

Preparación de Datos

Tratamiento Inteligente de Datos
Máster Universitario en Ingeniería Informática



UNIVERSIDAD
DE GRANADA

Gabriel Navarro (gnavarro@ugr.es, gnavarro@decsai.ugr.es)

Objetivos

- ❑ Entender los distintos problemas a resolver en procesos de recopilación y preparación de datos
- ❑ Conocer problemas presentes en la integración de datos de distintas fuentes y técnicas para resolverlos
- ❑ Conocer problemas a resolver para limpiar los datos y algunas técnicas que los resuelven
- ❑ Entender la necesidad, en ocasiones, de aplicar técnicas de transformación de datos
- ❑ Conocer las técnicas de reducción de datos y la necesidad de aplicación

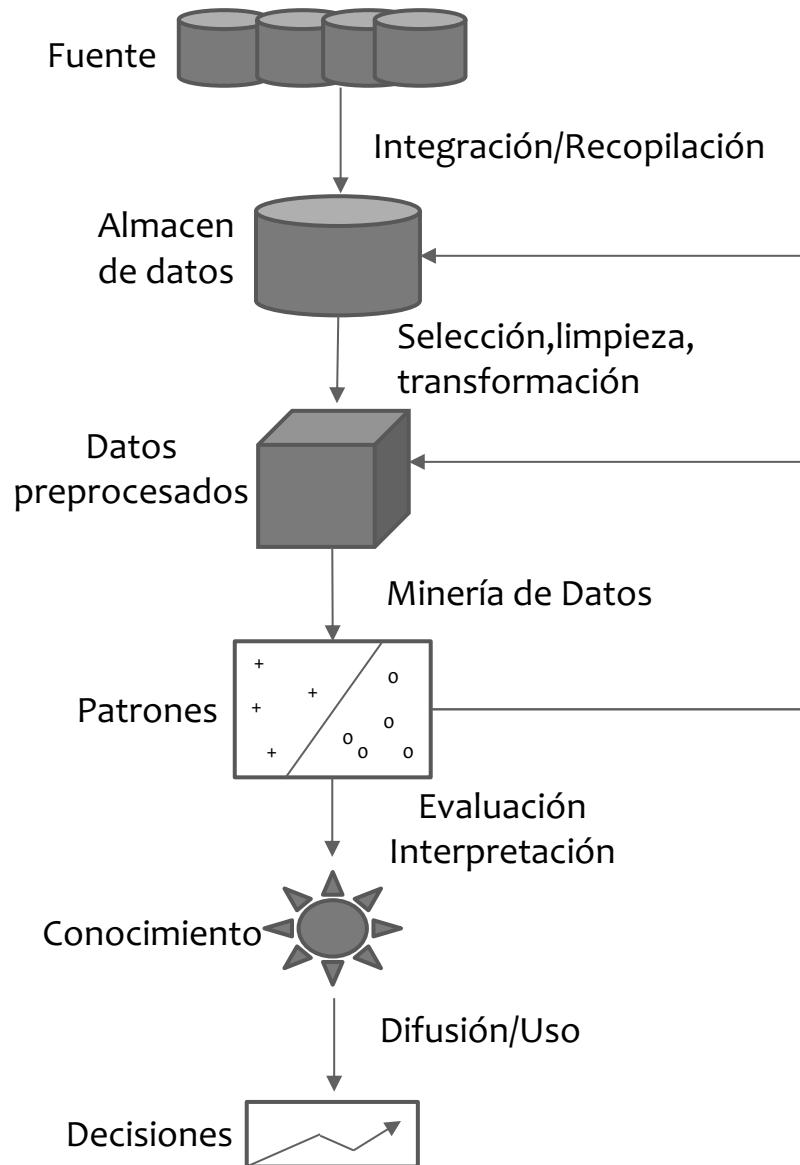
Índice

- Preparación de datos
- Recopilación
- Integración
- Limpieza
- Transformación
- Selección/Reducción

Fases del TID-KD

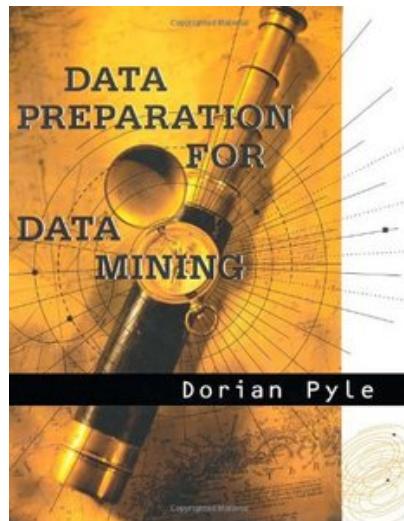
Fases del TID:

- Integración y recopilación
- Selección, limpieza y transformación
- Minería de datos
- Evaluación e interpretación
- Difusión y uso



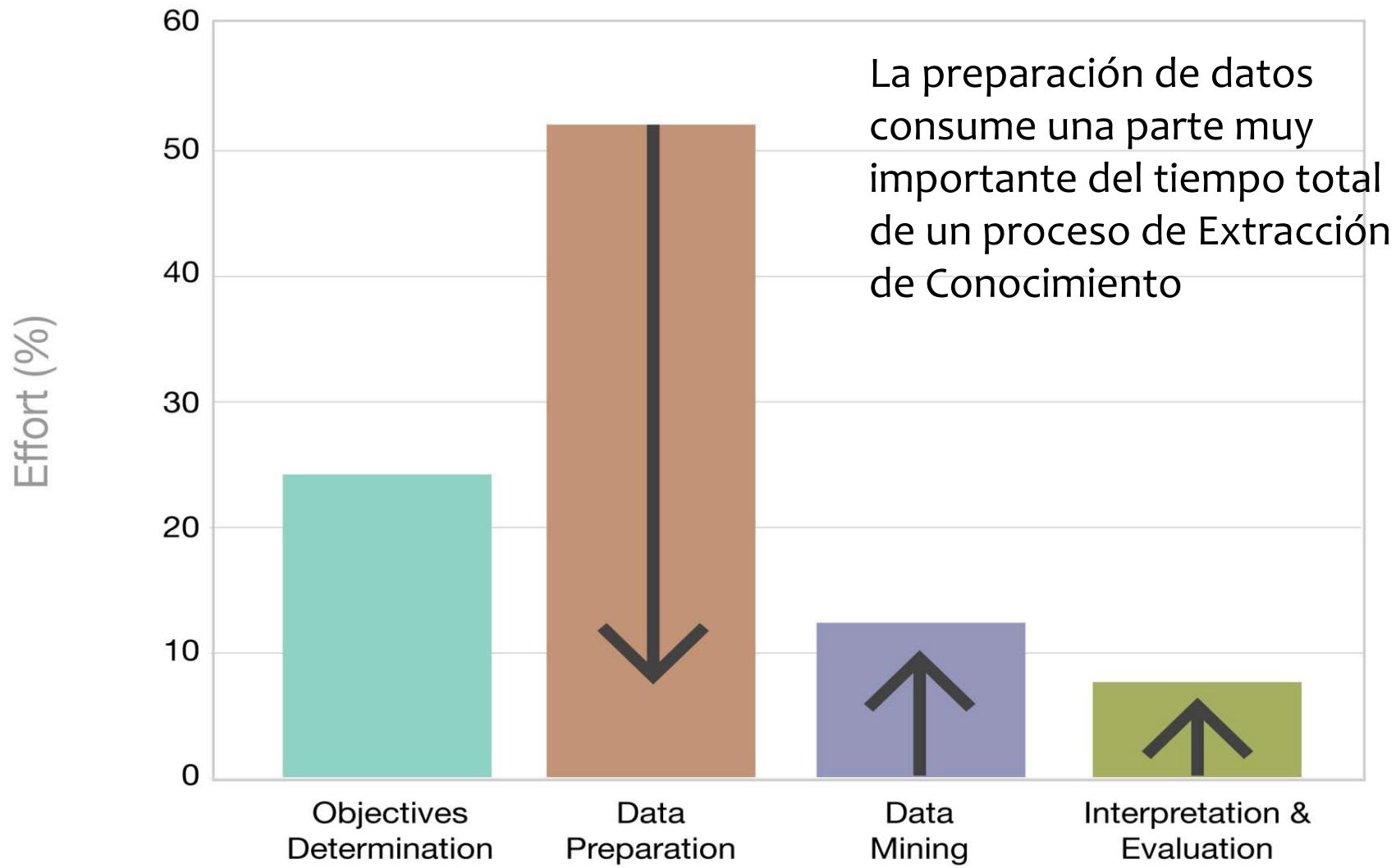
Preparación de datos

“El propósito fundamental de la preparación de los datos es la manipulación y transformación de los datos sin refinar para que la información contenida en el conjunto de datos pueda ser descubierta o estar accesible de forma más fácil”



Dorian Pyle,
Data Preparation for Data Mining,
Morgan Kaufmann Publishers, 1999

Preparación de datos



Preparación de datos

La preparación de datos es importante porque:

Los **datos reales pueden ser impuros**, esto puede conducir a la extracción de patrones/reglas poco útiles

- **Datos incompletos.** Falta de valores de atributos
- **Datos con ruido.** Malas mediciones,...
- **Datos inconsistentes.** Fusión de BBDD, integración desde distintas fuentes,...

La preparación de datos **genera “datos de calidad”**, los cuales pueden conducir a patrones/reglas de calidad

- Recuperar información incompleta
- Eliminar outliers
- Resolver conflictos
- Duplicados
- ...

Preparación de datos

La preparación de datos es importante porque:

La preparación de datos puede generar **un conjunto de datos más pequeño que el original**, lo cual puede mejorar la eficiencia del proceso de Minería de Datos

- **Selección relevante de datos**
 - eliminando registros duplicados
 - eliminando anomalías, ...
- **Reducción de datos**
 - Selección de características
 - muestreo o selección de instancias
 - discretización

Preparación de datos

Fases para la preparación de datos minables:

(Diferentes autores dan diferentes tareas y clasificaciones)

- Recopilación de datos
- Integración desde distintos orígenes/formato
- Limpieza. Eliminación de ruido e inconsistencias
- Transformación
 - Discretización, normalización,...
- Reducción/Selección
 - Tamaño, dimensión, eliminación redundancias,...

Recopilación de Datos

- Cuestionarios
- Entrevistas
- Informes
- Formularios
- BBDD de empresas adquiridas
- Compras por internet
- Clicks en anuncios de internet
- Búsquedas en Google
- ...

No trataremos eso aquí

Integración

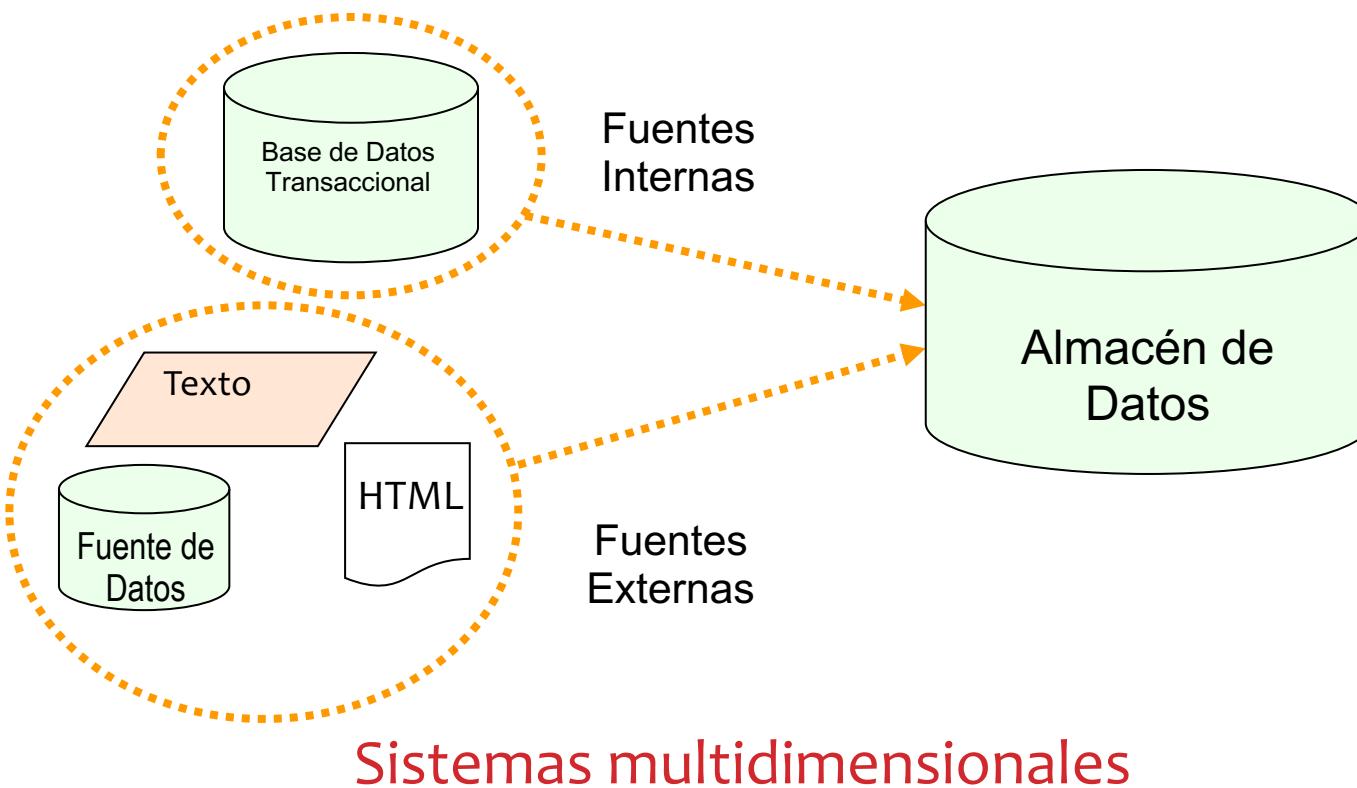
Gran cantidad de datos heterogéneos y, posiblemente, desestructurados



¿Cómo integrarlos para mejorar la búsqueda de patrones?

Integración

Una solución: **data warehouses** (almacenes de datos)



Integración

OLAP (OnLine Analytical Processing)

- Análisis en tiempo real
- Para cruzar gran cantidad de información
- Exclusivamente **de consulta**
- Para realizar informes y resúmenes
- Requieren tiempo y recursos en BBDD

Integración

No es lo mismo OLAP que Data Mining!

Análisis OLAP	Data Mining
¿Cuál es la proporción media de accidentes entre fumadores y no fumadores?	¿Cuál es la mejor predicción para accidentes?
¿Cuál es la factura telefónica media de mis clientes y de los que han dejado la compañía?	¿Dejara X la compañía? ¿Qué factores afectan a los abandonados?
¿Cuánto es la compra media diaria de tarjetas robadas y legítimas?	¿Cuáles son los patrones de compra asociados con el fraude de tarjetas?

Integración

	BASE DE DATOS TRANSACCIONAL	ALMACÉN DE DATOS
Propósito	Operaciones diarias. Soporte a las aplicaciones.	Recuperación de información, informes, análisis y minería de datos.
Tipo de datos	Datos de funcionamiento de la organización.	Datos útiles para el análisis, la summarización, etc.
Características de los datos	Datos de funcionamiento, cambiantes, internos, incompletos...	Datos históricos, datos internos y externos, datos descriptivos...
Modelo de datos	Datos normalizados.	Datos en estrella, en copo de nieve, parcialmente desnormalizados, multidimensionales...
Número y tipo de usuarios	Cientos/miles: aplicaciones, operarios, administrador de la base de datos.	Decenas: directores, ejecutivos, analistas (granjeros, mineros).
Acceso	SQL. Lectura y escritura.	SQL y herramientas propias (<i>slice & dice, drill, roll, pivot...</i>). Lectura.

No tienen la misma finalidad

Integración

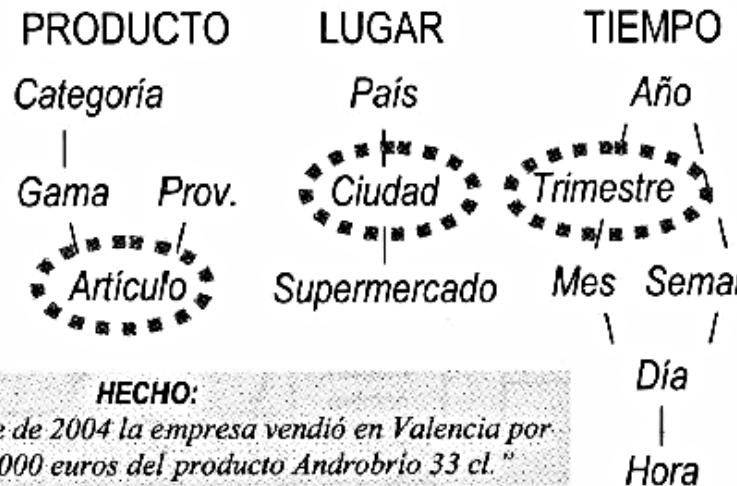
Representación tridimensional de un datamart

Ventas en miles de euros

PRODUCTO:
artículo

PRODUCTO: artículo	LUGAR: <i>ciudad</i>		TIEMPO: trimestre
	Zaragoza	Madrid	
	Barcelona	Valencia	
Zumo piña 1l.	17		
Cola 33 cl.	57		
Jabón Salitre	93		
Androbrio 33 cl	22		
Cerveza Kiel 20 cl	5		
Leche entera cabra 1l.	12		
	1 2 3 4	1 2	2004 2005

Jerarquía de dimensiones:



Integración

❑ Operador *roll*

CATEGORÍA	TRIMESTRE	IMPORTE
Refrescos	T1	150.323 euros
Refrescos	T2	233.992 euros
Refrescos	T3	410.497 euros
Refrescos	T4	203.400 euros
Congelados	T1	2.190.103 euros
Congelados	T2	1.640.239 euros
Congelados	T3	1.904.401 euros
Congelados	T4	2.534.031 euros



roll

un nivel por “tiempo”

CATEGORÍA	IMPORTE
Refrescos	998.212 euros
Congelados	10.458.877 euros

Integración

❑ Operador pivot

CATEGORÍA	TRIMESTRE	Valencia	León
Refrescos	T1	13.267	3.589
Refrescos	T2	27.392	4.278
Refrescos	T3	73.042	3.780
Refrescos	T4	18.391	3.629
Congelados	T1	150.242	4.798
Congelados	T2	173.105	3.564
Congelados	T3	163.240	4.309
Congelados	T4	190.573	4.812

pivot
categoría × ciudad

CATEGORÍA	TRIMESTRE	Refrescos	Congelados
Valencia	T1	13.267	150.242
Valencia	T2	27.392	173.105
Valencia	T3	73.042	163.240
Valencia	T4	18.391	190.573
León	T1	3.589	4.798
León	T2	4.278	3.564
León	T3	3.780	4.309
León	T4	3.629	4.812

Integración

❑ Operador *drill*

CATEGORÍA	TRIMESTRE	IMPORTE
Refrescos	T1	150.323 euros
Refrescos	T2	233.992 euros
Refrescos	T3	410.497 euros
Refrescos	T4	203.400 euros
Congelados	T1	2.190.103 euros
Congelados	T2	1.640.239 euros
Congelados	T3	1.904.401 euros
Congelados	T4	2.534.031 euros

drill
categoría= “refrescos”
ciudad= {“Valencia”, “León”}

CATEGORÍA	TRIMESTRE	CIUDAD	IMPORTE
Refrescos	T1	Valencia	13.267
Refrescos	T1	León	3.589
Refrescos	T2	Valencia	27.392
Refrescos	T2	León	4.278
Refrescos	T3	Valencia	73.042
Refrescos	T3	León	3.780
Refrescos	T4	Valencia	18.391
Refrescos	T4	León	3.629

Integración

❑ Operador *slice & dice*

CATEGORÍA	TRIMESTRE	Valencia	León
Refrescos	T1	13.267	3.589
Refrescos	T2	27.392	4.278
Refrescos	T3	73.042	3.780
Refrescos	T4	18.391	3.629
Congelados	T1	150.242	4.798
Congelados	T2	173.105	3.564
Congelados	T3	163.240	4.309
Congelados	T4	190.573	4.812



slice & dice

trimestre = {T1, T4}
ciudad = Valencia

CATEGORÍA	Trimestre	Valencia
Refrescos	T1	13.267
Refrescos	T4	18.391
Congelados	T1	150.242
Congelados	T4	190.573

Integración

- ❑ No es obligatorio realizar trabajos de Minería de Datos sobre un Almacén de Datos
- ❑ En esta asignatura trabajaremos sobre BBDD

KDD: Knowledge Discovery over DataBases

DIA: Data Intelligent Analysis

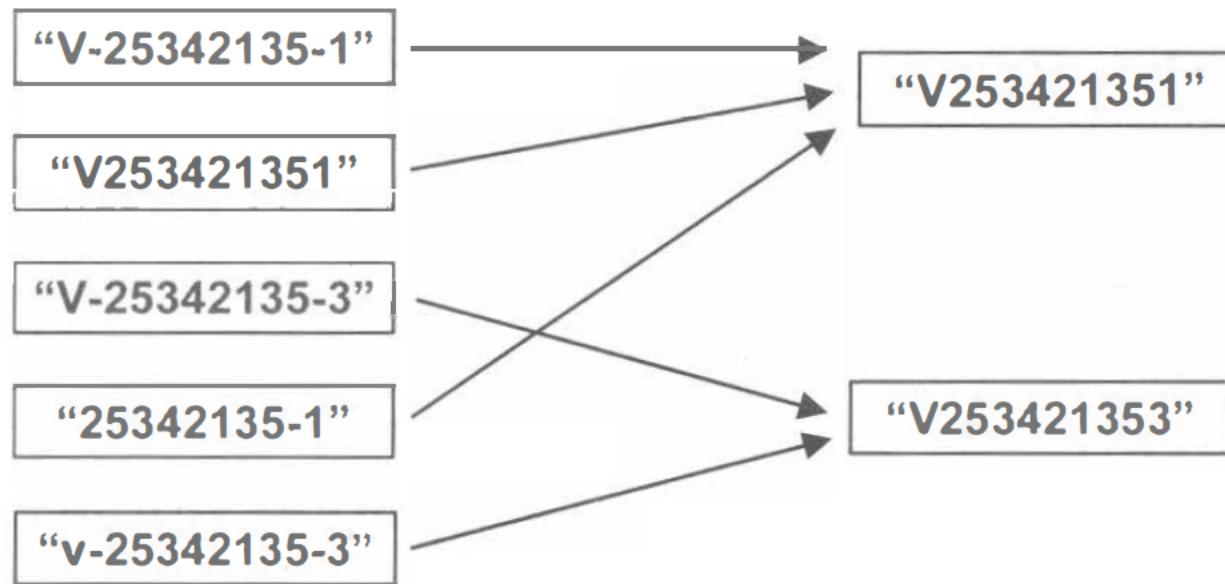
Y no trataremos los Almacenes de Datos

- ❑ Pero también existen problemas que tratar al realizar integración fuera de un DW

Integración

Algunos problemas con la integración:

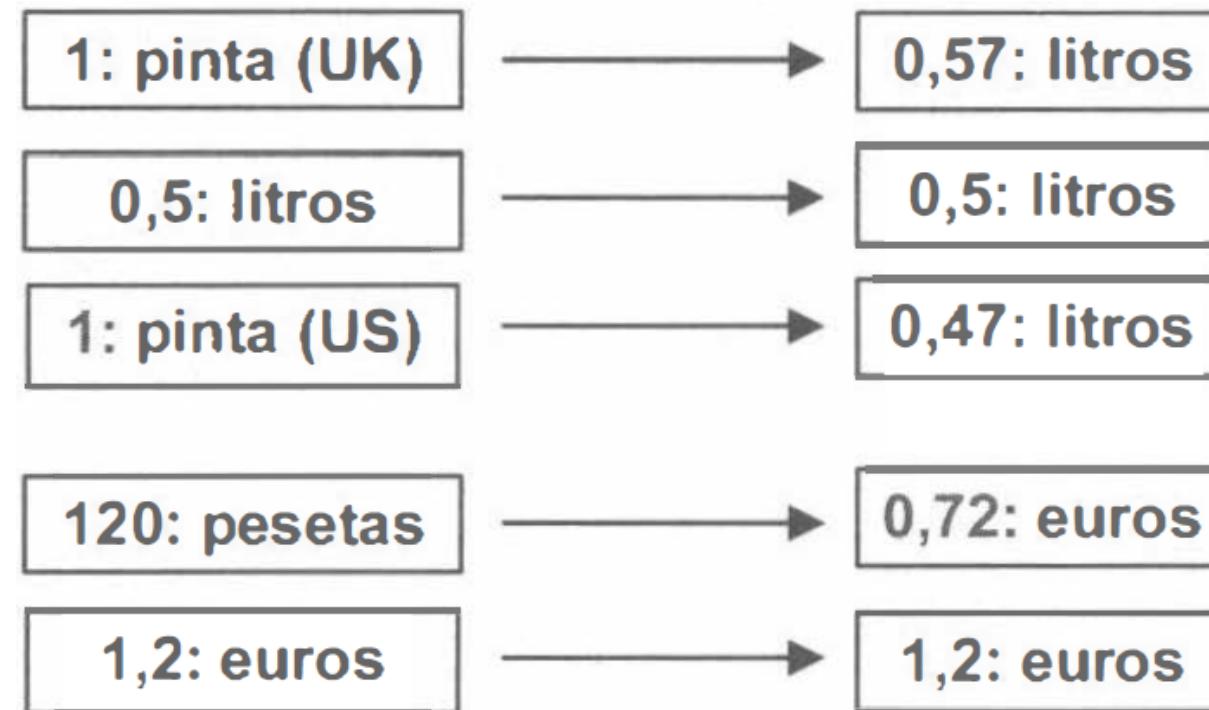
- Esclarecimiento de identidad:
 - Dos atributos diferentes se unifican
 - Dos atributos iguales se separan. Formato,...



Integración

Algunos problemas con la integración:

- Unificación de medidas



Integración

Soluciones a lo anterior:

- Utilizar los metadatos que normalmente se almacenan en las BBDD y los DW
- Con cierta supervisión
- En general, cuidar el proceso de integración a partir de múltiples fuentes reducirá y evitará redundancias e inconsistencias en los datos resultantes, mejorando la exactitud y velocidad del proceso de Data Mining

Una vez integrados los datos...

Normalmente, las técnicas y/o métodos utilizados dependen mucho de la naturaleza de los datos:

- Numéricos
 - Discretos
 - No discretos
- Nominales sin orden
- Nominales con orden

Siempre es bueno conocer el origen de los datos!

Análisis exploratorio!!!!

Análisis exploratorio

En algún momento (por ejemplo, antes de manipular los datos integrados) es conveniente realizar un **análisis exploratorio**

- Motivaciones para explorar los datos
 - Ayuda a elegir las mejores herramientas para preprocesar y analizar
 - Permite formular hipótesis iniciales sobre patrones a extraer ya que se explota la habilidad del ser humano para reconocer patrones
- El análisis exploratorio de Datos (EDA) 1977
 - Es debido a Tuckey
 - El EDA está enfocado a la visualización. Supone que con técnicas adecuadas se puede extraer conocimiento directo.
 - Para Tuckey, el clustering y la detección de anomalías forman parte
 - Más información
<http://www.itl.nist.gov/div898/handbook/eda/eda.htm>

Análisis exploratorio

Basado en estadística descriptiva

- ❑ Distribución de frecuencias. **Frecuencias absolutas:** número de veces que aparece un determinado valor
 - Atributos discretos: Categóricos y numéricos enteros y finitos (con no muchos datos en el dominio). Simple conteo sobre los valores del dominio
 - Atributos continuos. Se impone la discretización, el dominio se transforma en discreto mediante intervalos.
Normalmente, los intervalos se eligen de igual amplitud
 - El problema de la discretización puede ser complejo para temas de visualización y asociación. En algunos casos es un tema de preprocesamiento.

Análisis exploratorio

Basado en estadística descriptiva

- ❑ Distribución de frecuencias. **Frecuencias relativas**: razón entre la frecuencia absoluta y el número total de ítems. Se puede dar tambien en porcentajes

- ❑ Distribución de frecuencias. **Frecuencias acumulativas**: Se definen sólo para datos ordenados. Número de veces que aparece algo menor que cierto valor determinado. Cuando se utilizan porcentajes nos indica el porcentaje de la población con valores menores y conduce a concepto de **percentil**: el mayor valor del dominio que tiene por debajo al p por ciento de la población

Análisis exploratorio

Basado en estadística descriptiva

- **Medidas de centralización.** Para datos numéricos
 - Media, Moda, Mediana (percentil 50)
- **Medidas de dispersión.** Para datos numéricos
 - Varianza, desviación típica, media de la desviación absoluta, mediana de la desviación absoluta, rango intercuartil (p75-p25, para todos los tipos de datos)
- **Exploración de la relación entre atributos numéricos**
 - Matriz de covarianzas, matriz de correlación

Análisis exploratorio

Basado en gráficas

□ Visualización de ciertas gráficas

- Boxplot
- Histogramas
- Diagramas de líneas
- Nube de puntos (scatter plot)
- etc,...

Limpieza de datos

Objetivos:

- Resolver inconsistencias o datos erróneos
- Rellenar valores perdidos
- Suavizar el ruido de los datos
- Eliminar outliers

Algunos algoritmos de DM tienen métodos propios para tratar con datos incompletos o ruidosos. Pero en general estos métodos no son muy robustos, lo normal es realizar previamente la limpieza de los datos.

Limpieza de datos (valores perdidos)

- Debemos rellenarlos si
 - Nuestro algoritmo no funciona si hay valores perdidos
 - Nuestro algoritmo los rellena... malamente
 - Para calcular otro tipo de medidas: media, moda, etc...
- No necesitamos rellenarlos si:
 - Nuestro algoritmo los trata bien
- Detectar valores perdidos.
 - BBDD tiene campos nulos
 - Está marcado con algún símbolo especial, “-”
 - Valores “raros”
 - teléfono con 0000,
 - tarjeta crédito -1,
 - ...

Limpieza de datos (valores perdidos)

□ Qué hacer con los valores perdidos?

- **Ignorar la tupla.** Suele usarse cuando la variable a clasificar no tiene valor o el algoritmo va bien con valores perdidos
- **Ignorar la fila.** Sesga los datos, pero la falta de valor se puede deber a algún motivo relevante
- **Rellenar manualmente** los datos. En general es impracticable
- Utilizar una constante global para la sustitución. Por ejemplo “desconocido”,...
- **Rellenar con el valor más probable.**
 - utilizar alguna técnica de inferencia, por ejemplo, bayesiana o un árbol de decisión
 - utilizar la media/desviación del resto de las tuplas
 - utilizar la media/desviación del resto de las tuplas pertenecientes a la misma clase.

Limpieza de datos (valores perdidos)

Posición	Valor original	Pos. 11 perdida	Preservar la media	Preservar la desviación
1	0.0886	0.0886	0.0886	0.0886
2	0.0684	0.0684	0.0684	0.0684
3	0.3515	0.3515	0.3515	0.3515
4	0.9875	0.9875	0.9875	0.9875
5	0.4713	0.4713	0.4713	0.4713
6	0.6115	0.6115	0.6115	0.6115
7	0.2573	0.2573	0.2573	0.2573
8	0.2914	0.2914	0.2914	0.2914
9	0.1662	0.1662	0.1662	0.1662
10	0.4400	0.4400	0.4400	0.4400
11	0.6939	????	0.3731	0.6622
Media	0.4023	0.3731	0.3731	0.3994
SD	0.2785	0.2753	0.2612	0.2753
Error en la estimación			0.3208	0.0317

Limpieza de datos (valores perdidos)

Rellenar buscando relaciones entre variables

Por ejemplo, de los datos de las columnas X e Y, se podría estimar $Y = 1.06X$ y utilizarlo como estimador para valores perdidos de Y

X (orig.)	Y (orig.)	Y estimado	error
0.55	0.53	0.51	0.02
0.75	0.37	0.31	0.06
0.32	0.83	0.74	0.09
0.21	0.86	0.85	0.01
0.43	0.54	0.63	0.09

Limpieza de datos (valores perdidos)

Rellenar usando la moda dentro de la clase

X	Y	clase
1	A	1
1	A	0
?	B	0
2	B	1
2	B	0
2	B	0

$$P(1|clase = 0) = 0.33 \quad P(2|clase = 0) = 0.67$$

Ponemos un 2!

Limpieza de datos (valores erróneos)

Un dato erróneo es un dato que no es cierto aunque estadísticamente sea normal

□ Detección y corrección de datos erróneos:

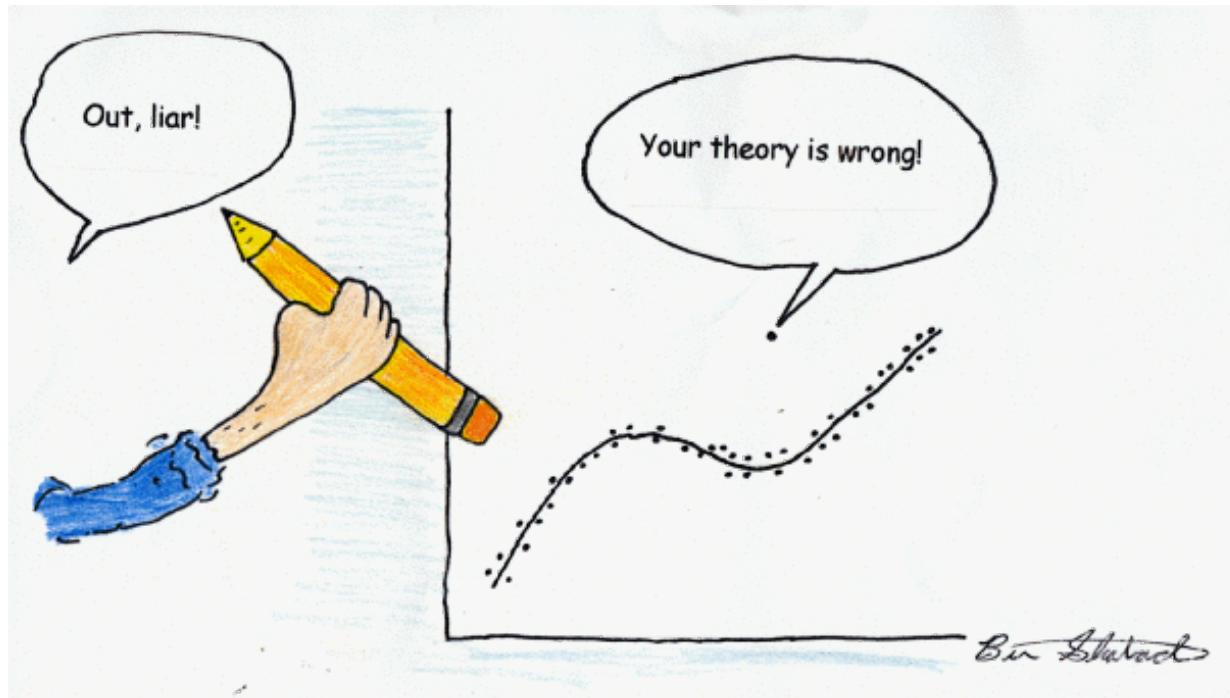
- Nominal. No se ajusta al formato, valor no permitido. Normalmente, eso está arreglado por restricciones de integridad

**Si se ajustan al formato es muy
difícil encontrar valores erróneos
salvo que sepamos algo del contenido**

- Numérico. Buscar outliers y mirar si son correctos

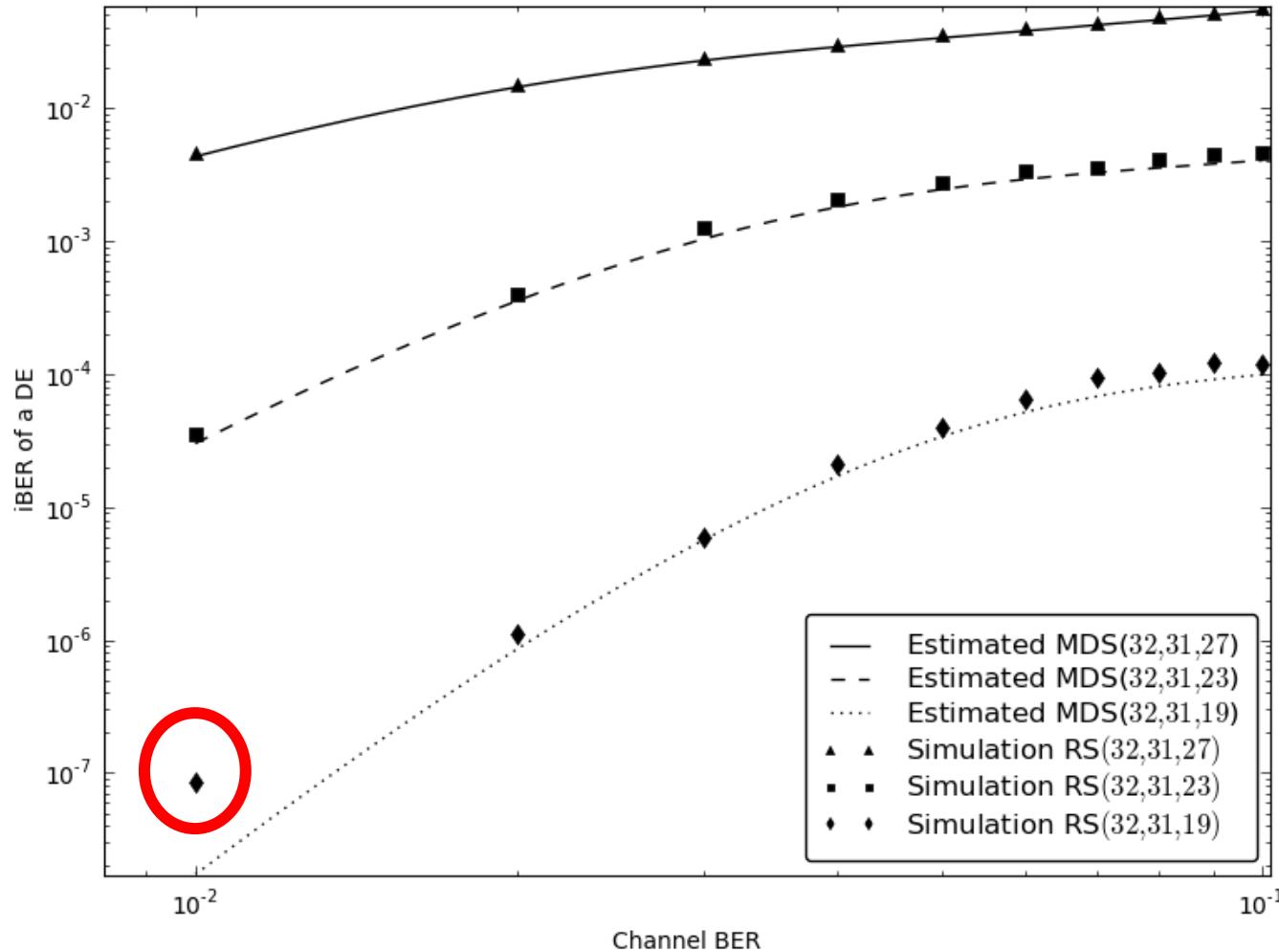
Limpieza de datos (Outliers)

Un dato anómalo o outlier es un dato que es cierto pero estadísticamente anómalo



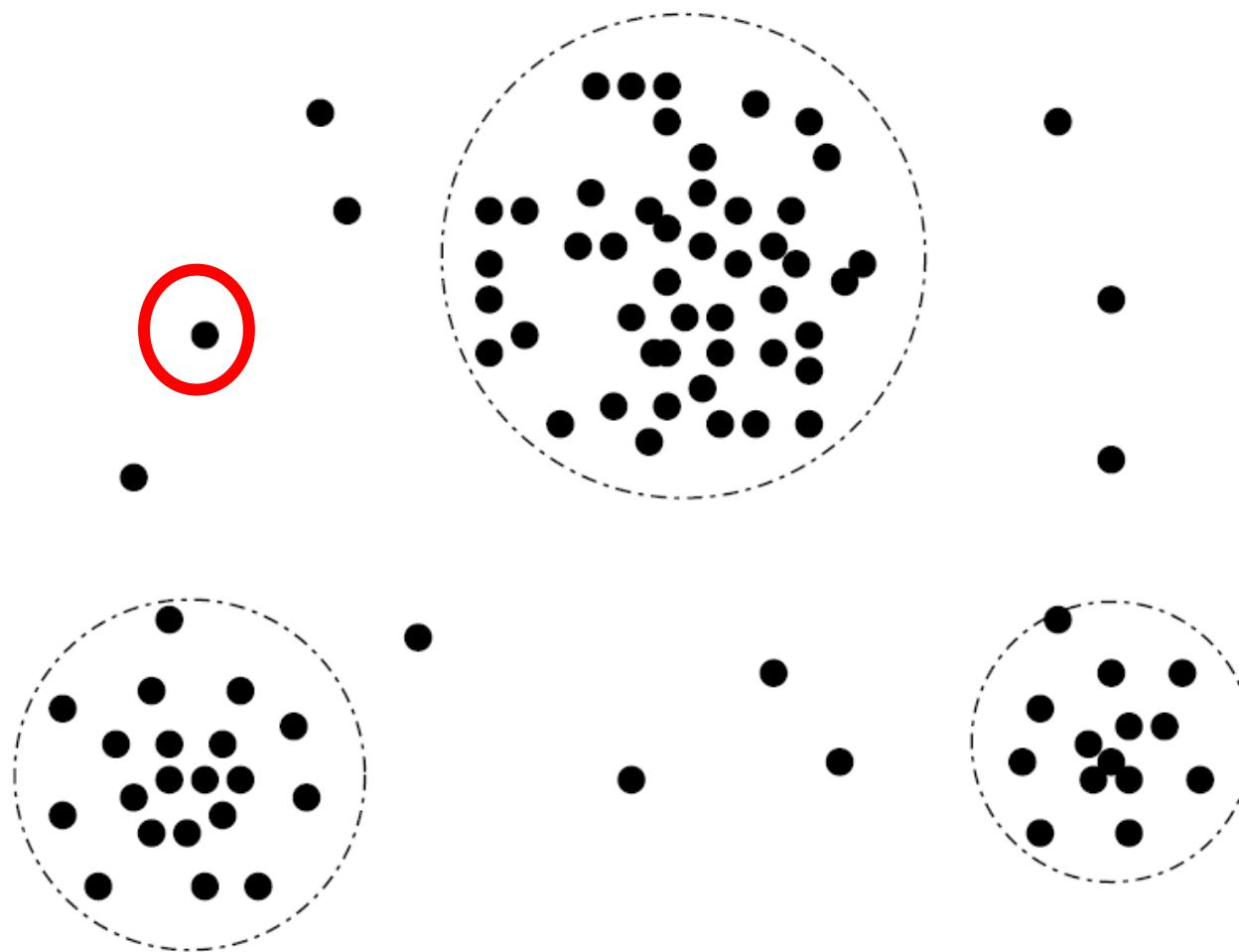
Limpieza de datos (Outliers)

Por ejemplo,



Limpieza de datos (Outliers)

Son objetos/datos con características que son considerablemente diferentes de la mayoría de los otros datos/objetos del conjunto



Limpieza de datos (Outliers)

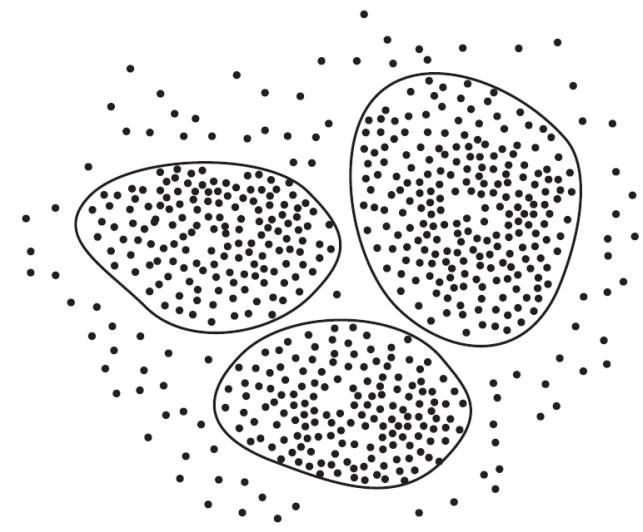
Aunque sean datos ciertos, **conviene eliminarlos** porque pueden

- distorsionar las estadísticas
- distorsionar la normalización
- no permitir un buen proceso de Minería de Datos
 - ser un inconveniente para métodos basados en ajuste de pesos
- deberse a elementos ajenos al experimento en sí

Limpieza de datos (Outliers)

Detección de outliers:

- Mediante técnicas estadísticas
 - Histogramas y otras visualizaciones
 - Seguimiento de distribuciones
 - Test de discordancias
- Definir una distancia y ver los individuos con mayor distancia media al resto de individuos
- Clustering parcial: los datos se agrupan en clusters y los datos que queden fuera pueden considerarse outliers



Limpieza de datos (Outliers)

Tratamiento de valores anómalos o erróneos:

- Ignorar**. Algunos algoritmos son robustos a outliers
- Filtrar la columna**. Solución extrema, conveniente si existe una columna dependiente con datos de mayor calidad.
- Filtrar la fila**. Puede sesgar los datos porque las causas de un dato erróneo están relacionadas con casos o tipos especiales
- Reemplazar el valor por valor nulo**, si el algoritmo de DM trabaja bien con datos nulos, máximo, mínimo o la media
- Discretizar**. Si transformamos un valor continuo en discreto (muy alto, alto, ..., muy bajo), los datos anómalos caen en la categoría muy alto o muy bajo y se tratan sin problemas

Limpieza de datos (suavizar ruido)

El ruido se elimina mediante técnicas de suavizado

- ❑ **Regresión:** Los datos se suavizan ajustándolos a una función con técnicas de regresión
- ❑ **Discretización (Binning):** Se suavizan valores ordenados consultando sus vecinos. Los valores se distribuyen en un conjunto de cajas o intervalos (*bins*). Realiza un suavizado local.
- ❑ Variantes:
 - *Binning* uniforme en los intervalos (*equiwidth*) o
 - en el contenido (*equidepth*)
 - Suavizar por la media o mediana
 - Suavizar por las fronteras

Transformación de datos

La transformación de datos engloba cualquier **proceso que modifique la forma de los datos**

- pero... prácticamente todos los procesos de preparación de datos entrañan algún tipo de transformación
- Aquí entenderemos transformación de atributos o quedarnos con algunos

Objetivo: poner los datos de la mejor forma posible para la aplicación de los algoritmos de Data Mining

Transformación de datos

Objetivo: poner los datos de la mejor forma posible para la aplicación de los algoritmos de Data Mining

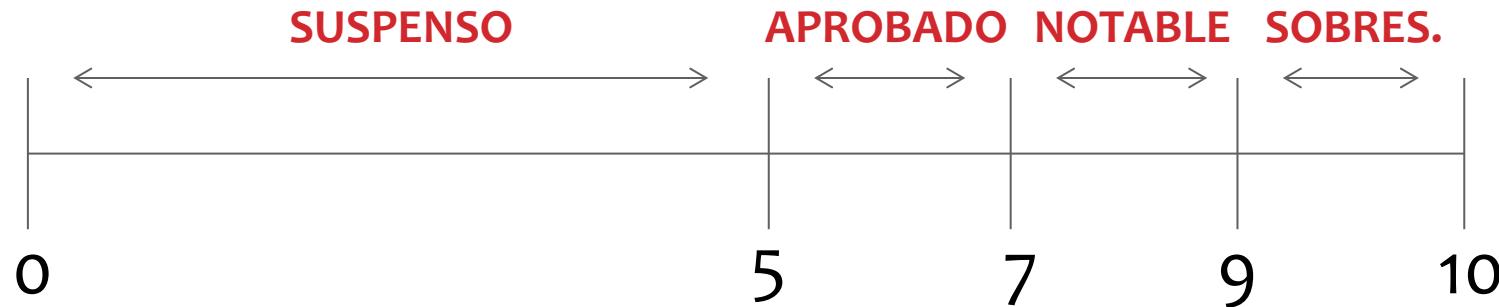
- Discretización o binning
- Numerización
- Normalización
- Intercambio de dimensiones
- ...

Transformación (binning)

Discretización: conversión de un valor numérico en un valor nominal ordenado (*bin*)

Se trata de agrupar el rango de valores para reducir su número

Ejemplo, notas de un examen



Transformación (binning)

Razones para discretizar:

- reducir el rango de valores
- el error en la medida puede ser grande (outliers)
- existen ciertos umbrales significativos
- pasar atributos numéricos a nominales
 - Algunos algoritmos sólo aceptan nominales
- representar la información de forma más concisa
 - Los datos son más fáciles de entender, más cercanos a la representación a nivel de conocimiento
- las diferencias en ciertas zonas del rango de valores sean más importantes que en otras

Transformación (binning)

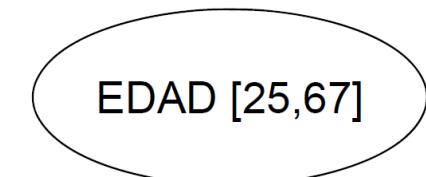
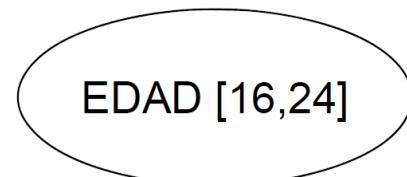
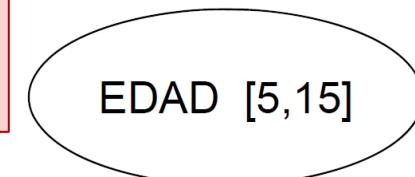
- ❑ Los valores discretos son muy útiles en Minería de Datos
- ❑ Los valores nominales tienen un dominio finito, por lo que también se considera una **técnica de reducción**
- ❑ La discretización puede hacerse antes de la obtención de conocimiento o durante dicha etapa

Transformación (binning)

- ❑ Divide el rango de atributos continuos (numéricicos) en intervalos
- ❑ Almacena solo las etiquetas de los intervalos

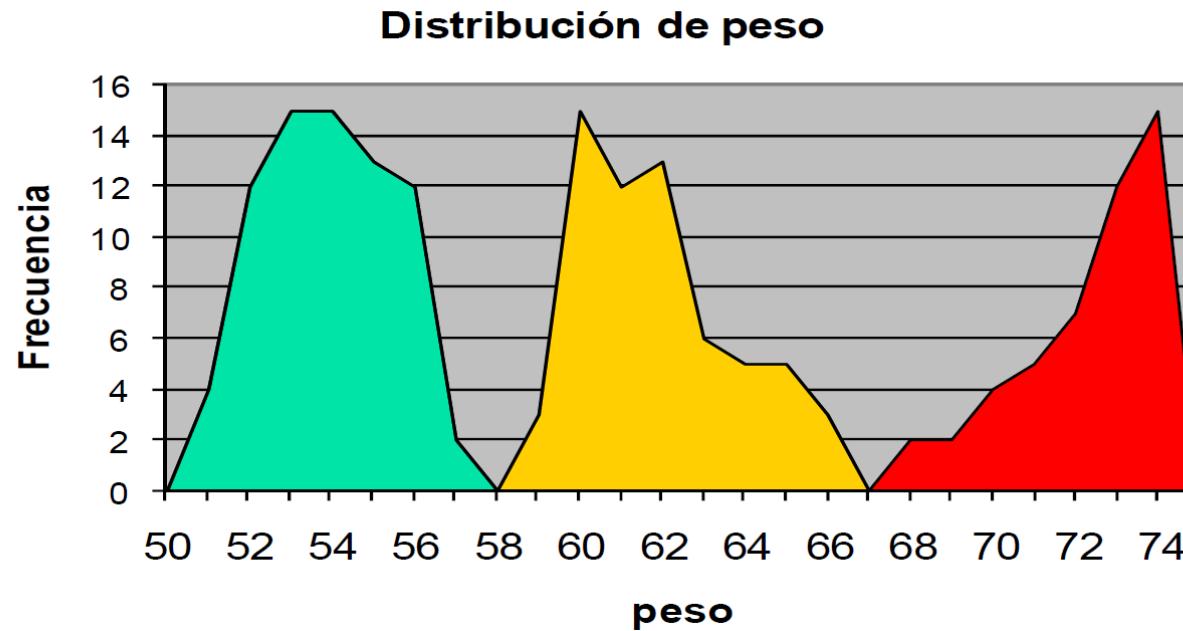
EDAD	5	6	6	9	...	15	16	16	17	20	...	24	25	41	50	65	...	67
COCHE EN PROPIEDAD	0	0	0	0	...	0	1	0	1	1	...	0	1	1	1	1	...	1

Es poco práctico
este rango de
valores!



Transformación (binning)

Existen ciertos umbrales significativos

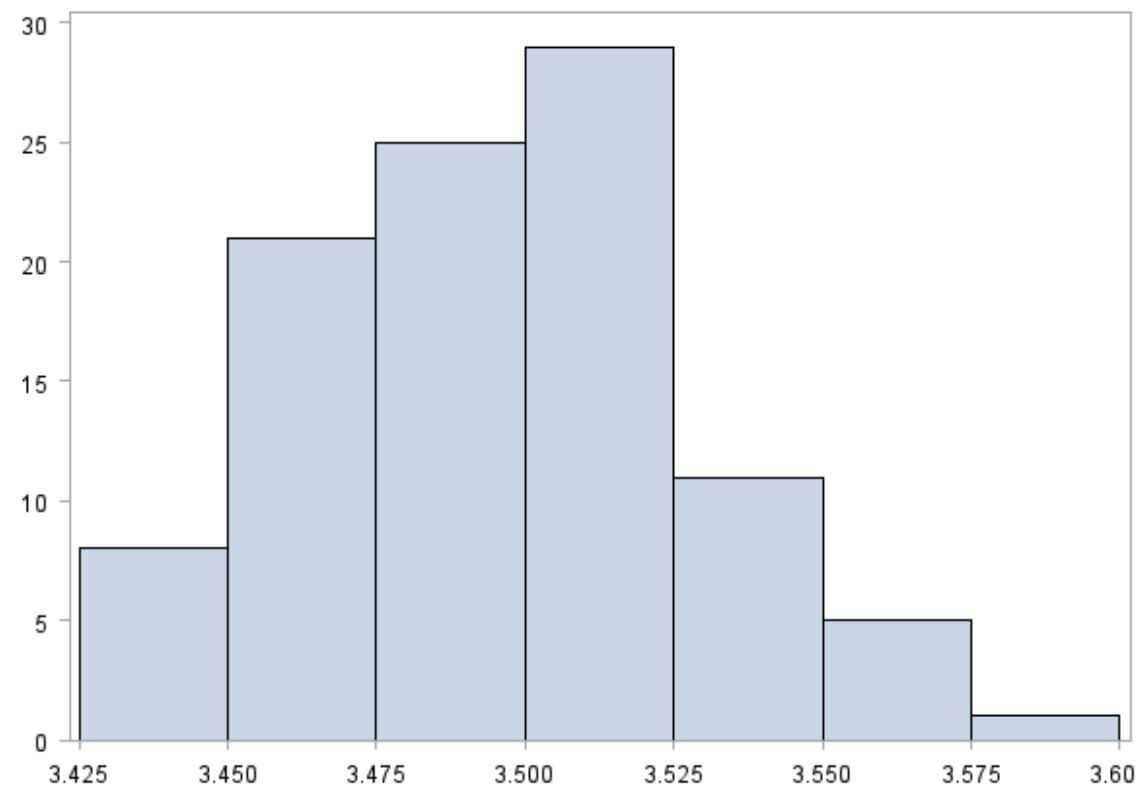


50 - 58 kg
59-67 kg
> 68 kg

Transformación (binning)

¿Cómo dividir el rango?

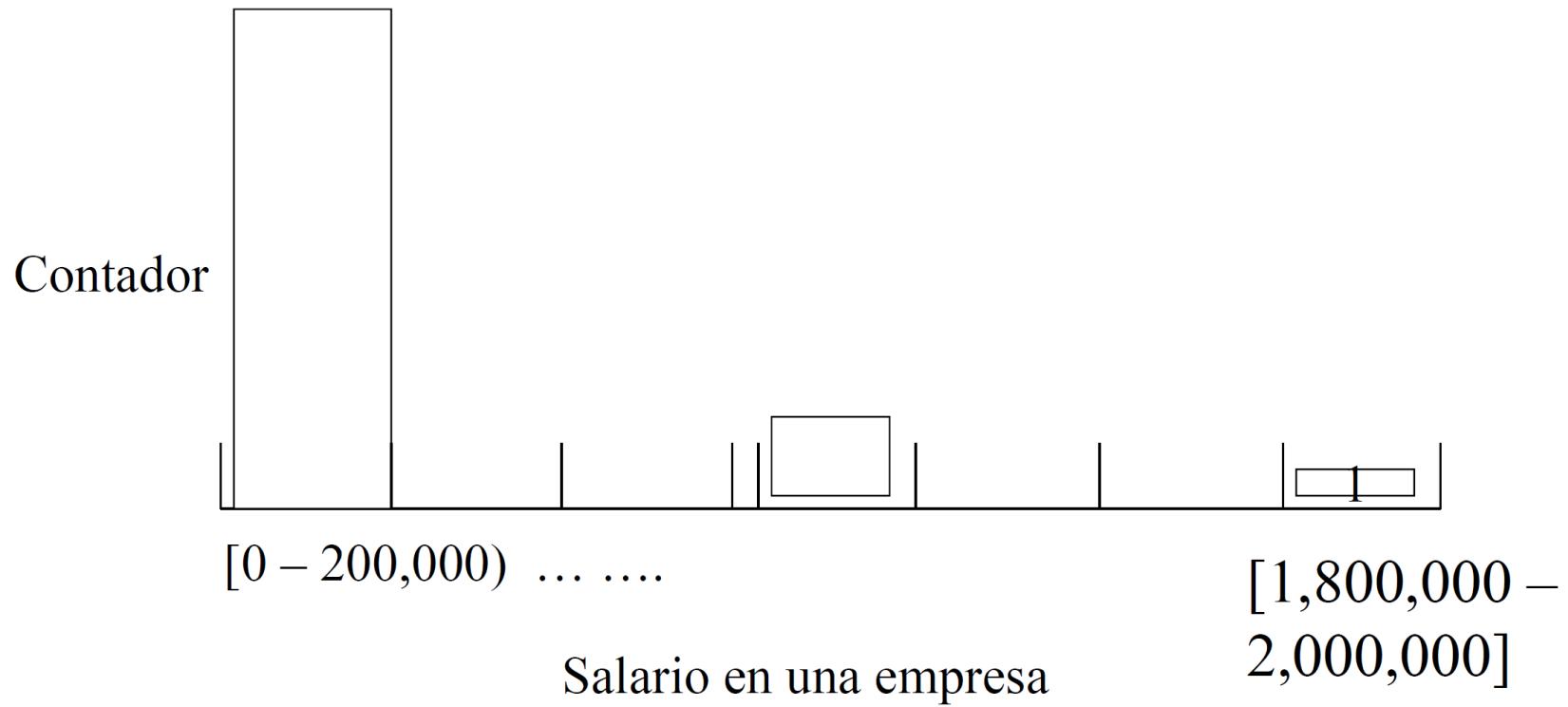
- Intervalos de igual longitud



Transformación (binning)

¿Cómo dividir el rango?

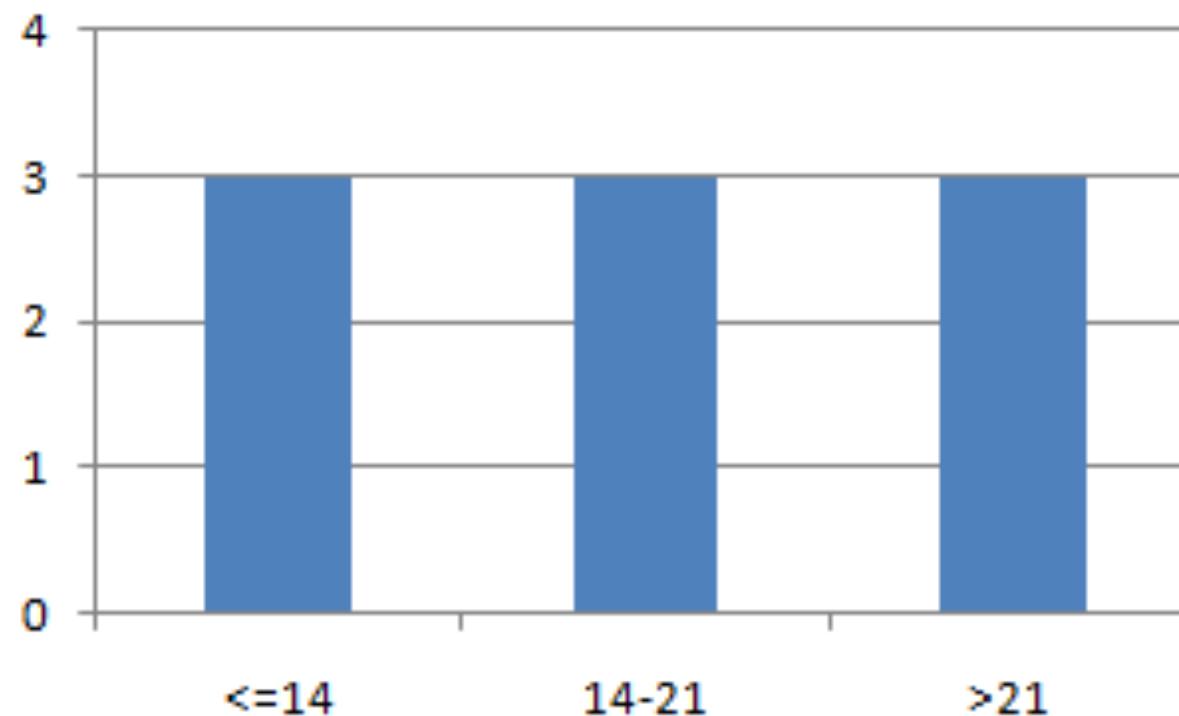
Problema: desequilibrio entre rangos



Transformación (binning)

¿Cómo dividir el rango?

- Igual frecuencia



Transformación (binning)

¿Cómo dividir el rango?

- Ventajas de la igualdad en frecuencia
 - Generalmente es preferible porque evita desequilibrios en el balanceo entre valores
 - En la práctica permite obtener puntos de corte más intuitivos
- Consideraciones adicionales:
 - Se deben crear cajas para valores especiales
 - Se deben tener puntos de corte interpretables

Transformación (binning)

En general, uno sigue alguno de los métodos siguientes

- Utilizar el conocimiento inherente de los datos y **elegir manualmente** los rangos (si se puede, es lo mejor)
- Dividir en intervalos de igual longitud, frecuencia,...
- Utilizar algún algoritmo basado en heurísticas

Transformación (binning)

¿Cómo dividir el rango?

Tipos de algoritmos para calcular intervalos

- **Supervisados vs. no supervisados**
- **Dinámicos vs. estáticos:** Mientras se construye o no el modelo
- **Locales vs. Globales:** Centrados en una subregión del espacio de instancias o considerando todas ellas
- **Top-down vs. Bottom-up:** Empiezan con una lista vacía o llena de puntos de corte
- **Directos vs. Incrementales:** Usan o no un proceso de optimización posterior

Transformación (binning)

¿Cómo dividir el rango?

Algunos algoritmos concretos:

- Maximum entropy
- IEM (Information Entropy Maximization)
- CADD (Class-Attribute Dependence Discretizer)
- CAIM (Class-Attribute Independence Maximization)
- PKID (Proportional K-Interval Discretizer)
- FFD (Fixed Frequency Discretizer)

<http://sci2s.ugr.es/publications/ficheros/2013-Garcia-IEEETKDE.pdf>

CAIM

L. A. Kurgan, K. J. Cios, CAIM Discretization Algorithm, IEEE Transactions on Knowledge and Data Engineering 16(2) 2004, 145-153

- ❑ Algoritmo de discretización **supervisado**
- ❑ Necesita de un **conjunto de entrenamiento**
- ❑ Busca el **menor número de intervalos**
 - Comienza con el intervalo total y lo va dividiendo (top-down)
- ❑ Maximiza el número de ejemplos de la misma clase en el mismo intervalo

CAIM

Para una discretización D, podemos calcular su “quanta matrix”

Class	Interval					Class Total
	$[d_0, d_1]$...	$(d_{r-1}, d_r]$...	$(d_{n-1}, d_n]$	
C_1	q_{11}	...	q_{1r}	...	q_{1n}	M_{1+}
:	:	...	:	...	:	:
C_i	q_{i1}	...	q_{ir}	...	q_{in}	M_{i+}
:	:	...	:	...	:	:
C_s	q_{s1}	...	q_{sr}	...	q_{sn}	M_{s+}
Interval Total	M_{+1}	...	M_{+r}	...	M_{+n}	M

CAIM

Para una discretización D, podemos calcular su “quanta matrix”

		Discretización D					Número de ejemplos en la clase i
Class	Interval						Class Total
		[d ₀ , d ₁]	...	(d _{r-1} , d _r)	...	(d _{n-1} , d _n)	
C ₁	q ₁₁	...	q _{1r}	...	q _{1n}		M ₁₊
:	:	...	:	...	:		:
C _i	q _{i1}	...	q _{ir}	...	q _{in}		M _{i+}
:	:	...	:	...	:		:
C _s	q _{s1}	...	q _{sr}	...	q _{sn}		M _{s+}
Interval Total	M ₊₁	...	M _{+r}	...	M _{+n}		M

Cada una de las Clases posibles
 Número de ejemplos en la clase i con F en el intervalo r
 Número de ejemplos en la clase i
 Número de ejemplos en el intervalo r
 Número de ejemplos

CAIM

Medida a maximizar (dependencia de C y D para el atributo F)

$$CAIM(C, D|F) = \frac{\sum_{r=1}^n \frac{\max_r^2}{M_{+r}}}{n}$$

Class	Interval					Class Total
	[(d ₀ , d ₁]	...	(d _{r-1} , d _r]	...	(d _{n-1} , d _n]	
C ₁	q ₁₁	...	q _{1r}	...	q _{1n}	M ₁₊
:	:	...	:	...	:	:
C _i	q _{i1}	...	q _{ir}	...	q _{in}	M _{i+}
:	:	...	:	...	:	:
C _S	q _{s1}	...	q _{sr}	...	q _{sn}	M _{S+}
Interval Total	M ₊₁	...	M _{+r}	...	M _{+n}	M

CAIM

Medida a maximizar (dependencia de C y D para el atributo F)

$$CAIM(C, D|F) = \frac{\sum_{r=1}^n \frac{\max_r^2}{M_{+r}}}{n}$$

Número de intervalos de la discretización

Class	Interval					Class Total
	[d ₀ , d ₁]	...	(d _{r-1} , d _r)	...	(d _{n-1} , d _n)	
C ₁	q ₁₁	...	q _{1r}	...	q _{1n}	M ₁₊
:	:	...	:	...	:	:
C _i	q _{i1}	...	q _{ir}	...	q _{in}	M _{i+}
:	:	...	:	...	:	:
C _S	q _{s1}	...	q _{sr}	...	q _{sn}	M _{S+}
Interval Total	M ₊₁	...	M _{+r}	...	M _{+n}	M

CAIM

Medida a maximizar (dependencia de C y D para el atributo F)

$$CAIM(C, D|F) = \frac{\sum_{r=1}^n \frac{\max_r^2}{M_{+r}}}{n}$$

Máximo de elementos de una clase en el intervalo r

Class	Interval					Class Total
	[d ₀ , d ₁]	...	(d _{r-1} , d _r)	...	(d _{n-1} , d _n)	
C ₁	q ₁₁	...	q _{1r}	...	q _{1n}	M ₁₊
:	:	...	:	...	:	:
C _i	q _{i1}	...	q _{ir}	...	q _{in}	M _{i+}
:	:	...	:	...	:	:
C _S	q _{s1}	...	q _{sr}	...	q _{sn}	M _{S+}
Interval Total	M ₊₁	...	M _{+r}	...	M _{+n}	M

CAIM

Algorithm 1 CAIM

Input: S conjunto de clases con l clases

Input: M conjunto de ejemplos clasificados cada uno a una única clase de S

Input: $F =$ un atributo continuo para discretizar

Output: D , discretización de F

- 1: Ordenar valores de F , $F = \{d_0, d_1, \dots, d_m\}$.
 - 2: $D \leftarrow [d_0, d_n]$
 - 3: GlobalCAIM $\leftarrow 0$
 - 4: **for** $0 < k \leq l$ **do**
 - 5: Calcular el valor CAIM para todas las particiones en subintervalos de D añadiendo un intervalo más
 - 6: CAIMmax \leftarrow valor máximo de CAIM para una partición D'
 - 7: **if** CAIMmax $>$ GlobalCAIM **then**
 - 8: GlobalCAIM \leftarrow CAIMmax
 - 9: $D \leftarrow D'$
 - 10: **else** Terminar bucle
 - 11: **return** D
-

Ejemplo CAIM

A1	A2	A3	Clase
5.1	3.5	1.4	A
4.9	3	1.4	A
4.7	3.2	1.3	A
4.6	3.1	1.5	A
7	3.2	4.7	B
6.4	3.2	4.5	B
6.9	3.1	4.9	B
5.5	2.3	4	B
6.3	3.3	6	C
5.8	2.7	5.1	C
7.1	3	5.9	C
6.3	2.9	5.6	C

Ejemplo CAIM

A1	A2	A3	Clase
4.7	3.2	1.3	A
5.1	3.5	1.4	A
4.9	3	1.4	A
4.6	3.1	1.5	A
5.5	2.3	4	B
6.4	3.2	4.5	B
7	3.2	4.7	B
6.9	3.1	4.9	B
5.8	2.7	5.1	C
6.3	2.9	5.6	C
7.1	3	5.9	C
6.3	3.3	6	C

Ejemplo CAIM

A1	A2	A3	Clase
4.7	3.2	1.3	A
5.1	3.5	1.4	A
4.9	3	1.4	A
4.6	3.1	1.5	A
5.5	2.3	4	B
6.4	3.2	4.5	B
7	3.2	4.7	B
6.9	3.1	4.9	B
5.8	2.7	5.1	C
6.3	2.9	5.6	C
7.1	3	5.9	C
6.3	3.3	6	C

D=[1.3,6]
GlobalCAIM=0

Ejemplo CAIM

A1	A2	A3	Clase
4.7	3.2	1.3	A
5.1	3.5	1.4	A
4.9	3	1.4	A
4.6	3.1	1.5	A
5.5	2.3	4	B
6.4	3.2	4.5	B
7	3.2	4.7	B
6.9	3.1	4.9	B
5.8	2.7	5.1	C
6.3	2.9	5.6	C
7.1	3	5.9	C
6.3	3.3	6	C

1.23

D=[1.3,6]
GlobalCAIM=0

	[1.3,1.4)	[1.4,6]	
A	1	3	4
B	0	4	4
C	0	4	4
	1	11	12

CAIM=27/22=1.23

Ejemplo CAIM

A1	A2	A3	Clase
4.7	3.2	1.3	A
5.1	3.5	1.4	A
4.9	3	1.4	A
4.6	3.1	1.5	A
5.5	2.3	4	B
6.4	3.2	4.5	B
7	3.2	4.7	B
6.9	3.1	4.9	B
5.8	2.7	5.1	C
6.3	2.9	5.6	C
7.1	3	5.9	C
6.3	3.3	6	C

D=[1.3,6]
GlobalCAIM=0

1.23
2.4

	[1.3,1.5)	[1.5,6]	
A	3	1	4
B	0	4	4
C	0	4	4
	3	9	12

CAIM=147/66=2.23

Ejemplo CAIM

A1	A2	A3	Clase
4.7	3.2	1.3	A
5.1	3.5	1.4	A
4.9	3	1.4	A
4.6	3.1	1.5	A
5.5	2.3	4	B
6.4	3.2	4.5	B
7	3.2	4.7	B
6.9	3.1	4.9	B
5.8	2.7	5.1	C
6.3	2.9	5.6	C
7.1	3	5.9	C
6.3	3.3	6	C

D=[1.3,6]
GlobalCAIM=0

1.23
2.4
3

	[1.3,4)	[4,6]	
A	4	0	4
B	0	4	4
C	0	4	4
	4	8	12

CAIM=(4+2)/2=3

Ejemplo CAIM

A1	A2	A3	Clase
4.7	3.2	1.3	A
5.1	3.5	1.4	A
4.9	3	1.4	A
4.6	3.1	1.5	A
5.5	2.3	4	B
6.4	3.2	4.5	B
7	3.2	4.7	B
6.9	3.1	4.9	B
5.8	2.7	5.1	C
6.3	2.9	5.6	C
7.1	3	5.9	C
6.3	3.3	6	C

D=[1.3,6]
GlobalCAIM=0

1.23

2.4

3

2.74

2.67

2.74

3

2.4

1.23

Ejemplo CAIM

A1	A2	A3	Clase
4.7	3.2	1.3	A
5.1	3.5	1.4	A
4.9	3	1.4	A
4.6	3.1	1.5	A
5.5	2.3	4	B
6.4	3.2	4.5	B
7	3.2	4.7	B
6.9	3.1	4.9	B
5.8	2.7	5.1	C
6.3	2.9	5.6	C
7.1	3	5.9	C
6.3	3.3	6	C

D=[1.3,4) [4,6]
GlobalCAIM=3

Ejemplo CAIM

A1	A2	A3	Clase
4.7	3.2	1.3	A
5.1	3.5	1.4	A
4.9	3	1.4	A
4.6	3.1	1.5	A
5.5	2.3	4	B
6.4	3.2	4.5	B
7	3.2	4.7	B
6.9	3.1	4.9	B
5.8	2.7	5.1	C
6.3	2.9	5.6	C
7.1	3	5.9	C
6.3	3.3	6	C

$D=[1.3,4) [4,6]$
GlobalCAIM=3

2

	[1.3,1.4)	[1.4,1.5)	[4,6]	
A	1	3	0	4
B	0	0	4	4
C	0	0	4	4
	1	3	8	12

CAIM=6/3=2

Ejemplo CAIM

A1	A2	A3	Clase
4.7	3.2	1.3	A
5.1	3.5	1.4	A
4.9	3	1.4	A
4.6	3.1	1.5	A
5.5	2.3	4	B
6.4	3.2	4.5	B
7	3.2	4.7	B
6.9	3.1	4.9	B
5.8	2.7	5.1	C
6.3	2.9	5.6	C
7.1	3	5.9	C
6.3	3.3	6	C

D=[1.3,4) [4,6]
GlobalCAIM=3

2
2
2.63
2.89
3.4
4
3.4
2.89

Ejemplo CAIM

A1	A2	A3	Clase
4.7	3.2	1.3	A
5.1	3.5	1.4	A
4.9	3	1.4	A
4.6	3.1	1.5	A
5.5	2.3	4	B
6.4	3.2	4.5	B
7	3.2	4.7	B
6.9	3.1	4.9	B
5.8	2.7	5.1	C
6.3	2.9	5.6	C
7.1	3	5.9	C
6.3	3.3	6	C

D=[1.3,4) [4,5.1)[5.1,6]
GlobalCAIM=4

Ejemplo CAIM

A1	A2	A3	Clase
4.7	3.2	1.3	A
5.1	3.5	1.4	A
4.9	3	1.4	A
4.6	3.1	1.5	A
5.5	2.3	4	B
6.4	3.2	4.5	B
7	3.2	4.7	B
6.9	3.1	4.9	B
5.8	2.7	5.1	C
6.3	2.9	5.6	C
7.1	3	5.9	C
6.3	3.3	6	C

D=[1.3,4) [4,5.1)[5.1,6]
GlobalCAIM=4

3
3
3
3
3
3
3
3
3
3
3

Ejemplo CAIM

A1	A2	A3	Clase
4.7	3.2	1.3	A
5.1	3.5	1.4	A
4.9	3	1.4	A
4.6	3.1	1.5	A
5.5	2.3	4	B
6.4	3.2	4.5	B
7	3.2	4.7	B
6.9	3.1	4.9	B
5.8	2.7	5.1	C
6.3	2.9	5.6	C
7.1	3	5.9	C
6.3	3.3	6	C

$$D = [1.3, 4) [4, 5.1) [5.1, 6]$$

Information Entropy Maximization

Fayyad, U.M., and Irani, K.B. *Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning*, Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence, San Francisco. Morgan Kaufmann, 1022-1027, 1993

- Busca la discretización que **maximice la ganancia de información**
- Es un algoritmo **top-down**
- Es un algoritmo **supervisado**

Information Entropy Maximization

Sigue el mismo procedimiento que CAIM pero maximizando la ganancia de información asociada a una división de un intervalo

$$INFO(C, D|F) = \sum_{i=1}^S \sum_{r=1}^n p_{ir} \log_2 \frac{p_{+r}}{p_{ir}}$$

$$\begin{cases} p_{+r} = p(D_r|F) = \frac{M_{+r}}{M} \\ p_{ir} = p(C_i, D_r|F) = \frac{q_{ir}}{M} \end{cases}$$

Class	Interval					Class Total
	[(d ₀ , d ₁]	...	(d _{r-1} , d _r]	...	(d _{n-1} , d _n]	
C ₁	q ₁₁	...	q _{1r}	...	q _{1n}	M ₁₊
:	:	...	:	...	:	:
C _i	q _{i1}	...	q _{ir}	...	q _{in}	M _{i+}
:	:	...	:	...	:	:
C _s	q _{s1}	...	q _{sr}	...	q _{sn}	M _{s+}
Interval Total	M ₊₁	...	M _{+r}	...	M _{+n}	M

Information Entropy Maximization

Sigue el mismo procedimiento que CAIM pero maximizando la ganancia de información asociada a una división de un intervalo

$$INFO(C, D|F) = \sum_{i=1}^S \sum_{r=1}^n p_{ir} \log_2 \frac{p_{+r}}{p_{ir}}$$

La ganancia de información es la diferencia entre la información de una discretización y la información de la discretización al dividir un intervalo

Information Entropy Maximization

Sigue el mismo procedimiento que CAIM pero maximizando la ganancia de información asociada a una división de un intervalo

- Se toma como punto de división el que produzca mayor ganancia de información (=el que tenga menor información)
- Cada nuevo intervalo de divide con el mismo criterio
- Hay que establecer un criterio para parar de dividir
 - Número preestablecido de intervalos
 - Ganancia de información mínima
 - ...

Ejercicio evaluable

Realiza una memoria y una presentación (3 personas, 30 minutos aprox.) explicando los siguientes algoritmos de discretización

- Maximum entropy
- IEM (Information Entropy Maximization)
- CADD (Class-Attribute Dependence Discretizer)
- CAIM (Class-Attribute Independence Maximization)
- FCAIM (Fast Class-Attribute Independence Maximization)
- PKID (Proportional K-Interval Discretizer)
- FFD (Fixed Frequency Discretizer)

Transformación (numerización)

Consiste en transformar datos nominales a numéricos

- Menos común que la discretización
- Ejemplo,
 - Suspenso –0
 - Aprobado – 1
 - Notable – 2
 - Sobresaliente – 3
 - Matrícula de Honor -- 4

Transformación (numerización)

¿Cuándo utilizar numerización?

- ❑ Cuando el algoritmo no admite datos nominales
 - Regresión
- ❑ Cuando los atributos nominales tienen un orden
 - Bajo, medio, alto
 - Suspenso, aprobado, notable, sobresaliente

Transformación (numerización)

Ventajas

- Se reduce espacio. Ej: apellido \Rightarrow entero
- Se pueden utilizar técnicas más simples

Desventajas

- Se necesita meta-information para distinguir los datos inicialmente no numéricos (la cantidad no es relevante) de los inicialmente numéricos (la cantidad es relevante: precios, unidades, etc.)
- A veces se puede “sesgar” el modelo (*biasing*)

Transformación (normalización)

Cambiar el rango de dato de un atributo de manera que se obtenga cierta uniformidad

- ❑ En general, no es necesario normalizar, o puede ser contraproducente
 - Reglas de asociación, árboles,...
- ❑ En otros casos sí puede ser conveniente normalizar el rango de los atributos
 - Análisis componentes principales,...

Transformación (normalización)

¿Cómo normalizar?

Normalización lineal uniforme (min-max). Se normaliza de forma genérica entre 0 y 1 con la fórmula

$$\mathcal{N}(v) = \frac{v - \min}{\max - \min}$$

- Se puede escoger otros valores que no sean el mínimo o el máximo (no se escala entre 0 y 1)
- Puede dar problemas con outliers

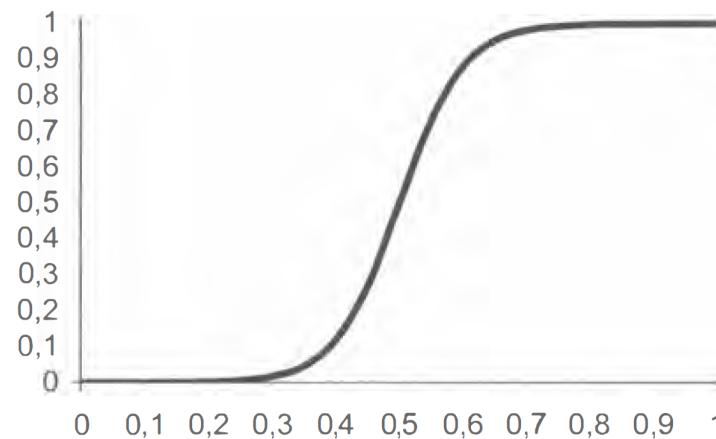
Transformación (normalización)

¿Cómo normalizar?

Escalado sigmoidal (softmax). Se normaliza utilizando una función sigmoidal

$$\mathcal{N}(v) = \frac{1}{1 + e^{-av+b}}$$

$$\mathcal{N}(v) = \text{arcotan}(av + b)$$



Transformación (normalización)

¿Cómo normalizar?

- **Centrado.** Restar la media a cada valor, para que la media sea cero
- **Tipificación.** Centrar y después dividir por la desviación típica, para que la media sea cero y la desviación típica 1.
 - Útil cuando se desconocen los límites o cuando los datos anómalos pueden dominar la normalización min-max

Transformación (normalización)

¿Cómo normalizar?

Normalización decimal. Normaliza moviendo el punto decimal de los valores del atributo. El número de puntos decimales movidos depende del valor absoluto máximo del rango de valores

$$\mathcal{N}(v) = \frac{v}{10^j}$$

donde j es el entero más pequeño que hace que la normalización sea menor que uno. Por ejemplo, para valores entre [-934,345], sería 3

Transformación (pivotamiento)

La operación de pivotamiento cambia filas por columnas

CATEGORÍA	TRIMESTRE	Valencia	León
Refrescos	T1	13.267	3.589
Refrescos	T2	27.392	4.278
Refrescos	T3	73.042	3.780
Refrescos	T4	18.391	3.629
Congelados	T1	150.242	4.798
Congelados	T2	173.105	3.564
Congelados	T3	163.240	4.309
Congelados	T4	190.573	4.812

CATEGORÍA	TRIMESTRE	Refrescos	Congelados
Valencia	T1	13.267	150.242
Valencia	T2	27.392	173.105
Valencia	T3	73.042	163.240
Valencia	T4	18.391	190.573
León	T1	3.589	4.798
León	T2	4.278	3.564
León	T3	3.780	4.309
León	T4	3.629	4.812

Ejemplo. Una tabla de cestas de la compra, donde cada atributo indica si el producto se ha comprado o no.

Objetivo: Ver si dos productos se compran conjuntamente (regla de asociación)

Problema

Es muy costoso: hay que mirar al menos la raíz cuadrada de todas las relaciones (cestas)

Y puede haber millones en una semana...

Sin embargo...productos sólo hay unos 10.000

Transformación (pivotamiento)

Si cambiamos filas por columnas...

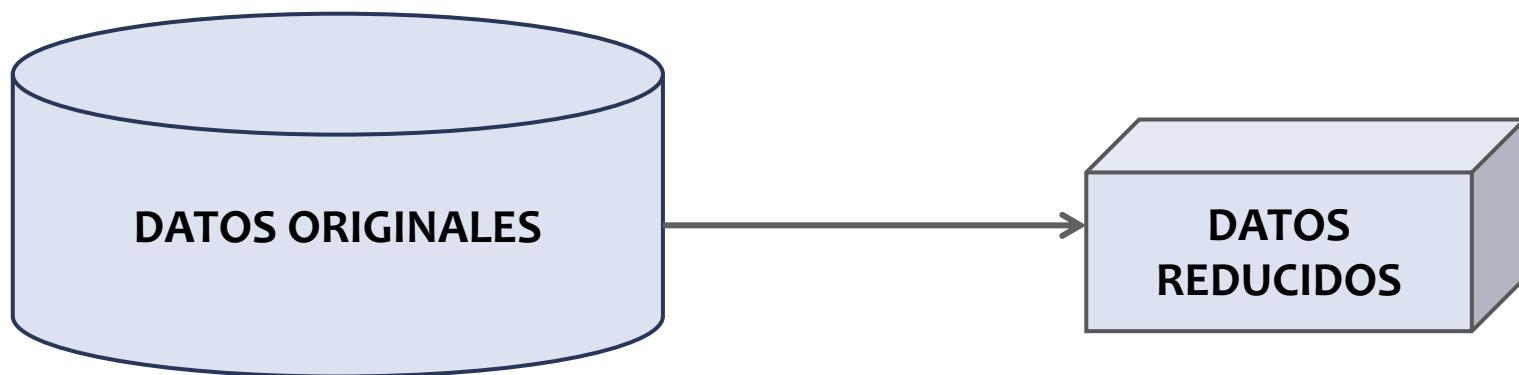
	B1	B2	B3	B4	B5	B6	...
Jabón	X		X				
Huevos		X			X		
Patatas Fritas		X			X		
Champú							
Jabón + Champú	X		X				
Huevos + Patatas							

Sólo es necesario hacer AND entre dos filas para saber si hay asociación

Selección/Reducción

Seleccionar/extraer datos relevantes para la tarea de la minería de datos/extracción de conocimiento

Objetivo: Trabajar con menor cantidad de datos, sin perder (o incluso ganar) eficacia, para mejorar la eficiencia de la técnica de Data Mining



Selección/Reducción

Objetivo: Trabajar con menor cantidad de datos, sin perder (o incluso ganar) eficacia, para mejorar la eficiencia de la técnica de Data Mining

- Discretización
- Selección de características
- Selección de instancias
- Agrupamiento/Compactación

Selección/Reducción (selección de características)

El problema de la selección de características (SC) o variables (*Feature Subset Selection, FSS*) consiste en **encontrar un subconjunto de las variables del problema que optimice la probabilidad de clasificar correctamente**

- Más rápido...
- y a veces, incluso mejor

Selección/Reducción (selección de características)

¿Por qué es necesaria la selección de variables?

- Más atributos no significa más éxito en la clasificación
- Trabajar con menos variables reduce la complejidad del problema y disminuye el tiempo de ejecución
- Con menos variables la capacidad de generalización aumenta
- Los valores para ciertos atributos pueden ser costosos de obtener (o hay muchos y es mejor eliminar el atributo)
- Resultados más simples, más fácil de entender
- Reducir a dos o tres atributos permite visualizar los datos

Selección/Reducción (selección de características)

Var. 1.

Var. 5

Var. 13

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
B	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
C	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
D	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
E	0	1	0	0	0	1	1	0	1	1	0	0	0	0	1	0
F	1	1	1	0	1	1	0	0	1	0	1	0	0	1	0	0

Selección/Reducción (selección de características)

Análisis de correlaciones

Intenta evitar la redundancia de atributos. Un atributo es redundante si puede obtenerse a partir de otros

La correlación entre dos atributos A y B mide la relación lineal entre dos listas de valores

$$r_{A,B} = \frac{\sigma_{A,B}}{\sigma_A \sigma_B} = \frac{\sum_i (A_i - \bar{A})(B_i - \bar{B})}{\sqrt{\sum_i (A_i - \bar{A})^2} \sqrt{\sum_i (B_i - \bar{B})^2}}$$

Selección/Reducción (selección de características)

Entonces:

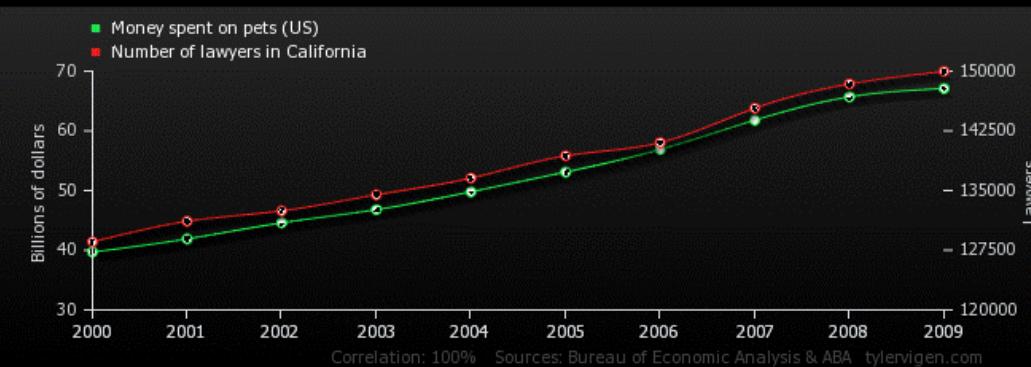
- El coeficiente de correlación está en $[-1,1]$
- Cuanto más se aproxime el valor absoluto a 1 mayor es el grado de conexión entre variables

Una alta correlación no significa necesariamente que haya una dependencia directa entre las variables

Selección/Reducción (selección de características)

Big Data y las relaciones espúreas

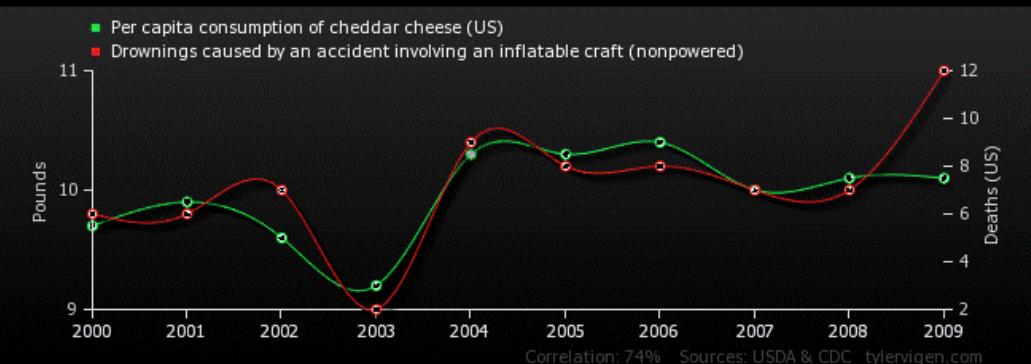
Dinero gastado en mascotas en Estados Unidos y número de abogados en California



Conclusión espúrea:

Es posible que los abogados en California sean contratados como mascotas.

Consumo per cápita de queso cheddar y fallecimientos en accidentes en los que interviene una embarcación inflable sin motor.



Conclusión espúrea:

Nunca te subas a una embarcación inflable si has desayunado queso cheddar.

Selección/Reducción (selección de características)

	Edad	Tensión	Obesidad	Colesterol	Tabaquismo	Alcoholismo	Pulsaciones	Hierro
Edad	1							
Tensión	0.63	1						
Obesidad	0.34	0.22	1					
Colesterol	0.42	0.56	0.67	1				
Tabaquismo	-0.02	0.72	0.72	0.52	1			
Alcoholismo	0.15	0.43	0.32	0.27	0.58	1		
Pulsaciones	0.12	0.27	0.32	0.40	0.39	0.23	1	
Hierro	-0.33	-0.08	0.21	0.45	-0.12	-0.22	-0.15	1

Atributos más relacionados:

- obesidad/tabaquismo
- tensión/tabaquismo
- colesterol/obesidad

se podrían utilizar los que sean más fáciles de obtener con fiabilidad (p.e. tensión y obesidad) y eliminar el tabaquismo para obtener otros modelos

Selección/Reducción (selección de características)

	Edad	Tensión	Obesidad	Colesterol	Tabaquismo	Alcoholismo	Pulsaciones	Hierro
Edad	1							
Tensión	0.63	1						
Obesidad	0.34	0.22	1					
Colesterol	0.42	0.56	0.67	1				
Tabaquismo	-0.02	0.72	0.72	0.52	1			
Alcoholismo	0.15	0.43	0.32	0.27	0.58	1		
Pulsaciones	0.12	0.27	0.32	0.40	0.39	0.23	1	
Hierro	-0.33	-0.08	0.21	0.45	-0.12	-0.22	-0.15	1

algunos parecen independientes como tabaquismo y edad lo que puede sugerir que para predecir el tabaquismo, la edad se puede eliminar

Selección/Reducción (selección de características)

Análisis de componentes principales

- a.k.a, principal component analysis
 - a.k.a., método Karhunen-Loeve
-
- Consiste en reducir la dimensión de los datos
 - Porque muchas veces algunas están correlacionadas
 - Porque si son muchas no podemos visualizarlas
 - Porque muchas variables puede afectar a la eficiencia
 - Introducida por Pearson (s. XIX) y Hotelling (30's)

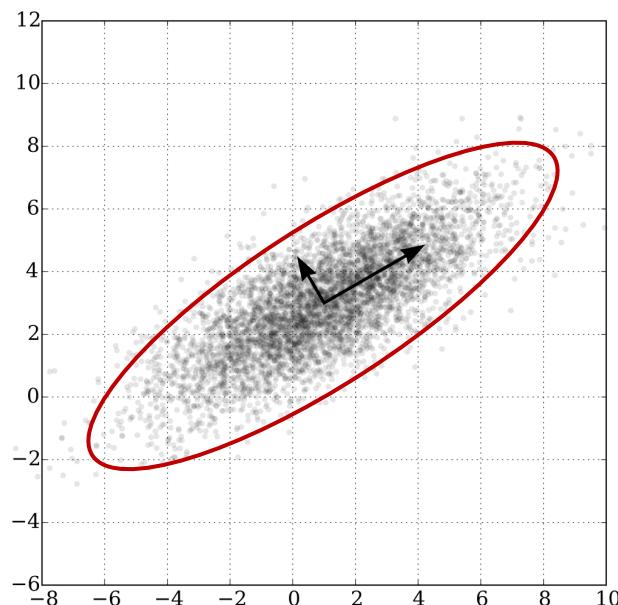
Se trata de encontrar transformaciones lineales normalizadas (SLC) que "resuman" lo mejor posible los datos, capturando la mayor varianza de los mismos

Selección/Reducción (selección de características)

Análisis de componentes principales

Idea Intuitiva. Si se consideran los items como una nube de puntos en \mathbb{R}^n , todos ellos se pueden encerrar en un elipsoide, de centro la media, cuya matriz es la matriz de covarianzas.

Los ejes del elipsoide son un sistema de coordenadas ortogonal. Si realizamos un cambio a este sistema de coordenadas, los puntos se dispersan a lo largo de los ejes



Selección/Reducción (selección de características)

Análisis de componentes principales

La idea es algebraica: una transformación lineal a un espacio con menos dimensiones

$$\phi : \mathbb{R}^m \longrightarrow \mathbb{R}^p \quad p < m$$



Viene determinado por una matriz $A \in \mathcal{M}_{m \times p}$

¿Cómo determinar esa matriz?

Selección/Reducción (selección de características)

Análisis de componentes principales

El cálculo de la matriz se realiza mediante técnicas estadísticas

1. Matriz de correlaciones

- Los valores de las variables no son similares

$$a_{i,j} = \frac{\sigma_{i,j}}{\sigma_i \sigma_j} \quad \text{para } i, j = 1, \dots, m$$

2. Matriz de covarianzas

- Los valores de las variables son similares

$$a_{i,j} = \sigma_{i,j} \quad \text{para } i, j = 1, \dots, m$$

Selección/Reducción (selección de características)

Análisis de componentes principales

En ambos casos, la matriz es simétrica y por tanto, diagonalizable con todos los valores propios reales

$$e_1, e_2, \dots, e_m$$

Calculamos valores propios

$$e_1 \geq e_2 \geq \dots \geq e_m$$

Ordenamos valores propios

$$v_1, v_2, \dots, v_m$$

Calculamos vectores propios

Nos quedamos con los vectores propios de los valores propios dominantes

Selección/Reducción (selección de características)

Análisis de componentes principales

¿Con cuántos nos quedamos?

- Depende de con cuánto valor de la varianza total queramos quedarnos (90% normalmente)
- Por ejemplo,

$$e = (3.65, 0.93, 0.22, 0.13, 0.07)$$



$$d = (3.65, 0.93, 0.22)$$

Los tres primeros suman el 96% de la varianza total

Selección/Reducción (selección de características)

Análisis de componentes principales

ALGORITHM 7.1. Principal Component Analysis

PCA (\mathbf{D}, α):

- 1 $\mu = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$ // compute mean
 - 2 $\mathbf{Z} = \mathbf{D} - \mathbf{1} \cdot \mu^T$ // center the data
 - 3 $\Sigma = \frac{1}{n} (\mathbf{Z}^T \mathbf{Z})$ // compute covariance matrix
 - 4 $(\lambda_1, \lambda_2, \dots, \lambda_d) = \text{eigenvalues}(\Sigma)$ // compute eigenvalues
 - 5 $\mathbf{U} = (\mathbf{u}_1 \quad \mathbf{u}_2 \quad \dots \quad \mathbf{u}_d) = \text{eigenvectors}(\Sigma)$ // compute eigenvectors
 - 6 $f(r) = \frac{\sum_{i=1}^r \lambda_i}{\sum_{i=1}^d \lambda_i}$, for all $r = 1, 2, \dots, d$ // fraction of total variance
 - 7 Choose smallest r so that $f(r) \geq \alpha$ // choose dimensionality
 - 8 $\mathbf{U}_r = (\mathbf{u}_1 \quad \mathbf{u}_2 \quad \dots \quad \mathbf{u}_r)$ // reduced basis
 - 9 $\mathbf{A} = \{\mathbf{a}_i \mid \mathbf{a}_i = \mathbf{U}_r^T \mathbf{x}_i, \text{ for } i = 1, \dots, n\}$ // reduced dimensionality data
-

Selección/Reducción (selección de características)

Análisis de componentes principales

Ejemplo. Base de datos Iris (clasificación de tipos de lirios)

Row ID	Col0	Col1	Col2	Col3	Col4
Row0	5.1	3.5	1.4	0.2	Iris-setosa
Row1	4.9	3	1.4	0.2	Iris-setosa
Row2	4.7	3.2	1.3	0.2	Iris-setosa
Row3	4.6	3.1	1.5	0.2	Iris-setosa
Row4	5	3.6	1.4	0.2	Iris-setosa
Row5	5.4	3.9	1.7	0.4	Iris-setosa
Row6	4.6	3.4	1.4	0.3	Iris-setosa
Row7	5	3.4	1.5	0.2	Iris-setosa
Row8	4.4	2.9	1.4	0.2	Iris-setosa
Row9	4.9	3.1	1.5	0.1	Iris-setosa
Row10	5.4	3.7	1.5	0.2	Iris-setosa
Row11	4.8	3.4	1.6	0.2	Iris-setosa
Row12	4.8	3	1.4	0.1	Iris-setosa
Row13	4.3	3	1.1	0.1	Iris-setosa
Row14	5.8	4	1.2	0.2	Iris-setosa
Row15	5.7	4.4	1.5	0.4	Iris-setosa
Row16	5.4	3.9	1.3	0.4	Iris-setosa

Datos:
Longitud sépalo
Anchura sépalo
Longitud pétalo
Anchura pétalo

Selección/Reducción (selección de características)

Análisis de componentes principales

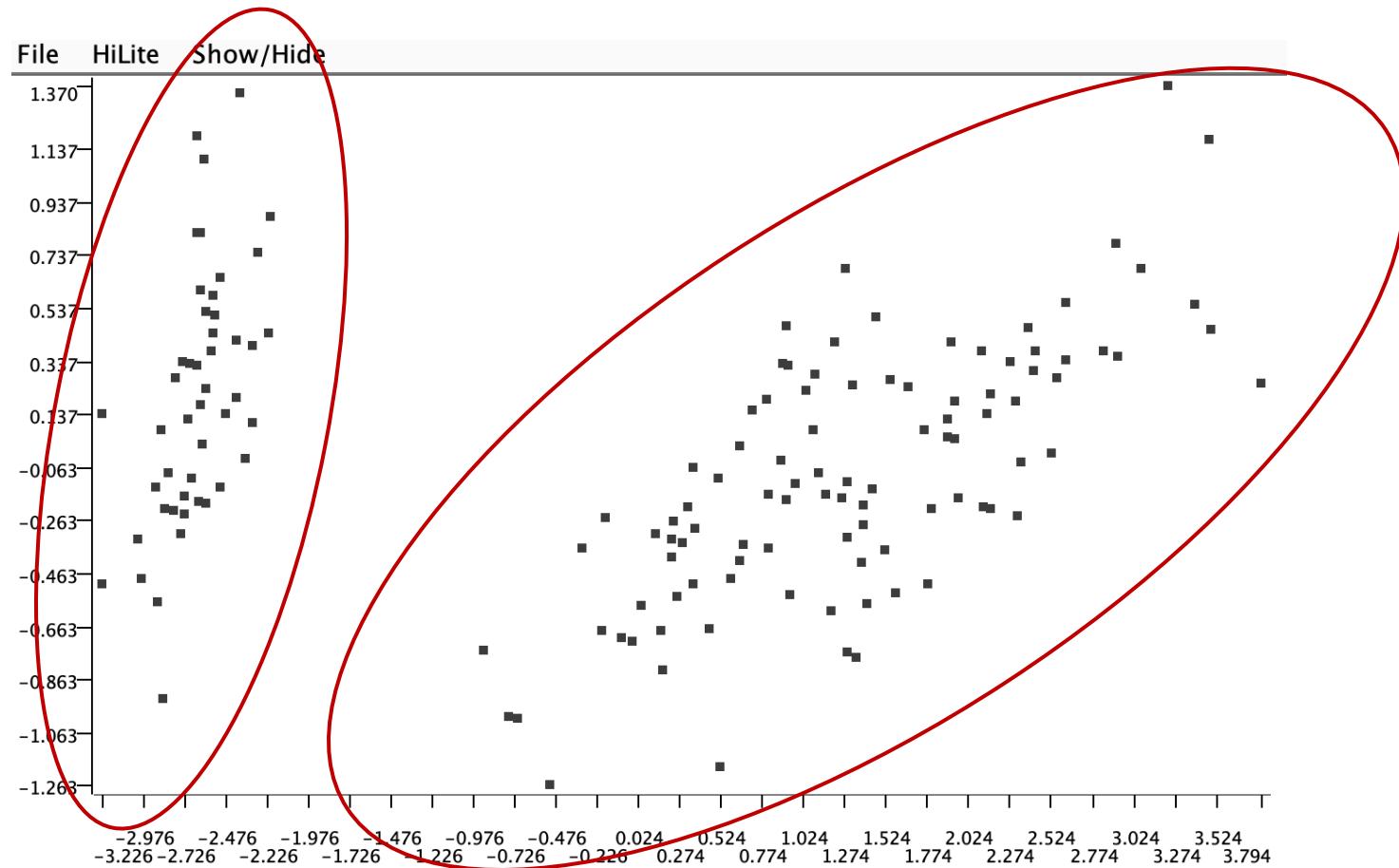
Ejemplo. Base de datos Iris (clasificación de tipos de lirios)

Row ID	Col0	Col1	Col2	Col3	PCA d...	PCA d...
Row0	5.1	3.5	1.4	0.2	-2.684	0.327
Row1	4.9	3	1.4	0.2	-2.715	-0.17
Row2	4.7	3.2	1.3	0.2	-2.89	-0.137
Row3	4.6	3.1	1.5	0.2	-2.746	-0.311
Row4	5	3.6	1.4	0.2	-2.729	0.334
Row5	5.4	3.9	1.7	0.4	-2.28	0.748
Row6	4.6	3.4	1.4	0.3	-2.821	-0.082
Row7	5	3.4	1.5	0.2	-2.626	0.17
Row8	4.4	2.9	1.4	0.2	-2.888	-0.571
Row9	4.9	3.1	1.5	0.1	-2.674	-0.107
Row10	5.4	3.7	1.5	0.2	-2.507	0.652
Row11	4.8	3.4	1.6	0.2	-2.613	0.022
Row12	4.8	3	1.4	0.1	-2.787	-0.228
Row13	4.3	3	1.1	0.1	-3.225	-0.503
Row14	5.8	4	1.2	0.2	-2.644	1.186
Row15	5.7	4.4	1.5	0.4	-2.384	1.345
Row16	5.4	3.9	1.3	0.4	-2.623	0.818

Selección/Reducción (selección de características)

Análisis de componentes principales

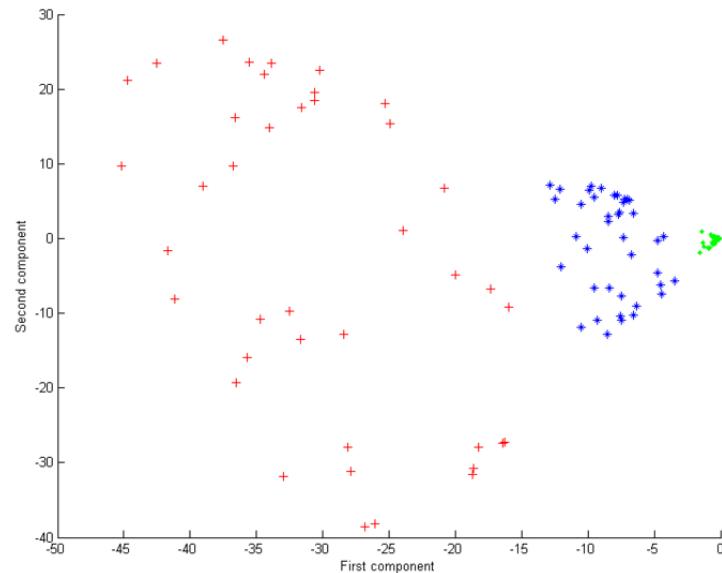
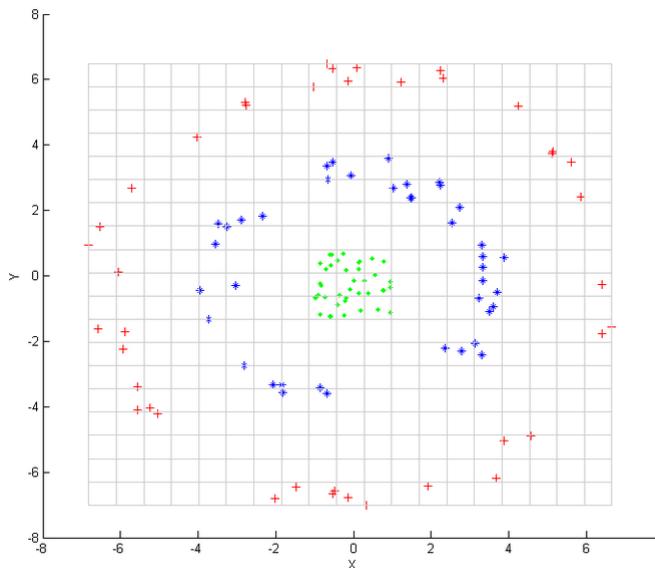
Ejemplo. Base de datos Iris (clasificación de tipos de lirios)



Selección/Reducción (selección de características)

Kernel Principal Components Analysis

Generalización del Análisis de Componentes Principales cuando los atributos no siguen una relación lineal



Selección/Reducción (selección de características)

Selección sin transformación

La selección de atributos se puede ver como un problema de búsqueda de un subconjunto óptimo de atributos

- $C=\{A_1, A_2, A_3\}$
- Posibilidades
 $\{A_1\}, \{A_2\}, \{A_3\}, \{A_1A_2\}, \{A_1A_3\}, \{A_2A_3\}, \{A_1A_2A_3\}$

El espacio de búsqueda crece exponencialmente, 2^n

Búsqueda exhaustiva es impracticable!

Selección/Reducción (selección de características)

Dos formas de evaluar cada subconjunto:

- ❑ Filtro (**Filter**, más rápidas menos exactas)

Se evalúan según la información que contienen

- Correlaciones
- Medidas basadas en Teoría de la Información
- ...

- ❑ Envolvente (**Wrapper**, menos rápidas más exactas)

Se utiliza la técnica de aprendizaje y se evalúan los modelos obtenidos (más exactos)

Selección/Reducción (selección de características)

Búsqueda hacia adelante

Se comienza con el conjunto vacío y, en cada paso, se va añadiendo el “mejor” atributo de los no seleccionados

Búsqueda hacia atrás

Se comienza con el conjunto total de atributos y en cada paso se elimina el “peor” atributo de los que quedan

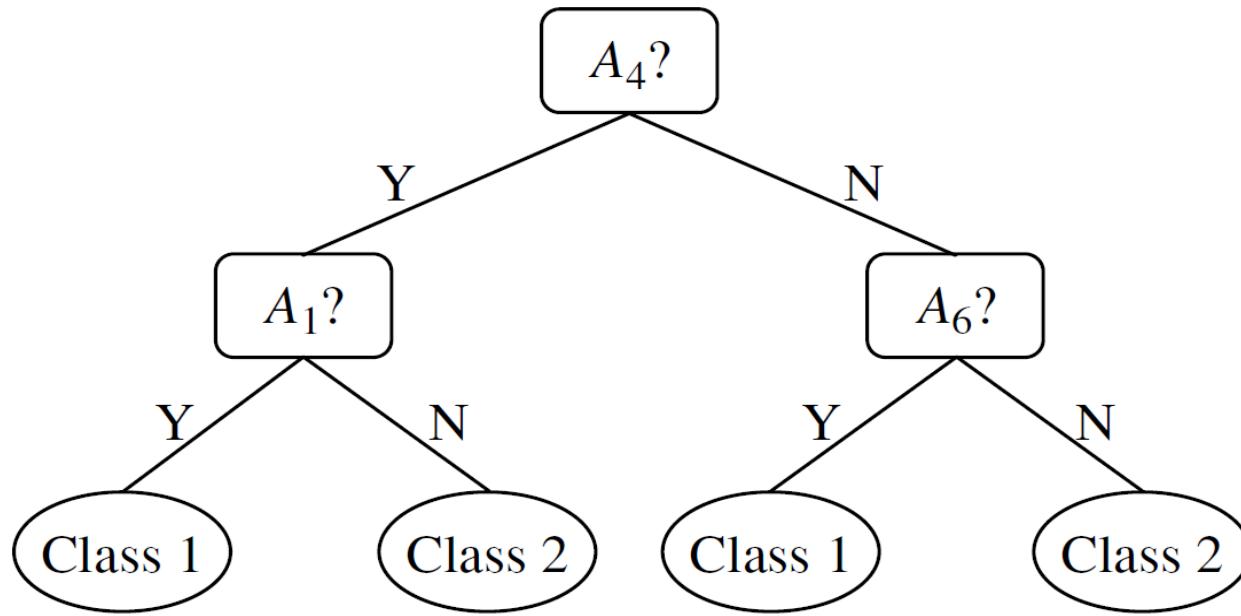
¿Cómo elegir el “mejor” o el “peor”? ¿Cuándo parar?

Selección/Reducción (selección de características)

Mediante árboles de decisión

Initial attribute set:

$$\{A_1, A_2, A_3, A_4, A_5, A_6\}$$



=> Reduced attribute set:
 $\{A_1, A_4, A_6\}$

Los atributos que no aparecen
se consideran irrelevantes

Selección/Reducción (selección de instancias)

Selección horizontal. Se eliminan algunos individuos (filas) del total de los datos recolectados

¿Por qué?

- Porque posiblemente sea **intratable**, tanto en tiempo como espacio, el **considerar todos los datos**
- Porque seleccionando sólo **individuos relevantes** podemos obtener incluso **mejores resultados** (más simples, más fáciles de entender, el algoritmo va mejor,...)
- Porque, normalmente, es **más barato**

Selección/Reducción (selección de instancias)

Muestreo

- ❑ Una muestra es un subconjunto de la población bajo estudio
- ❑ Es un concepto básico para las medidas y técnicas estadísticas
 - Las características de la muestra se extrapolan al total
 - Por ejemplo, encuestas
- ❑ El muestreo es el proceso de seleccionar una muestra de la población

Selección/Reducción (selección de instancias)

Muestreo

Dos situaciones:

Disponemos de toda la población

- ¿Qué cantidad de datos necesito?
- ¿Cómo hacer la muestra? ¿aleatoria? ¿Un mínimo de individuos de cada subgrupo?

Los datos ya son una muestra

- ¿La muestra está balanceada?
- ¿Siguen siendo demasiados datos?

Selección/Reducción (selección de instancias)

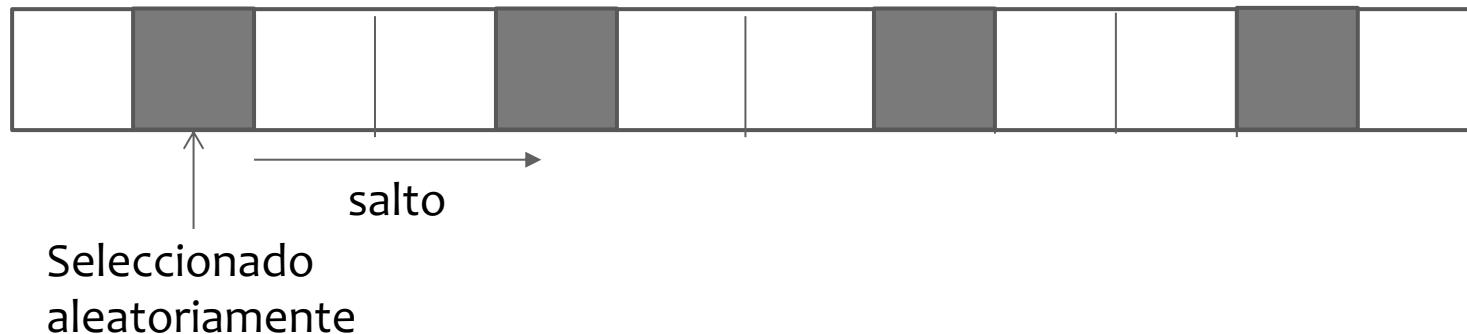
Muestreo

Muestreo aleatorio simple

Se selecciona un número de individuos de forma aleatoria y cada individuo tiene la misma probabilidad de ser elegido

- Sin reemplazo
- Con reemplazo

No vale con escoger los primeros! Puede crear sesgos...



Selección/Reducción (selección de instancias)

Muestreo

Muestreo aleatorio simple

Problema: Seleccionando aleatoriamente siempre se pueden crear sesgos si existen ciertos subgrupos

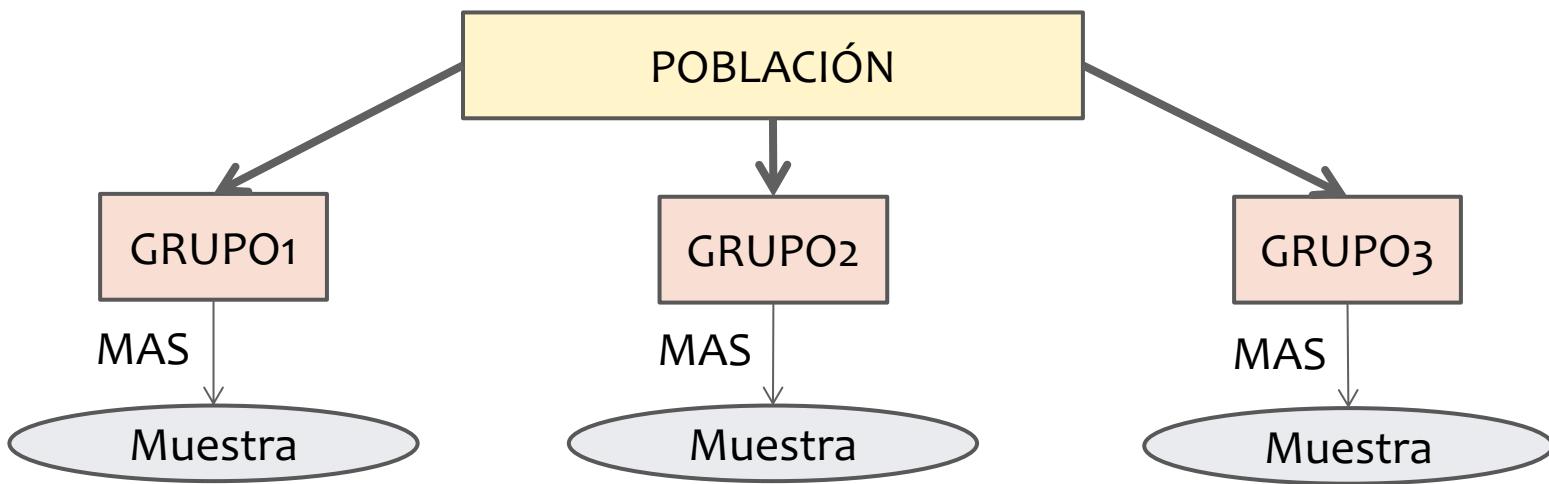
- En una encuesta de intención de voto, seleccionar más votantes de un barrio afín al PP
- En una encuesta sobre abandono escolar, seleccionar más individuos de barrios marginales
- ...

Selección/Reducción (selección de instancias)

Muestreo

Muestreo aleatorio estratificado

Es aquel en el que se divide la población en subpoblaciones o **estratos**, atendiendo a criterios que puedan ser importantes en el estudio. Después realizamos en cada una de estas subpoblaciones muestreos aleatorios simples



Selección/Reducción (selección de instancias)

Muestreo

- Muestreo aleatorio estratificado

¿Cuántos individuos de cada subgrupo?

- **Asignación proporcional.** Se elige un número proporcional al tamaño del subgrupo

$$n_i = n \frac{N_i}{N}$$

The diagram illustrates the formula for proportional allocation. At the bottom is a red box labeled "Tamaño de la muestra". An upward-pointing arrow originates from this box and points to the variable n in the formula. To the right of the formula, two horizontal arrows originate from red boxes labeled "Tamaño del subgrupo" and "Tamaño de la población". These arrows point to the N_i and N terms in the formula, respectively.

Selección/Reducción (selección de instancias)

Muestreo

□ Muestreo aleatorio estratificado

¿Cuántos individuos de cada subgrupo?

- **Asignación óptima.** Se eligen más individuos si:

- El estrato es más grande
- El estrato posee mayor variabilidad interna (varianza)
- El muestreo es más barato en ese estrato

$$n_i = n \frac{N_i \bar{\sigma}_i}{\sum_{j=1}^k N_j \bar{\sigma}_j}$$

$$\bar{\sigma}_i = \frac{1}{N-1} \sum_{j=1}^{N_i} (x_{ij} - \bar{x}_i)$$

Asignación de Neyman

Cuasi-varianza

Selección/Reducción (selección de instancias)

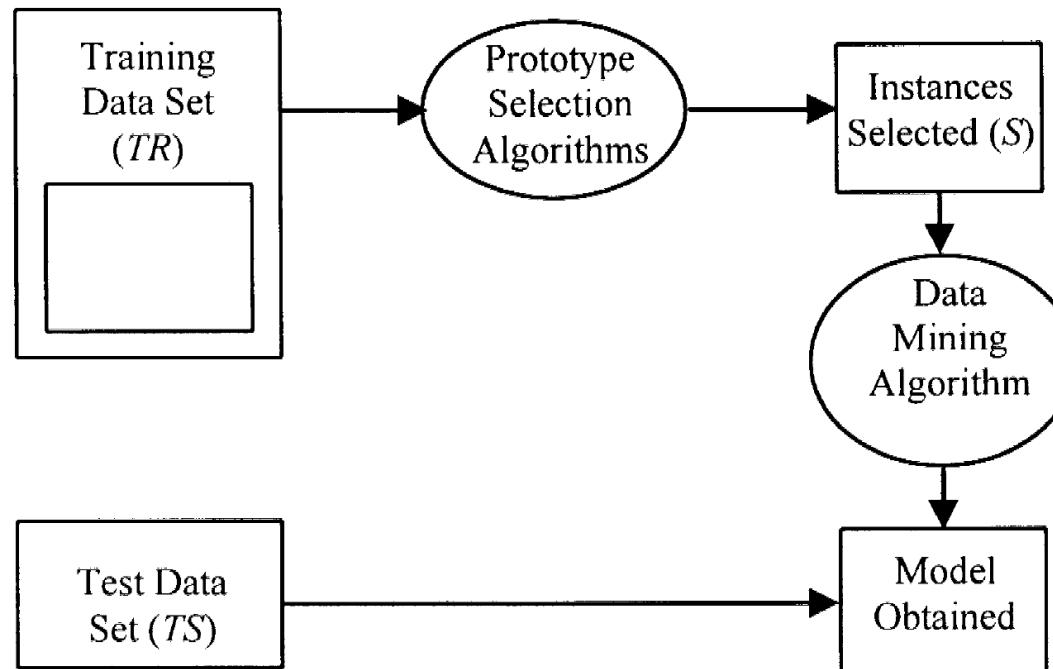
Muestreo

Otros

- Muestreo por agrupamiento
- Muestreo sistemático
- Muestreo doble
- Muestro enlazado
- Muestreo inverso
- Muestreo progresivo

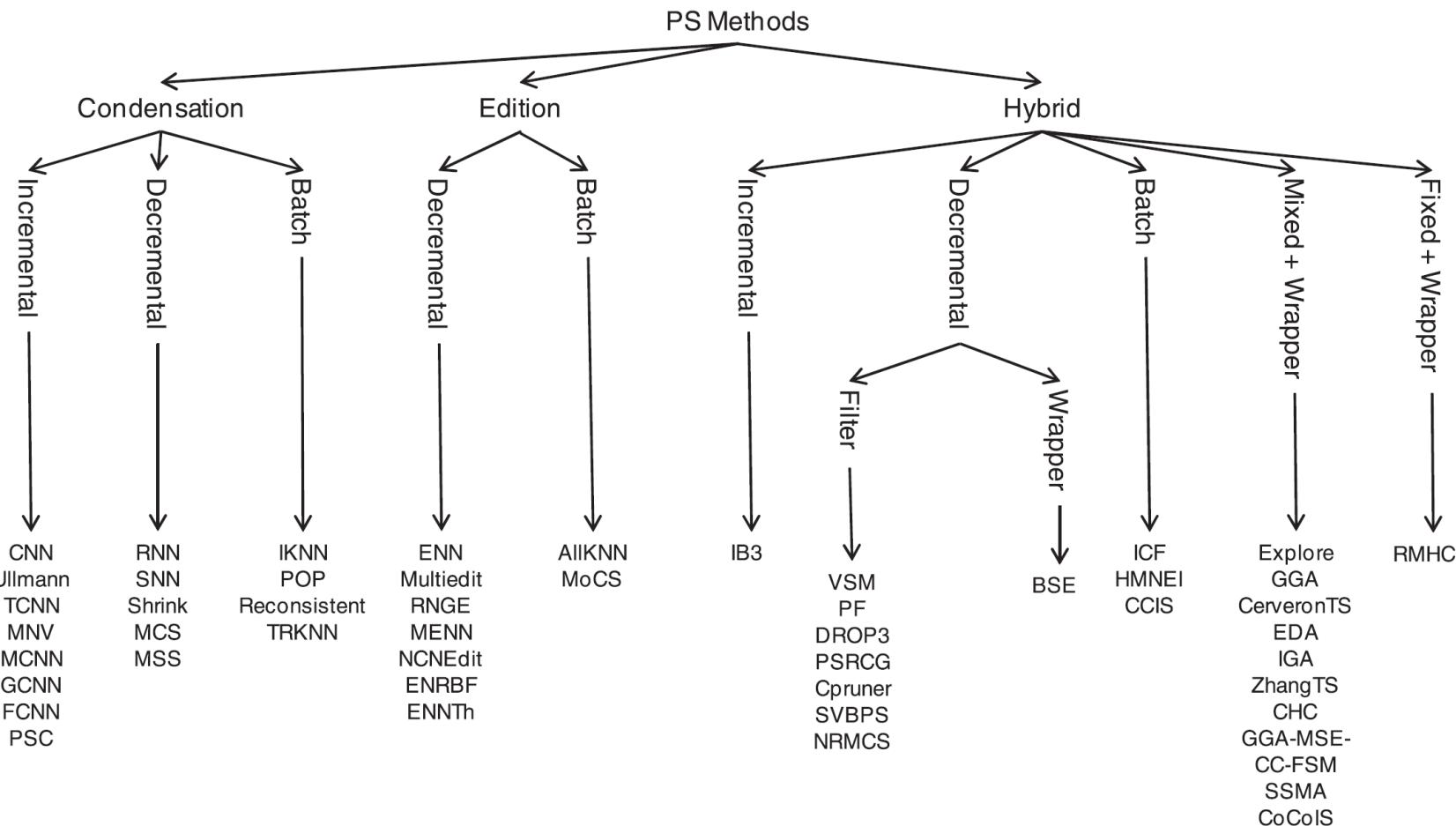
Selección/Reducción (selección de instancias)

Selección de Prototipos/Aprendizaje basado en instancias



sci2s.ugr.es/pr/

Selección/Reducción (selección de instancias)



[doi: 10.1109/TPAMI.2011.142](https://doi.org/10.1109/TPAMI.2011.142) sci2s.ugr.es/pr/

Selección/Reducción (selección de instancias)

Selección de Prototipos/Aprendizaje basado en instancias

Algorithm 1 Condensed k-nearest neighbor

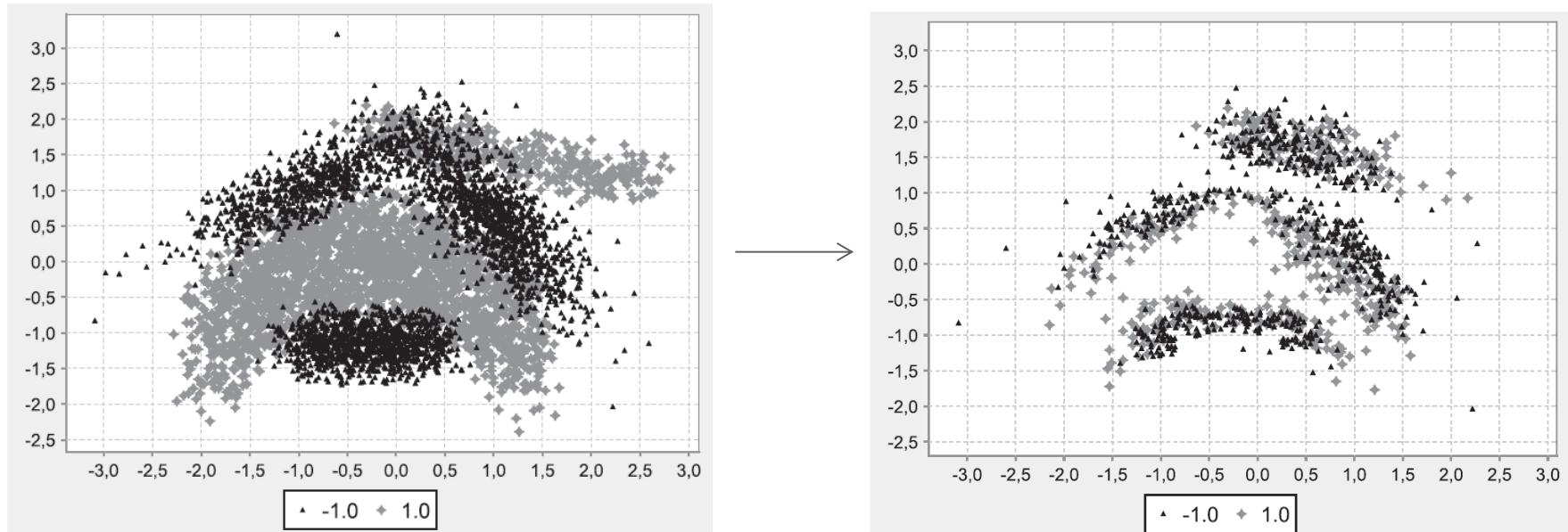
Input: X conjunto total de instancias

Output: S conjunto de prototipos

- 1: $S \leftarrow \emptyset$, seguir \leftarrow TRUE
 - 2: Seleccionar $a_1, \dots, a_t \in X$ cada uno en una clase diferente
 - 3: $S = S \cup \{a_1, \dots, a_t\}$
 - 4: **while** seguir = TRUE **do**
 - 5: seguir \leftarrow FALSE
 - 6: **for** $x \in X \setminus S$ **do**
 - 7: **if** k -NN no clasifica bien a x con S **then**
 - 8: $S \leftarrow S \cup \{x\}$, $X \leftarrow X \setminus \{x\}$
 - 9: seguir \leftarrow TRUE
 - 10: **return** S
-

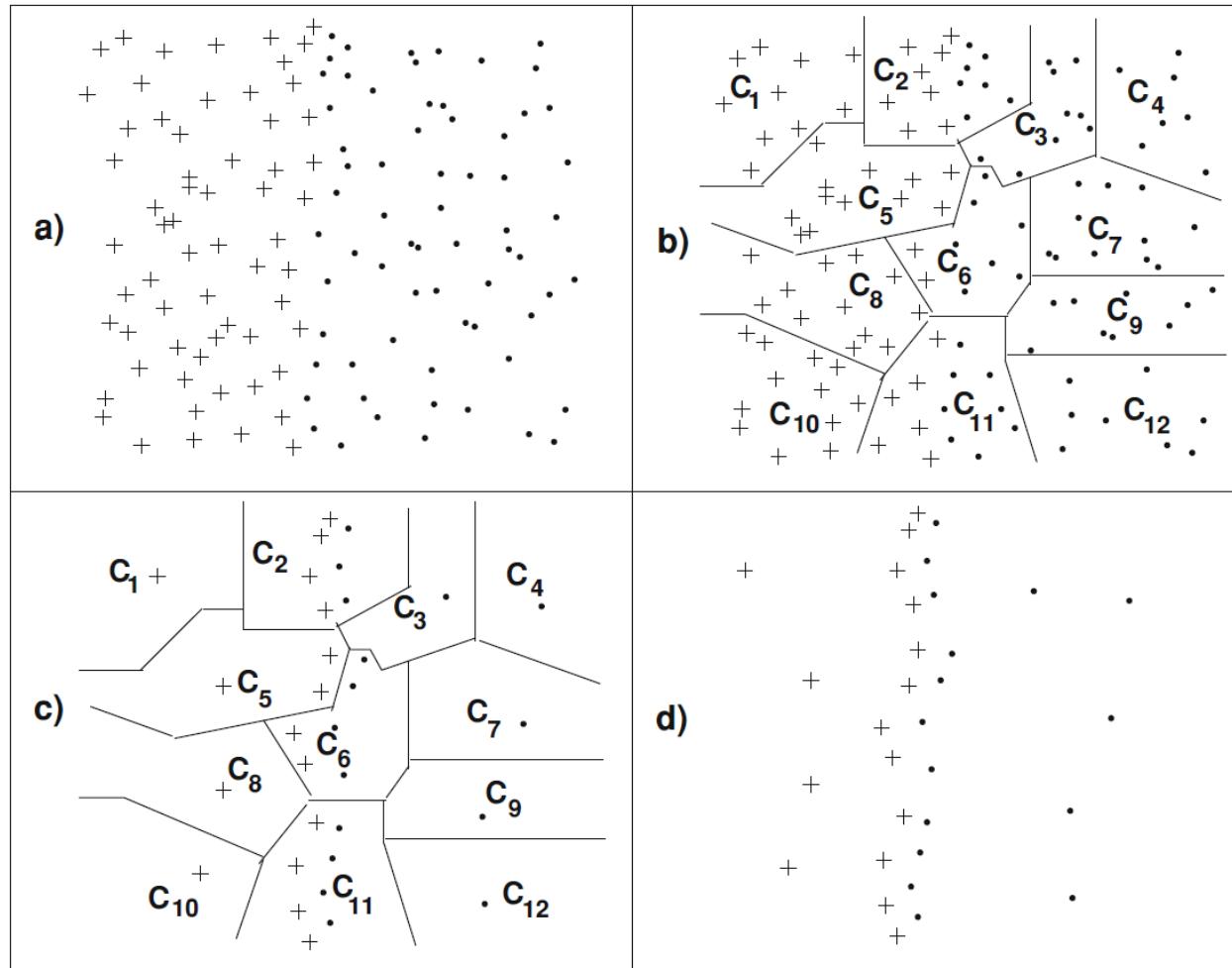
Selección/Reducción (selección de instancias)

Selección de Prototipos/Aprendizaje basado en instancias



[doi: 10.1109/TPAMI.2011.142](https://doi.org/10.1109/TPAMI.2011.142)

Selección/Reducción (selección de instancias)



<https://link.springer.com/article/10.1007/s10044-008-0142-x>

Bibliografía

- ❑ Introducción a la Minería de Datos. José Hernández Orallo, M.José Ramírez Quintana, Cèsar Ferri Ramírez. Pearson, 2004. **Capítulos 3,4 y 5**
- ❑ J. Han, M. Kamber and J. Pei. Data Mining, Second Edition: Concepts and Techniques. Morgan Kaufmann, 2006. **Capítulo 3.**
- ❑ Mohammed J. Zaki, Wagner Meira, Jr., Data Mining and Analysis: Fundamental Concepts and Algorithms, Cambridge University Press, May 2014. **Capítulo 7**
- ❑ Dorian Pyle, Data Preparation for Data Mining, Morgan Kaufmann Publishers, 1999.
- ❑ Salvador García, Julián Luengo and Francisco Herrera, Data Preprocessing in Data Mining, Springer, 2015.

Algunas transparencias y gráficos tomados de:

- <http://sci2s.ugr.es/docencia/in/>
- <http://users.dsic.upv.es/~jorallo/indexcas.htm>

Trabajos evaluables

- ❑ 3 personas, 30 minutos. **Outlier Detection and Analysis.**

Bibliografía:

- Han, Kamber and Pei, Data Mining Concepts and Techniques, 3rd edition, Morgan Kaufmann, 2012. Capítulo 12
- C. Aggarwal, Data Mining: The textbook, Springer, 2015. Capítulos 8 y 9

- ❑ 4 personas, 30 minutos. **Diferentes tipos de muestreo.**

Bibliografía:

- Scheaffer, Mendenhall, Ott and Gerow, Elementary Survey Sampling, Brooks/Cole, Cengage Learning, 2012.
- S. K. Thompson, Sampling, John Wiley & Sons, 2012.
- Levy and Lemeshow, Sampling of Populations. Methods and Applications, 4th Edition, John Wiley & Sons, 2012.
- C. Pérez, Técnicas de Muestreo Estadístico, Ibergarceta Publicaciones S.L, 2010.

Trabajos evaluables

- ❑ 3 personas, 30 minutos. **Algoritmos de discretización.**

Bibliografía:

- Buscar: Maximum entropy, IEM (Information Entropy Maximization), CADD (Class-Attribute Dependence Discretizer), CAIM (Class-Attribute Independence Maximization), PKID (Proportional K-Interval Discretizer), FFD (Fixed Frequency Discretizer)
- S. García, J. Luengo, J.A. Sáez, V. López and F. Herrera, *A Survey of Discretization Techniques: Taxonomy and Empirical Analysis in Supervised Learning*. IEEE Transactions on Knowledge and Data Engineering 25:4 (2013) 734-750, [doi: 10.1109/TKDE.2012.35](https://doi.org/10.1109/TKDE.2012.35)
COMPLEMENTARY MATERIAL to the paper