
- Reglas **Si** At_1=a y At_2=b y ... **Entonces** At_n=c
  - **Si** pañales=si, **entonces** leche=si (100%, 37%)
  - **Si** huevos=si, **entonces** aceite=si (50%, 25%)
  - **Si** vino=si, **entonces** lechugas=si (33%, 12%)
- Las reglas también pueden ser:
  - **Si** At_1=a y At_2=b **Entonces** At_n=c, **At_4=D**
- (a,b) = (precisión, cobertura)
  - Precisión: veces que la regla es correcta
  - Cobertura: frecuencia de ocurrencia de la regla en los datos

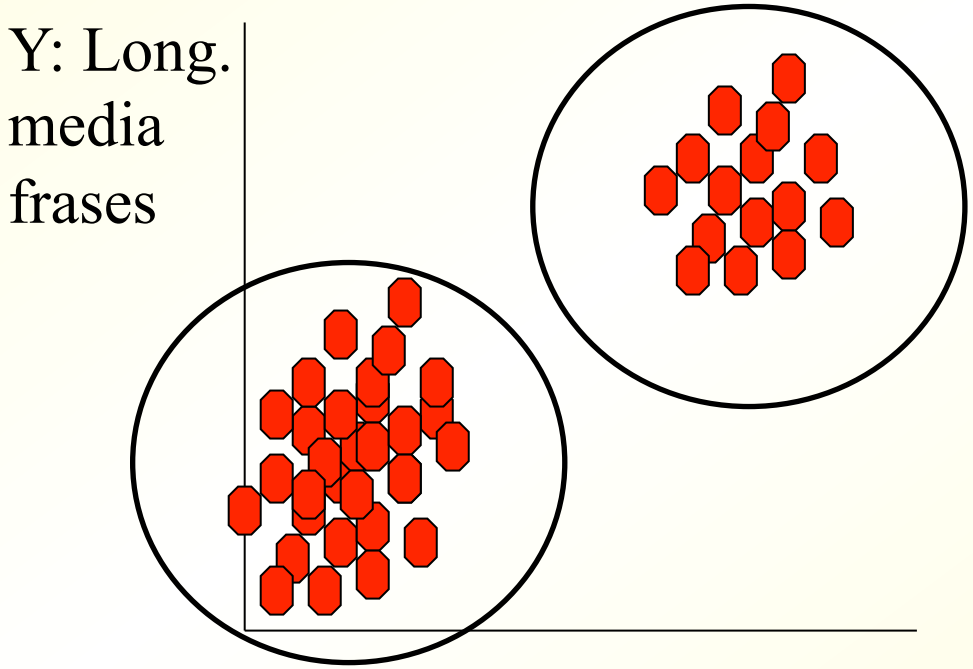
Minería de Datos. Tareas

- Predicción
  - Clasificación
  - Regresión
- Asociación
- **Agrupación** (*clustering*)

Tareas en Minería de Datos

Idea del clustering

- Detectar agrupaciones naturales en los datos
  - Agrupación (o "clustering") = aprendizaje no supervisado: se parte de una tabla, como en clasificación, pero sin la clase



Ejemplo: clustering de libros. 2 grupos

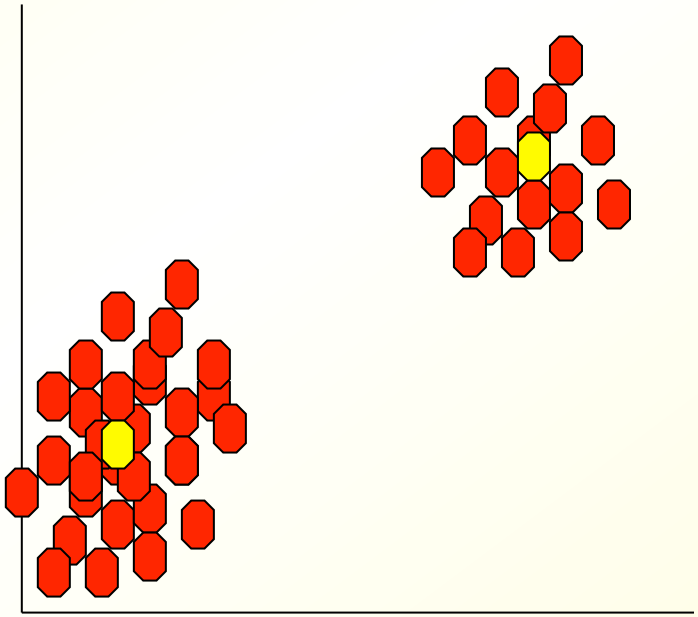
* Palabras y frases largas (¿filosofía?)

* Palabras y frases cortas (¿novela?)

Tareas en Minería de Datos

Representación de los clusters

- Por sus centroides (ej: algoritmo k-medias)
  - La pertenencia a un cluster puede ser probabilística (ej: algoritmo EM)



Tareas en Minería de Datos

Ejemplo de clustering: RRHH (I)

- Situación
  - El departamento de RRHH de una empresa desea categorizar a sus empleados en distintos grupos
  - El objetivo es entender mejor su comportamiento para tratarlos de manera adecuada (p.e., incentivos)
- Problema
  - No existen categorías *a priori*
  - Aprender las categorías es parte de la tarea

Tareas en Minería de Datos

Ejemplo de clustering: RRHH (II)

Id	Sueldo	Casado	Coche	Hijos	Alq/Prop	Sindicado	Bajas	Antigüedad	Sexo
1	1000	Si	No	0	Alq	No	7	15	H
2	2000	No	Si	1	Alq	Si	3	3	M
3	1500	Si	Si	2	Prop	Si	5	10	H
4	3000	Si	Si	1	Alq	No	15	7	M
5	1000	Si	Si	0	Prop	Si	1	6	H
...	...	...	...	...		...	...	...	...

Tareas en Minería de Datos

Ejemplo de clustering: RRHH (III)

- Conocimiento obtenido

	GRUPO 1	GRUPO 2	GRUPO 3
Sueldo	1535	1428	1233
Casado (No/Si)	77%/22%	98%/2%	0%/100%
Coche	82%/18%	1%/99%	5%/95%
Hijos	0.05	0.3	2.3
Alq/Prop	99%/1%	75%/25%	17%/83%
Sindicado	80%/20%	0%/100%	67%/33%
Bajas	8.3	2.3	5.1
Antigüedad	8.7	8	8.1
Sexo (H/M)	61%/39%	25%/75%	83%/17%

Tareas en Minería de Datos

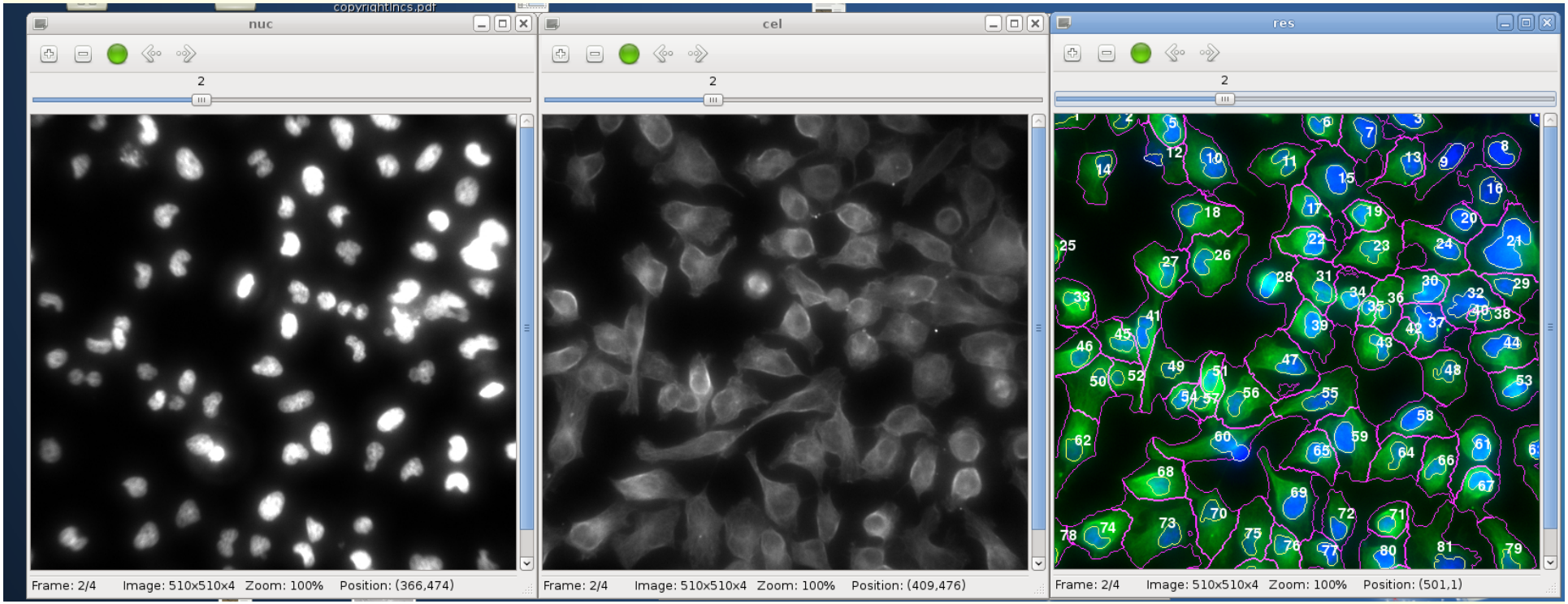
Ejemplo de clustering: RRHH (IV)

- Grupo 1
  - Sin hijos y con vivienda de alquiler. Poco sindicados. Muchas bajas
- Grupo 2
  - Sin hijos y con coche. Muy sindicados. Pocas bajas. Normalmente son mujeres y viven en alquiler
- Grupo 3
  - Con hijos, casados y con coche. Mayoritariamente hombres propietarios. Poco sindicados.

Tareas en Minería de Datos

Ejemplo de clustering: Visión artificial

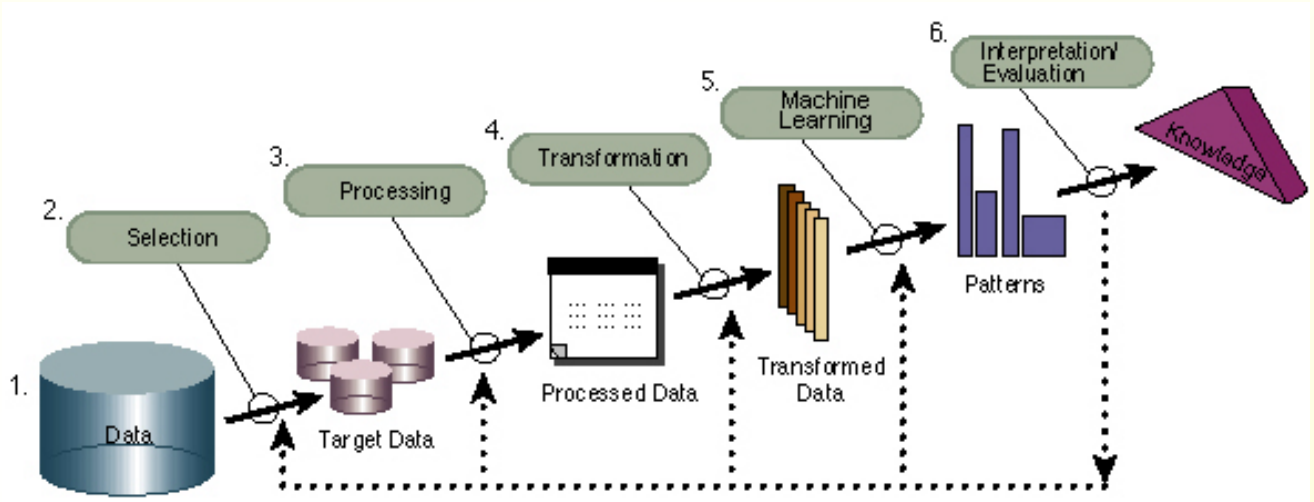
- Contar número de células



Fases de la Minería de Datos

Vistazo general

1. Integración y recopilación de datos
2. Selección, limpieza y transformación -> Datos
3. **Minería de datos -> Patrones (ej: clasificador)**
4. Evaluación e interpretación -> Conocimiento
5. Difusión y uso -> Decisiones

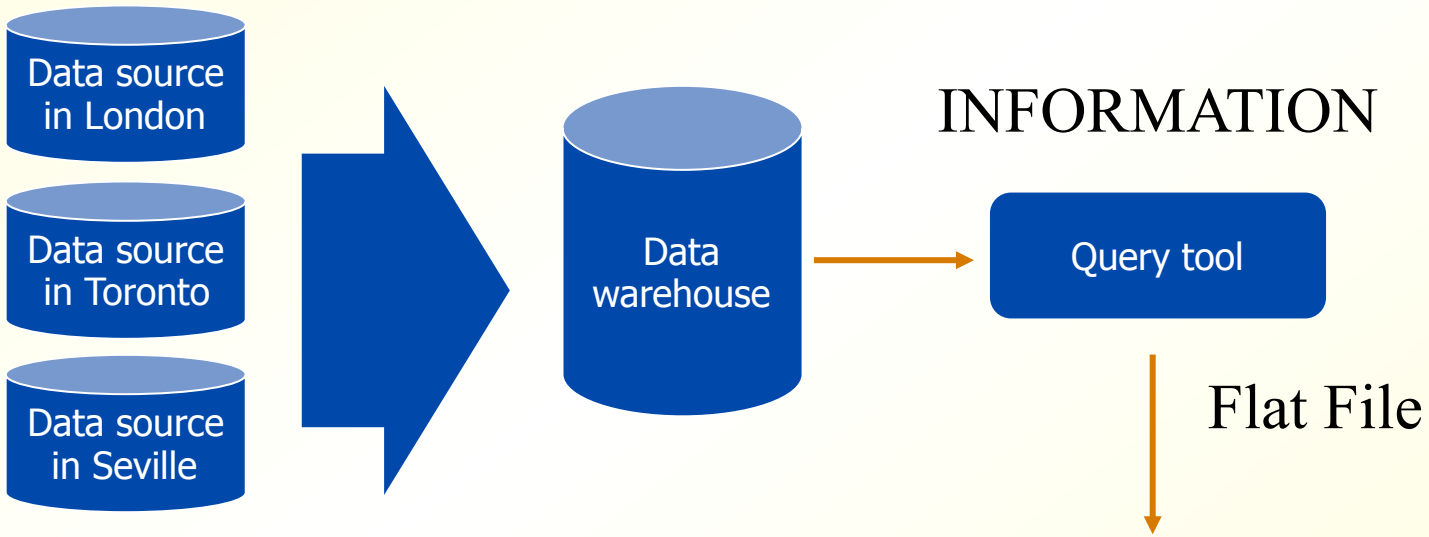


An Overview of the Steps That Compose the KDD Process

Fases de la Minería de Datos

Integración y recopilación (I)

- Se hace minería sobre tablas simples
  - Construcción de la tabla: Datawarehouse, etc



customer

Cust-ID	name	address	age	income	credit-info	.
C1	Smith, Sandy	5463 E Hasting, Burnaby BC V5A 459, Canada	21	$27000	1	…

Fases de la Minería de Datos

Preproceso: Selección, limpieza y transformación

- Instancias
  - O*utliers*: Eliminar o dejar
  - Muestreo de datos (si hay muchos)
- Atributos
  - Valores faltantes (*missing values*)
  - Eliminar atributos redundantes o irrelevantes (ej: sueldo y clase social)
  - Calcular nuevos atributos que sean más relevantes (área, población -> densidad de población)
  - Discretización, numerización, normalización, …

Fases de la Minería de Datos

Minería de Datos: Aprendizaje supervisado

- Árboles de decisión
  - ID3, C4.5 (J48), ...
- Árboles de regresión
  - LMT (M5), ...
- Reglas
  - PART, CN2, AQ, ...
- Funciones
  - Redes neuronales, regresión, SVMs, ...
- Técnicas perezosas
  - IB1, IBK, ...
- Técnicas Bayesianas
- Metatécnicas

A_1	...	A_n	C
$a_{1,1}$	...	$a_{1,n}$	c_1
...	...	...	...
$a_{m,1}$	...	$a_{m,1}$	c_m

Fases de la Minería de Datos

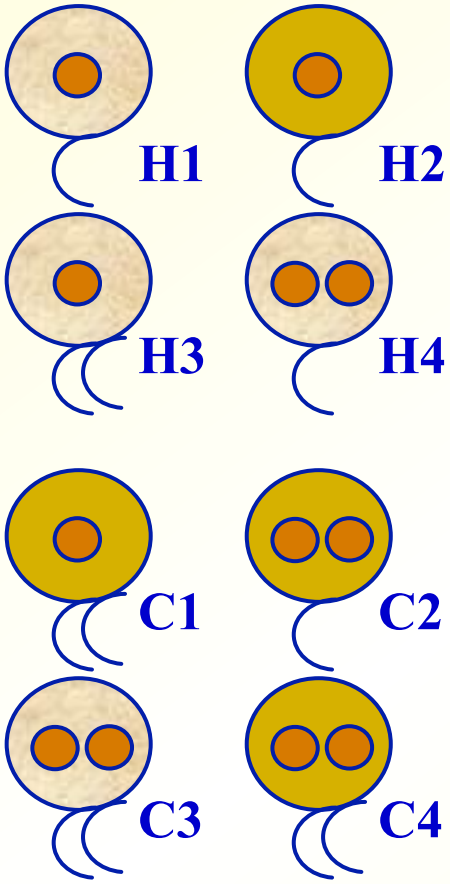
Minería de Datos: Aprendizaje no supervisado

- Asociación
  - Association Rules, A PRIORI
- *Clustering*
  - *Clustering* jerárquico
  - Algoritmos: K-Means, EM
- No hay una clase definida

A_1	...	A_n
$a_{1,1}$	...	$a_{1,n}$
...	...	...
$a_{m,1}$		$a_{m,1}$

Fases de la Minería de Datos

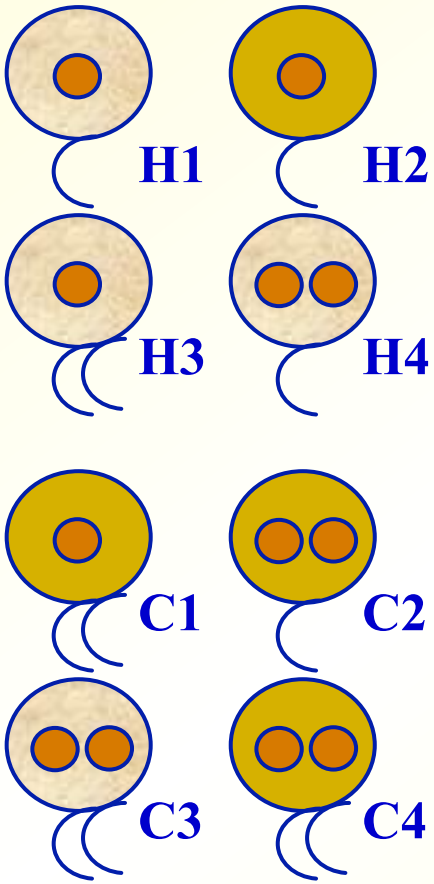
Minería de Datos: Ejemplos



	colour	#nuclei	#tails	class
H1	light	1	1	healthy
H2	dark	1	1	healthy
H3	light	1	2	healthy
H4	light	2	1	healthy
C1	dark	1	2	cancerous
C2	dark	2	1	cancerous
C3	light	2	2	cancerous
C4	dark	2	2	cancerous

Fases de la Minería de Datos

Minería de Datos: Reglas de decisión



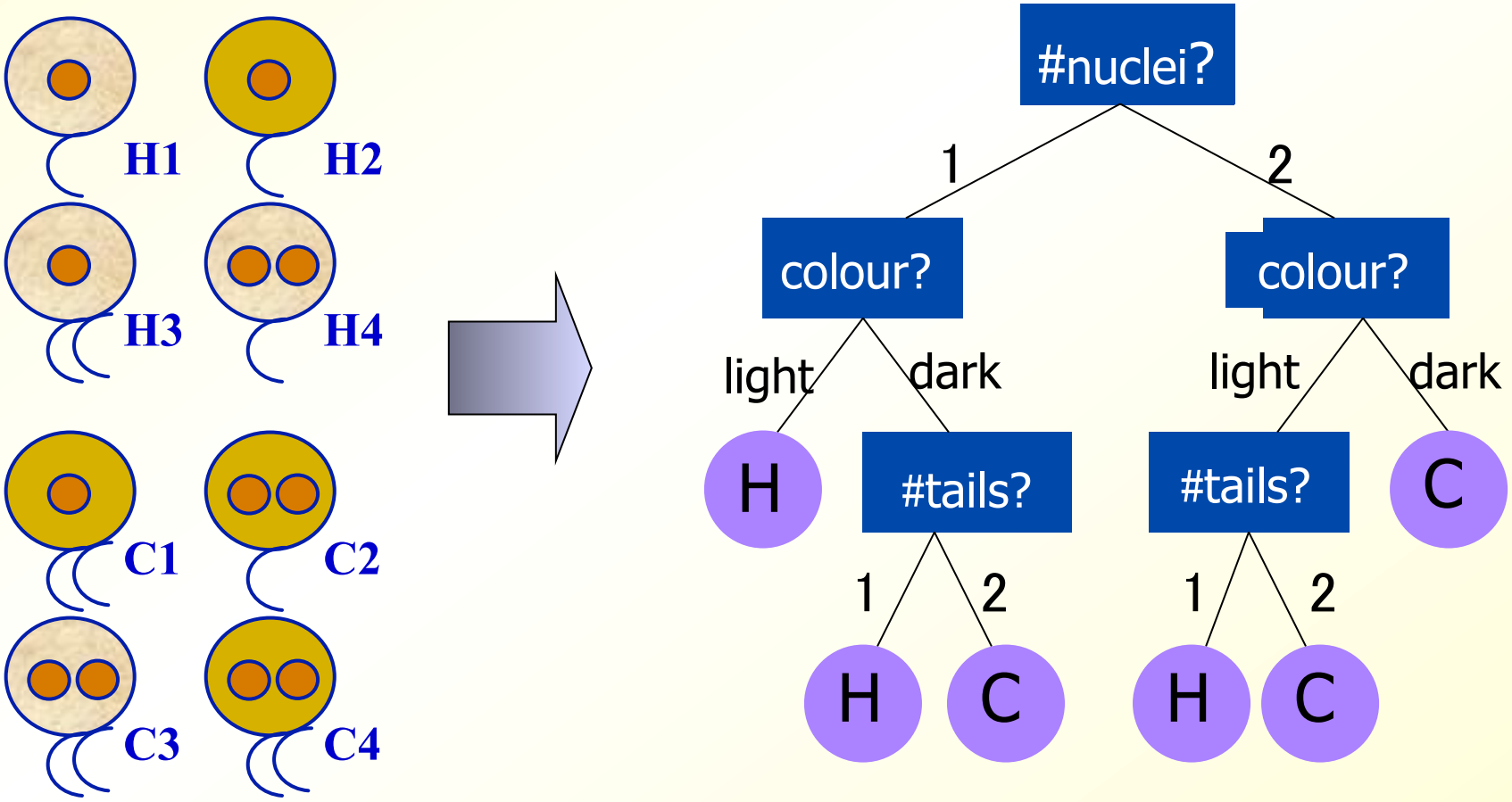
If colour = light **and** # nuclei = 1
Then cell = healthy

If #nuclei = 2 **and** colour = dark
Then cell = cancerours

(and 4 rules more)

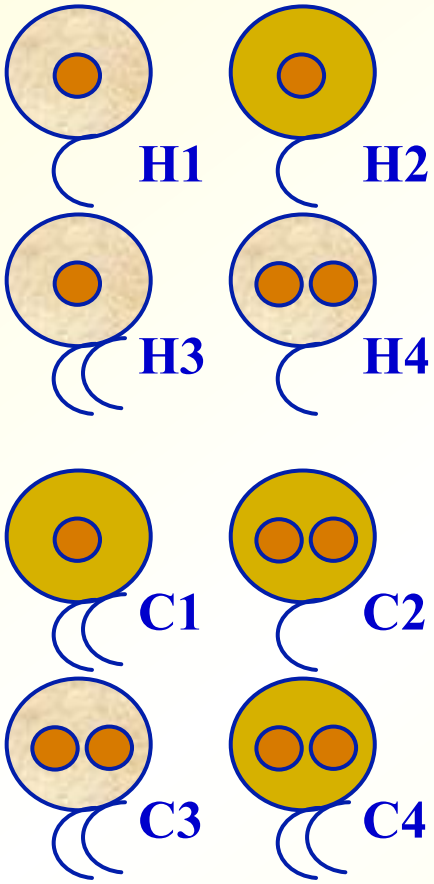
Fases de la Minería de Datos

Minería de Datos: Árboles de decisión



Fases de la Minería de Datos

Minería de Datos: Reglas de decisión jerárquicas



If colour = light **and** # nuclei = 1
Then cell = healthy

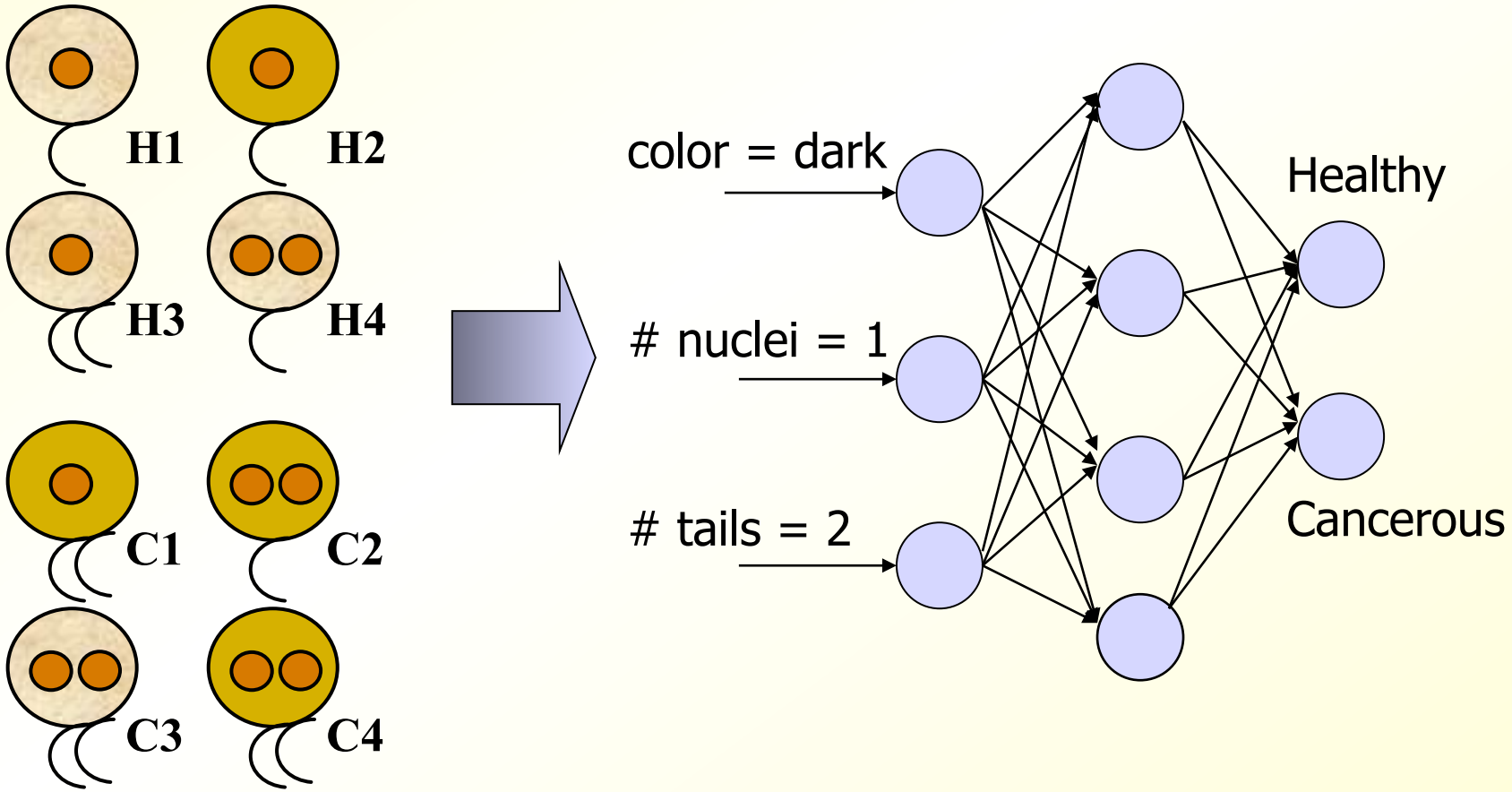
Else
  If #nuclei = 2 **and** colour = dark
  Then cell = cancerous

  Else
    If #tails = 1
    Then cell = healthy

    Else cell = cancerous

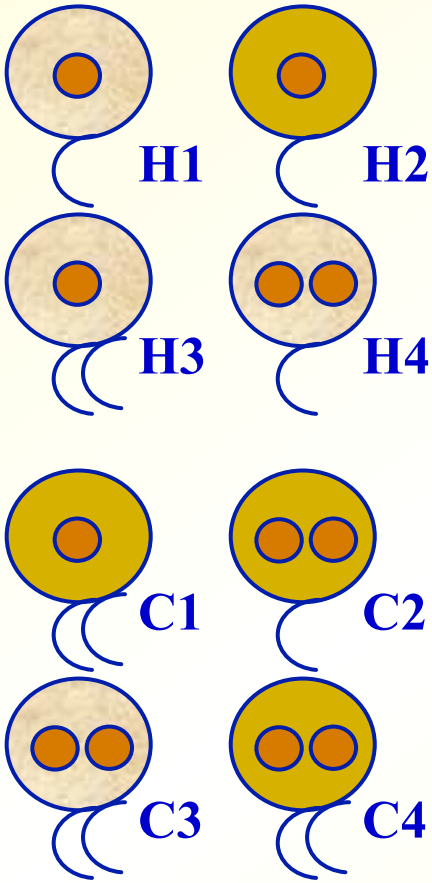
Fases de la Minería de Datos

Minería de Datos: Redes neuronales



Fases de la Minería de Datos

Minería de Datos: Reglas de asociación



If colour = light
and # nuclei = 1
Then # tails = 1
(support = 25%; confidence = 50%)

association among A and B means that the presence of A in a record implies the presence of B in the same record

If # nuclei = 2
and cell = cancerous
Then # tails = 2
(support = 3/8%; confidence = 2/3%)

Support: the proportion of times that the rule applies.
Confidence: the proportion of times that the rule is correct

Apriori algorithm, Agrawal 1993

Fases de la Minería de Datos

Evaluación (I)

- Es necesario validar el conocimiento obtenido
  - Observar su comportamiento con datos no vistos
  - Ejemplo: Si a un alumno se le evalúa (examen) con los mismos problemas con los que aprendió, no se demuestra su capacidad de generalización
- Solución: Dividir el conjunto de datos en un subconjunto para entrenamiento (66%) y otro para test (33%)
- Problema: Es posible que por azar, los datos de entrenamiento y test estén sesgados
  - Ejemplo de sesgo: Sea un problema para determinar qué tipo de personas compran aparatos de DVD. Puede ocurrir por casualidad que en los datos de entrenamiento aparezcan muchas más mujeres que hombres. El sistema creerá que hay una correlación entre el sexo y la clase.

Fases de la Minería de Datos

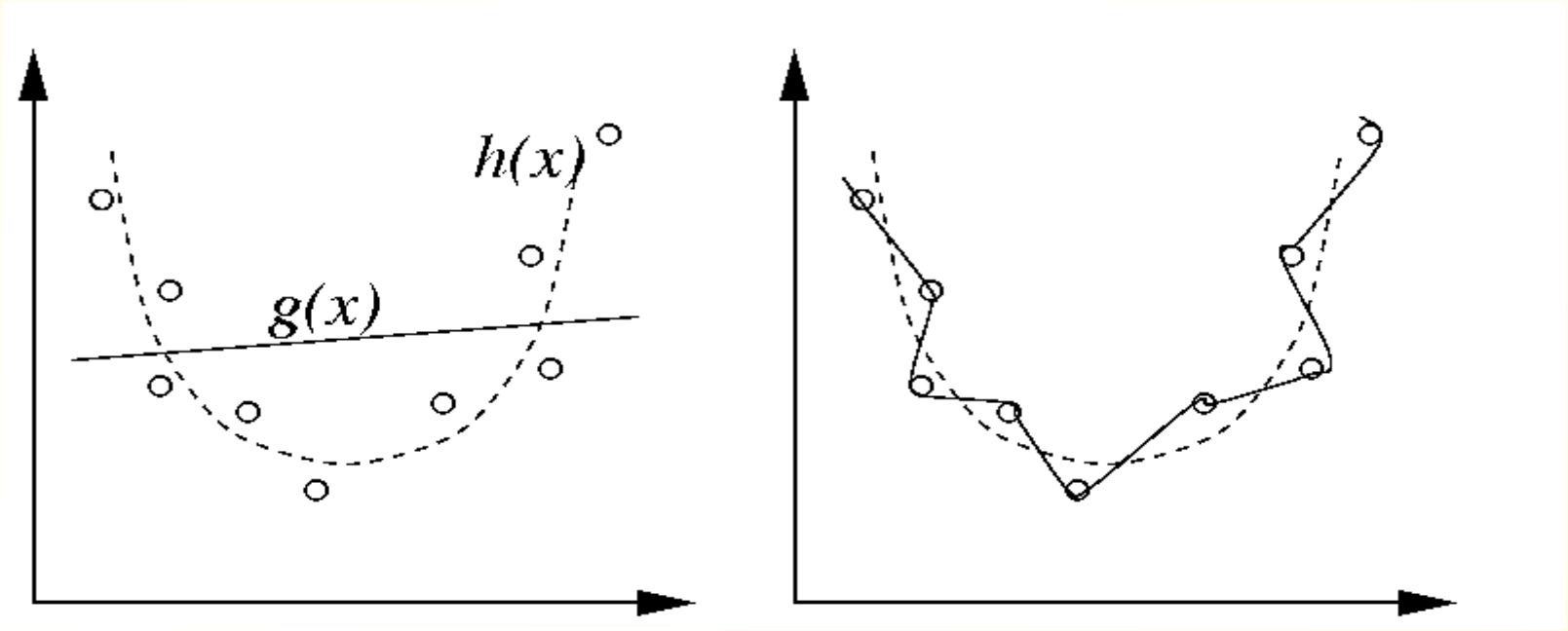
Evaluación (II): Validación cruzada

- Solución:
  - Dividir **varias veces** el mismo conjunto de datos en entrenamiento y test y calcular la media.
- Se divide el conjunto de datos original en k partes. Con k=3 tenemos los subconjuntos A, B, y C.
- Tres iteraciones:
  - Aprender con A, B y test con C (T1 = % aciertos con C)
  - Aprender con A, C y test con B (T2 = % aciertos con B)
  - Aprender con B, C y test con A (T3 = % aciertos con A)
  - % aciertos final T = (T1+T2+T3)/3
- El clasificador final CF se construye **con todos los datos (los tres conjuntos** A, B y C).
  - Se supone que T es una estimación de los aciertos
- Se suele utilizar k=10 (***10*-fold cross validation**)

Fases de la Minería de Datos

Evaluación (III)

- Derecha: El modelo se ha sobreadaptado al ruido porque es demasiado complejo (*overfitting*)
- Izquierda: El modelo lineal **g(x)** es demasiado simple para aproximar una parábola y subadapta los datos
- Conclusión: Tiene que haber un equilibrio en la complejidad del clasificador (o del modelo en general)



Práctica

DATASETS EN

http://atc1.aut.uah.es/~david/titanic.arff

http://atc1.aut.uah.es/~david/Drug1n.arff

http://atc1.aut.uah.es/~david/Employees.arff

Práctica

- Predecir el tipo de fármaco (*drug*) que se debe administrar a un paciente afectado de rinitis alérgica según distintos parámetros/ variables.
- Las variables que se recogen en los historiales clínicos son:
  - Age: Edad
  - Sex: Sexo
  - BP (Blood Pressure): Tensión sanguínea
  - Cholesterol: nivel de colesterol
  - Na: Nivel de sodio en la sangre
  - K: Nivel de potasio en la sangre
- Hay cinco fármacos posibles: DrugA, DrugB, DrugC, DrugX, DrugY.
- Se han recogido los datos del medicamento idóneo para muchos pacientes en cuatro hospitales. Se pretende, para nuevos pacientes, determinar el mejor medicamento a probar
- Fichero "Drug1n.arff" de:
  - http://atc1.aut.uah.es/~david/Drug1n.arff

Práctica

1. Cargar los datos
2. Probar con el algoritmo ZeroR para obtener un resultado base
3. Probar con varios algoritmos de clasificación, con validación cruzada de 10. Intentar obtener el mejor porcentaje de aciertos
4. Probar a quitar atributos, para ver si el porcentaje de aciertos mejora