Universidad de Los Andes
Facultad de Ciencias Económicas y Sociales
Escuela de Estadística

Minería de Datos.

Tema 1: Datos, Tareas y Patrones

Prof. Angel A. Zambrano

Agenda

- Conjuntos de Datos. Tipos.
- Tareas de la Minería de Datos.
- Tipos de Patrones.

Datos, Tareas y Patrones.

La Minería de Datos: Conjunto de técnicas y herramientas aplicadas al proceso no trivial de extraer y presentar conocimiento implícito, previamente desconocido, potencialmente útil y humanamente comprensible, a partir de conjuntos de datos.

Cómo son los Datos (tipos de Conjuntos Datos)

Qué problemas puede resolverse (tareas de la Minería de Datos)

Cómo expresar el Conocimiento (Tipo de Patrones)

Tipos de conjunto de datos.

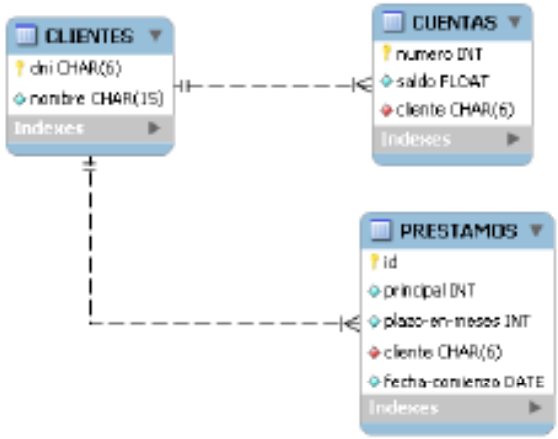
La Minería de Datos puede aplicarse a una gran cantidad de conjunto de datos:

- Almacén de Datos
- Base de Datos Relacionales
- Datos transaccionales
- Hojas de Cálculo
- Multimedia: Imágenes, vídeo y audio
- Espaciales, satelitales,
- Temporales y series de tiempo
- Señales cardiológicas, ecosonogramas, etc
- Documentos: textos y páginas web

Tipos de conjunto de datos.

Almacén de datos

Bases de datos relacionales



Datos transaccionales

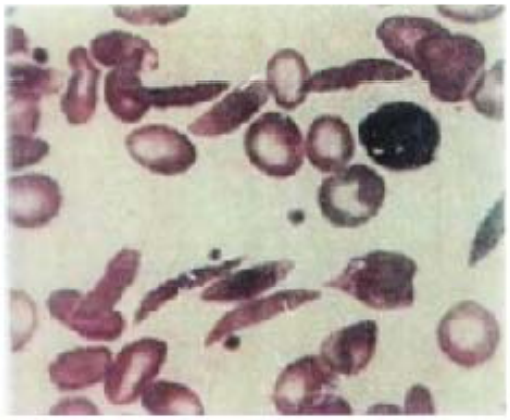
No	Contenido canasta
1	Brócoli, pimienta, maíz
2	Espárragos, calabaza, maíz
3	Maíz, Tomates, frijoles, calabazas
4	Pimienta, maíz, tomates, frijoles
5	Frijoles, espárragos, frijoles, tomates
6	Calabaza, espárragos, frijoles, tomates
7	Tomates, maíz
8	Brócoli, tomates, pimienta
9	Calabaza, espárragos, frijoles
10	Frijoles, maíz
11	Pimienta, brócoli, frijoles, calabaza
12	Espárragos, frijoles, calabaza
13	Calabaza, maíz, espárragos, frijoles
14	Maíz, pimienta, tomates, frijoles, brócoli

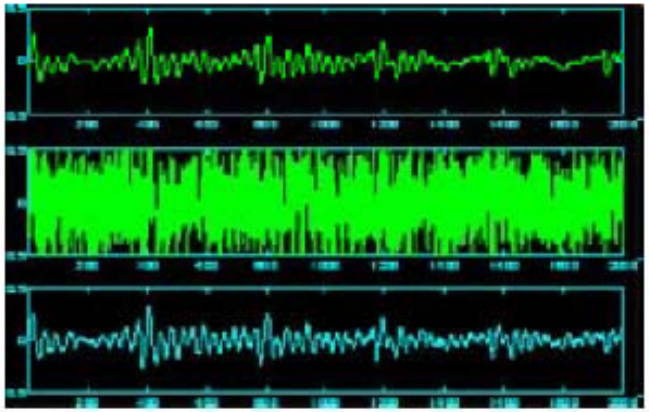
Hojas de cálculo

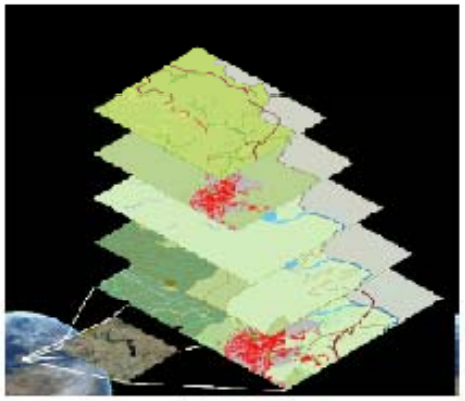
	A	B	C	D	E	F	G	H
1	DULCES DE LA CLASE					Grupo:		
2	COLOR	Naranja	Café	Amarillo	Verde	Rojo	Azul	TOTAL
3	Total de la Clase	40	104	83	67	29	40	363
4	Juan	2	8	5	4	2	3	24
5	María	1	7	8	3	2	3	24
6	Antonio	3	6	4	5	4	3	25
7	Andrea	4	7	4	6	1	2	24
8	Carlos	3	7	3	5	3	3	24
9	Jorge	2	6	7	5	1	1	23
10	Luisa	3	6	6	6	1	3	25
11	Mary	4	7	6	4	2	2	25
12	Guillermo	2	6	7	3	2	3	23
13	Felipe	1	7	8	4	1	4	25
14	Carmen	2	8	4	5	1	3	23
15	Cristina	4	6	6	3	3	3	25
16	Gustavo	3	8	4	5	2	2	24
17	Ana	1	8	5	4	2	4	24
18	Pedro	4	7	6	5	2	1	25

Tipos de conjunto de datos.

Multimedia: Imagenes, audio, vídeo. Espaciales. Temporales. Textos. Sitios web







Tipos de conjunto de datos.

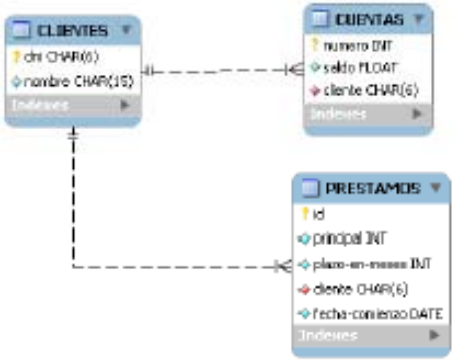
En general los conjuntos de datos hay que convertirlos en una tabla atributo-valor.

Algunos de ellos (multimedia, documentos, espaciales y médicos) hay que pre-procesarlos para convertirlos en tabla atributo-valor

Cada columna de la tabla atributo-valor es una propiedad y se denomina variable, atributo, característica o campo; y cada fila es una colección de atributos que describen un objeto, corresponde a una instancia, ejemplo, registro, caso u observación.

Tipos de conjunto de datos.

En general necesitamos Tabla Atributo-valor

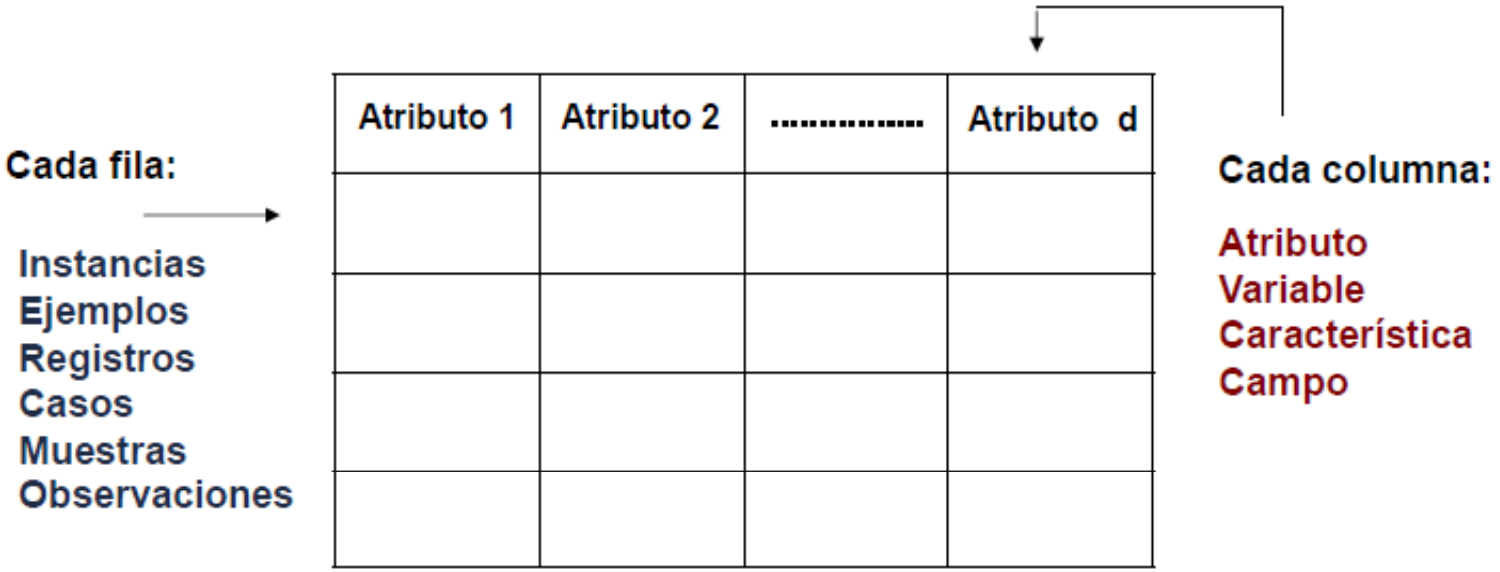


No	Contenido canasta
1	Brócoli, pimienta, maíz
2	Espárragos, calabaza, maíz
3	Maíz, Tomates, frijoles, calabazas
4	Pimienta, maíz, tomates, frijoles
5	Frijoles, espárragos, frijoles, tomates
6	Calabaza, espárragos, frijoles, tomates
7	Tomates, maíz
8	Brócoli, tomates, pimienta
9	Calabaza, espárragos, frijoles
10	Frijoles, maíz
11	Pimienta, brócoli, frijoles, calabaza
12	Espárragos, frijoles, calabaza
13	Calabaza, maíz, espárragos, frijoles
14	Maíz, pimienta, tomates, frijoles, brócoli

DULCES DE LA CLASE							
Color de la Clase	Naranja	Café	Amarillo	Verde	Rojo	Azul	TOTAL
	40	104	83	67	29	40	363
Juan	7	8	5	4	2	3	24
María	1	7	8	3	2	3	24
Antonio	3	6	4	5	4	3	25
Andrea	4	7	4	6	1	2	24
Carlos	3	7	3	5	3	3	24
Jorge	3	6	7	5	1	1	23
Lina	3	6	6	6	1	3	25
Mary	4	7	6	4	2	2	25
Guillermo	2	6	7	3	2	3	23
Felipe	1	7	8	4	1	4	25
Carmen	2	8	4	5	1	3	23
Cristina	4	6	6	3	3	3	25
Gustavo	3	6	4	5	2	2	24
Ana	1	6	5	4	2	4	24
Pedro	4	7	6	5	2	1	25

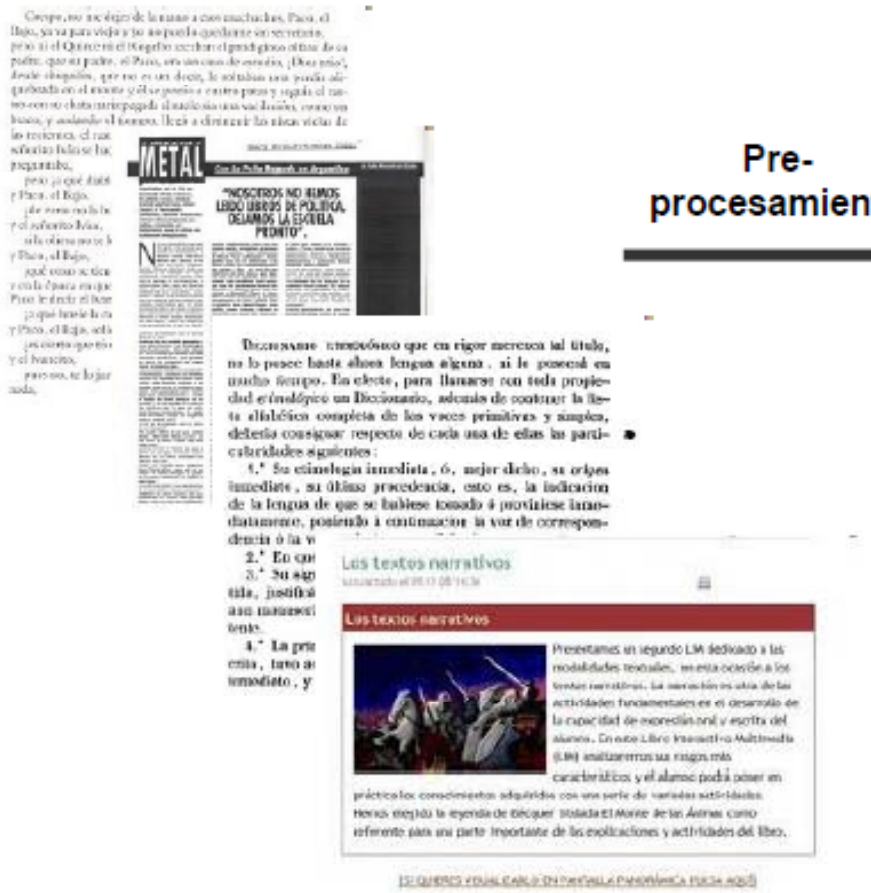
Tipos de conjunto de datos.

Tabla Atributo – Valor:



Tipos de conjunto de datos.

Si los datos son complejos:



Pre-procesamiento

	t_1	...	t_j	...	t_M
d_1	a_{11}	...	a_{1j}	...	a_{1M}
...	...	...	...	...	...
d_i	a_{i1}	...	a_{ij}	...	a_{iM}
...	...	...	...	...	...
d_N	a_{N1}	...	a_{Nj}	...	a_{NM}

Conjunto de datos (Tabla atributo-valor)

Tipos de conjunto de datos.

Cada atributo puede ser de uno de los siguientes tipos de datos:

❖ **Nominales o categóricos**: Los valores son un numero finito de letras, dígitos o palabras:

- Sexo: [M,F]
- Preferencia película: [comedia, acción, suspenso, ciencia ficción]
- Nacionalidad [V, E]
- Categoría Profesor [instructor, asistente, agregado, asociado, titular]

❖ **Binarios**: es un atributo nominal que solo puede tomar dos valores, o (ausencia) Y 1 (presencia)

- Fumador
- Trabaja

Tipos de conjunto de datos.

Cada atributo puede ser de uno de los siguientes tipos de datos:

❖ **Ordinales**: conjunto finito de valores o símbolos que tienen algún orden:

- Satisfacción del cliente: [muy satisfecho, medianamente satisfecho, poco satisfecho]
- Grado académico: [Primaria, Bachiller, Técnico, Universitario, Postgrado]
- Tamaño: [Pequeño, Mediano, Grande, Extra Grande]

❖ **Numéricos**: O cuantitativos, asumen valores enteros o reales:

- Edad
- Temperatura
- Calificación
- Salario

Tareas de la Minería de Datos

Qué tipo de problemas pueden resolverse con la Minería de Datos?

Información del Problema

Lo que se quiere aprender: Función o modelo que explique los datos , con el cual se pueda hacer predicciones válidas para nuevos datos

Tareas de la Minería de Datos

El tipo de conocimiento que se desea extraer va a marcar claramente la *técnica de minería de datos* o *tarea a utilizar*. Según como sea la búsqueda del conocimiento se puede distinguir:

- ***Aprendizaje Supervisado****: se sabe claramente lo que se busca;* generalmente predecir unos ciertos datos o clases.
- ***Aprendizaje No Supervisado:*** *no se sabe lo que se busca, se trabaja con* los datos (hasta que confiesen!).

En el primer caso, los propios sistemas de minería de datos se encargan generalmente de elegir el *algoritmo más idóneo entre los disponibles* para un determinado tipo de patrón a buscar.

Tareas de la Minería de Datos

Tipos de conocimiento:

- **Asociaciones**: Una asociación entre dos atributos ocurre cuando la frecuencia con la que se dan dos valores determinados de cada uno conjuntamente es relativamente alta.
  Ejemplo: en un supermercado se analiza si los pañales y los potitos de bebé se compran conjuntamente.
- **Dependencias**: Una dependencia funcional (aproximada o absoluta) es un patrón en el que se establece que uno o más atributos determinan el valor de otro. Ojo! Existen muchas dependencias nada interesantes (ojo con causalidades inversas).
  Ejemplo: que un paciente haya sido ingresado en maternidad determina su sexo

La búsqueda de asociaciones y dependencias se conoce a veces como análisis exploratorio.

Tareas de la Minería de Datos

Tipos de conocimiento:

- **Clasificación**: Una clasificación se puede ver como el esclarecimiento de una dependencia, en la que el atributo dependiente puede tomar un valor entre varias clases, ya conocidas.

  Ejemplo: se sabe (por un estudio de dependencias) que los atributos edad, número de dioptrías y astigmatismo han determinado los pacientes para los que su operación de cirugía ocular ha sido satisfactoria.

  Podemos intentar determinar las reglas exactas que clasifican un caso como positivo o negativo a partir de esos atributos.

- **Segmentación**: La segmentación (o clustering) es la detección de grupos de individuos. Se diferencia de la clasificación en el que no se conocen ni las clases ni su número (aprendizaje no supervisado), con lo que el objetivo es determinar grupos o racimos (clusters) diferenciados del resto.

Tareas de la Minería de Datos

Tipos de conocimiento:

- **Tendencias**: El objetivo es predecir los valores de una variable continua a partir de la evolución de otra variable continua, generalmente el tiempo.

  Ejemplo, se intenta predecir el número de clientes o pacientes, los ingresos, llamadas, ganancias, costes, etc. a partir de los resultados de semanas, meses o años anteriores.

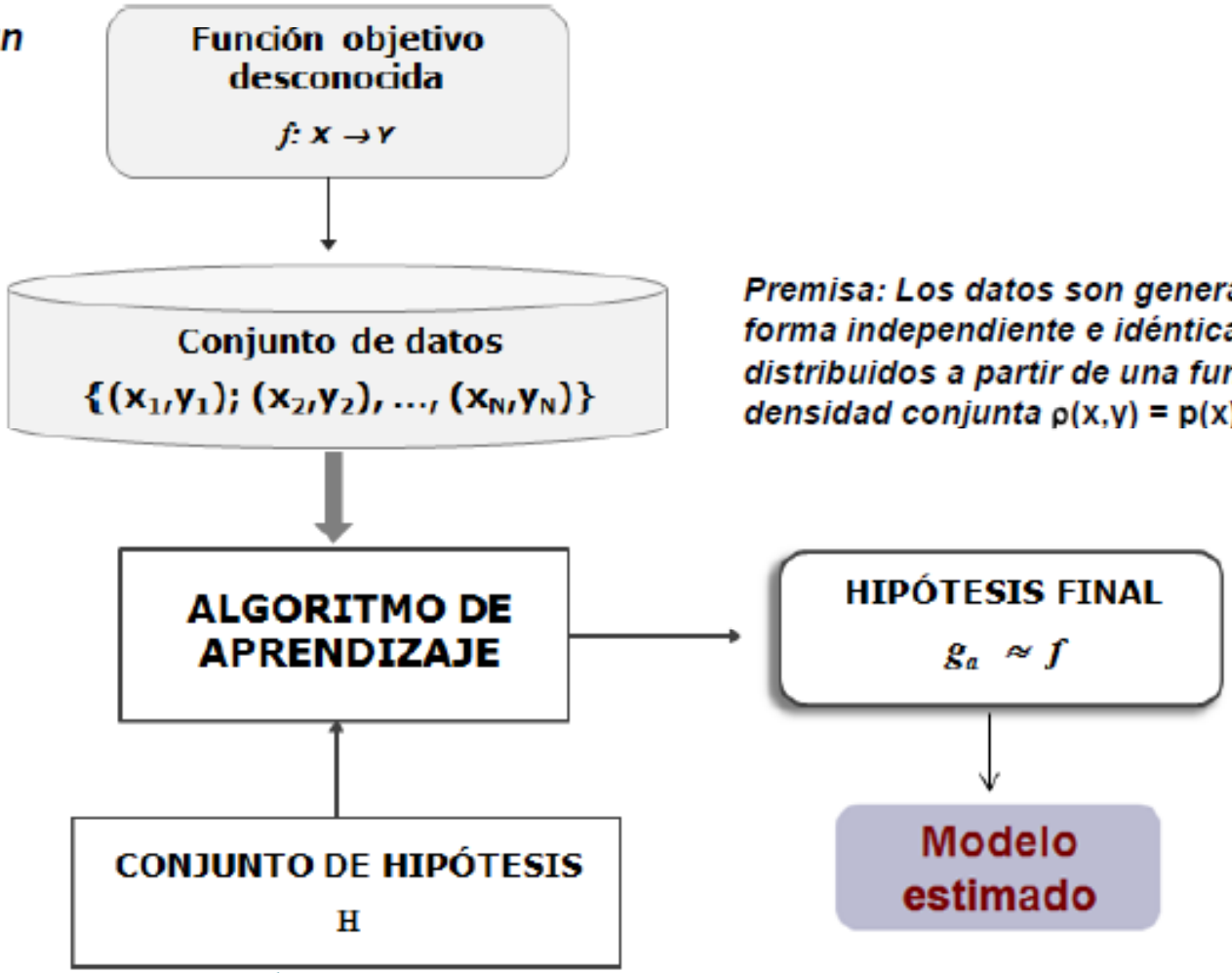
- **Información del Esquema**: (descubrir claves primarias alternativas, R.I.).

- **Reglas Generales**: patrones que no se ajustan a los tipos anteriores. Recientemente los sistemas incorporan capacidad para establecer otros patrones más generales.

Tareas de la Minería de Datos

El aprendizaje a partir de ejemplos = adquisición automática de conocimiento a partir de la experiencia

Elementos de un sistema de aprendizaje:



Premisa: Los datos son generados de forma independiente e idénticamente distribuidos a partir de una función de densidad conjunta $p(x,y) = p(x).p(y|x)$

Tareas de la Minería de Datos

- Los datos constituyen una colección de ejemplos (registros o instancias), (xi,yi) i =1..N, donde cada vector xi pertenece al espacio de entrada (el espacio definido por las variables, atributos o características de entrada), y la salida yi pertenece a (el conjunto de todas las posibles salidas o respuestas).
- La función f: , tal que f(xi) = yi, es la función objetivo o target del aprendizaje.
- Durante la fase de aprendizaje, el algoritmo de aprendizaje tratará de capturar o aprender la relación f(x) entre las variables del sistema, utilizando el conjunto de datos.
- El resultado de este aprendizaje es una hipótesis (modelo), o estimado ga(x) de f(x), seleccionado del conjunto de todas las posibles hipótesis

Tareas de la Minería de Datos

Básicamente, se pueden generar dos tipos de modelos:

Modelos predictivos

(con aprendizaje supervisado)

- **Predicen el valor de un atributo particular basado en los valores de otros atributos.**
- **Estiman valores futuros o desconocidos de variables de interés (= variable dependiente u objetivo), a partir de otras variables (= variables independientes o predictivas).**

Modelos descriptivos

(con aprendizaje no supervisado)

- **Patrones que explican o describen los datos.**
- **Patrones (correlaciones, grupos, anomalías, trayectorias) que sumarizan las relaciones fundamentales de los datos.**
- **Sirven para explorar las propiedades de los datos**

Tareas de la Minería de Datos

Tareas de la minería de datos

Tareas predictivas

- **Clasificación**
- **Regresión**

***Aprendizaje supervisado*:**

Se generan modelos predictivos

Tareas descriptivas

- **Agrupación (*clustering*)**
- **Análisis de asociación (descubrimiento de reglas de asociación)**
- **Detección de anomalías**

Aprendizaje no supervisado:

Se generan modelos descriptivos

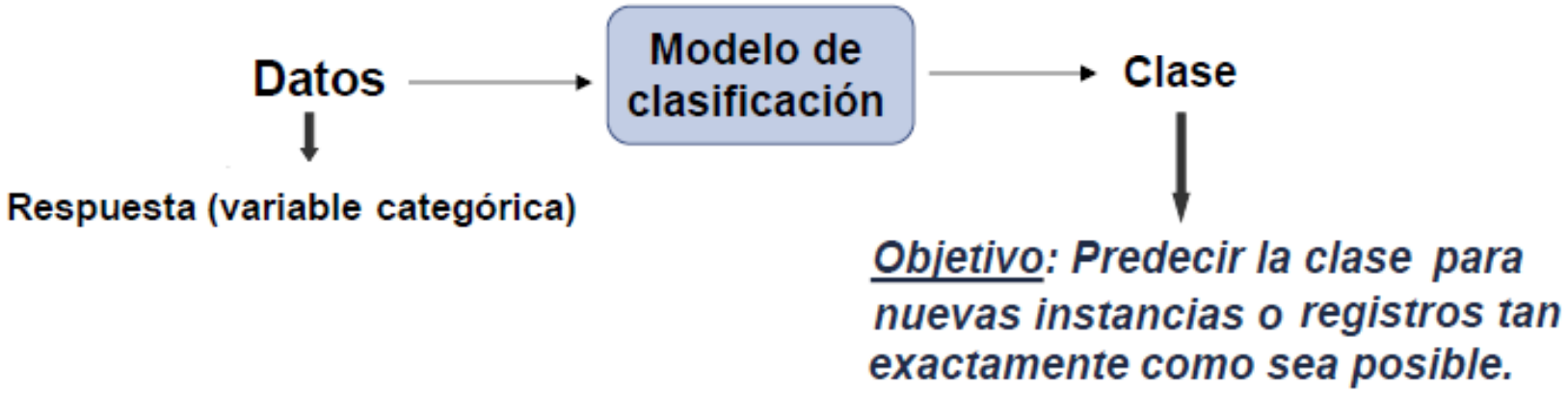
Tareas de la Minería de Datos

A) Clasificación:

- **Dada una colección de registros o instancias**

  Donde cada registro contiene los valores de un conjunto de atributos, uno de los cuales es la clase (cada instancia pertenece a una clase)

- **Encontrar un modelo para el atributo clase como una función de los valores de los otros atributos.**



Tareas de la Minería de Datos

Clasificación: ejemplo

Determinar el rechazo o aceptación de una solicitud de crédito basado en la información financiera y personal del solicitante.

- Conjunto de datos: solicitudes realizadas anteriormente, con la información del cliente y la decisión sobre la aceptación o no de la solicitud por parte de la entidad bancaria.

Edad	Años en el banco	Tarjeta de crédito	Saldo en cuenta	Asignación
30	4	SI	> 2000	SI
19	1	NO	< 2000	NO
⋮				
44	10	SI	< 2000	SI

↑
Clase

- A partir de esta información estimar un modelo para el atributo clase, que permita determinar si dadas las características de un solicitud nueva es viable la concesión del crédito

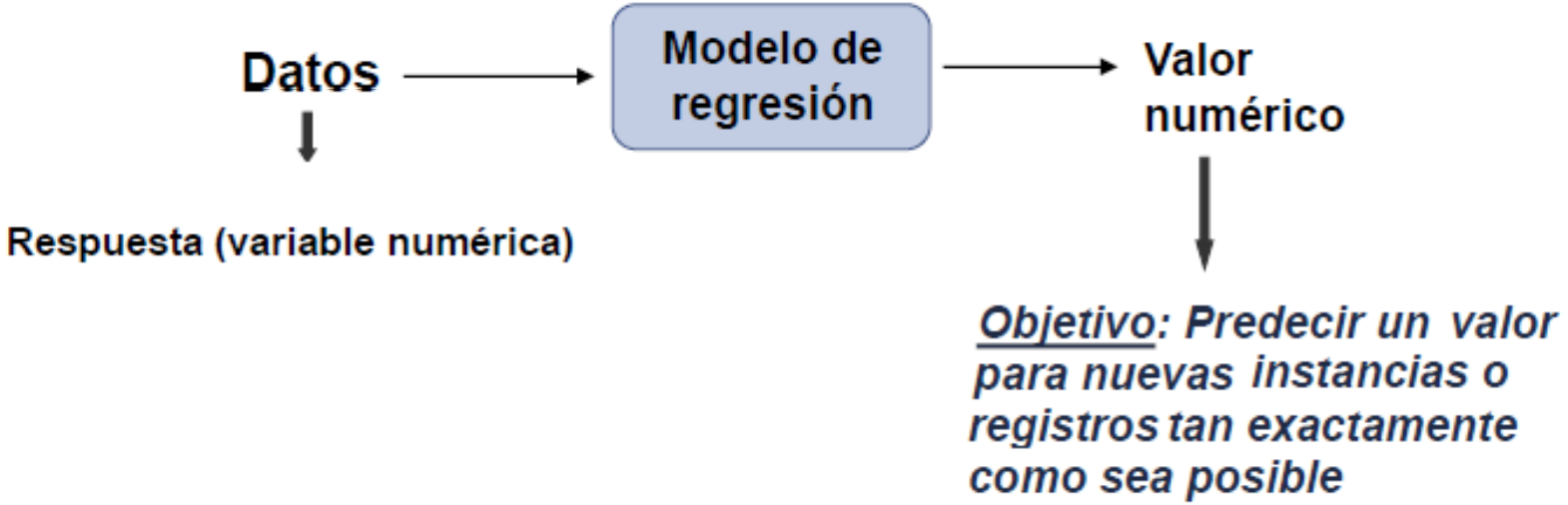
Tareas de la Minería de Datos

B) Regresión:

- Dada una colección de registros o instancias

Donde cada instancia tiene asociado un valor real

- Encontrar un modelo para el atributo real como una función de los valores de los otros atributos.



Tareas de la Minería de Datos

Regresión: ejemplo

Determinar el valor de un inmueble a partir de las características de la zona de ubicación y su arquitectura.

- Conjunto de datos: registros asociados a diferentes tipos de inmuebles en diferentes zonas. A partir de esta información, encontrar un modelo de regresión que permita estimar el valor de un nuevo inmueble.

No. de habitaciones	No. de baños	índice de criminalidad	Índice de acceso a TP	. . . .	Valor (x 1000 Bs.F)
3	2	0.06	1		1000
1	1	0.01	2		600
2	1	0.80	1		250
3	3	0.40	3		400
.					

Atributo de Salida

Tareas de la Minería de Datos

C) Agrupación (*clustering*):

- Dada una colección de registros o instancias

  Donde cada registro tiene asociado un conjunto de atributos, no hay salida definida.

- Encontrar grupos naturales a partir de los datos

- Objetivo: Los objetos de un grupo son muy similares entre sí y muy diferentes a los objetos de otros grupos.

  Se utilizan medidas de similitud, que dependerán del tipo de variable presente en el conjunto de datos

Tareas de la Minería de Datos

Agrupación: ejemplo

Determinar diferentes tipos de documentos basados en su contenido

- Conjunto de datos: conjunto de documentos preprocesados para transformarlos en una tabla donde cada columna representa la frecuencia relativa de términos claves en los documentos.

Doc.	Términos										
	Industria	Mercado	Trabajo	País	Inflación	Precio	Salud	....	Droga	Vacuna	Médico
1	4	2	3	1	0	0	0	....	0	0	0
2	2	3	0	2	2	0	0	....	0	0	0
3	0	2	1	1	0	2	0	....	0	0	0
4	0	0	0	0	0	0	0	....	2	3	0
5	0	0	0	0	0	0	3	....	3	0	2

Utilizando una medida de similitud adecuada, encontrar grupos de documentos que son similares entre si basados en los términos importantes que aparecen en ellos, para *posteriormente* identificar el tema de un nuevo documento.

Tareas de la Minería de Datos

D) Análisis de asociación:

- Dada una colección de registros o instancias

Donde cada registro contiene los valores de un conjunto de atributos o items, no hay salida definida

- Encontrar combinaciones o asociaciones de items (atributos) que ocurren frecuentemente.

- Objetivo: descubrir patrones que describen características fuertemente asociadas en los datos. Identificar relaciones no explícitas entre atributos categóricos.

Regla de asociación:

"Si el atributo X toma el valor A entonces el atributo Y toma el valor B"

Tareas de la Minería de Datos

Reglas de asociación: ejemplo

Manejo de los estantes de un supermercado

- Conjunto de datos: registros de las compras de los clientes provenientes de los puntos de ventas.

- Identificar items que son comprados juntos por un número suficiente de clientes, para mejorar la organización física del almacén.

Transacción 1	Pan, Leche
Transacción 2	Pan, Servilletas, Cerveza, Huevos
Transacción 3	Leche, Servilletas, Cerveza, Agua
Transacción 4	Pan, Leche, Servilletas, Cerveza
Transacción 5	Pan, Servilletas, Cerveza, Refresco
⋮	

Una regla clásica sería:

Si el cliente compra pan y servilletas muy probablemente comprará cerveza

Tipos de Patrones

¿Cómo se representa el Conocimiento?

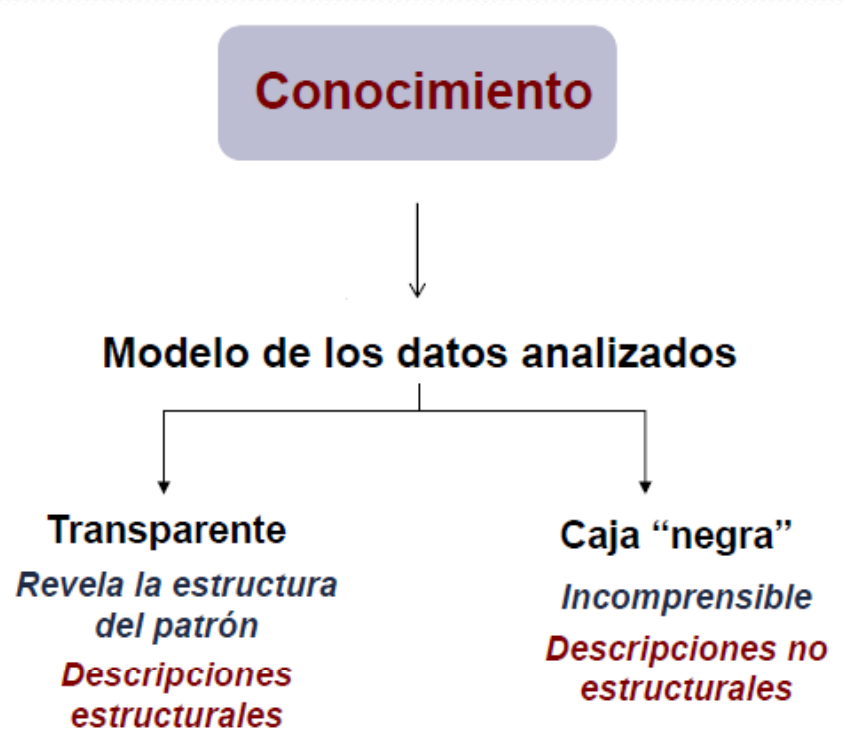
En un sistema de aprendizaje:

- La función objetivo f(x) y el conjunto de datos son dictados por el problema a resolver.
- El algoritmo de aprendizaje y el espacio de hipótesis, son herramientas de solución de libre elección (modelo de aprendizaje)
- El conjunto de hipótesis □ se especifica a través de una forma funcional o forma de representación que se utiliza para describir la dependencia f(x).
- Por ejemplo, el estimado ga(x) puede representarse mediante un conjunto de reglas, una función polinomial, una red neuronal, un árbol de decisión o una sumatoria de funciones núcleo, entre otras.

Tipos de Patrones

¿Cómo se representa el Conocimiento?

Dependiendo de la Técnica de Aprendizaje o de minería de datos existen diferentes lenguajes de representación para expresar el conocimiento extraído



Tipos de Patrones

Ejemplo: Sistema de apoyo al diagnóstico en el área de oftalmología

Conjunto de datos:

Clase ↓

ID	EDAD	PRESCRIPCIÓN	ASTIGMATISMO	PRODUCCIÓN DE LÁGRIMAS	RECOMENDACIÓN
1	Joven	Miopía	No	Reducida	Ninguno
2	Joven	Miopía	No	Normal	Blandos
	⋮	⋮	⋮	⋮	⋮
9	Pre-presbicia	Miopía	No	Reducida	Ninguno
10	Pre-presbicia	Hipermetropía	No	Normal	Blandos
	⋮	⋮	⋮	⋮	⋮
23	Presbicia	Miopía	Si	Normal	Duros
24	Presbicia	Hipermetropía	No	Normal	Blandos

Tipos de Patrones

Descripciones estructurales

- Reglas de clasificación:

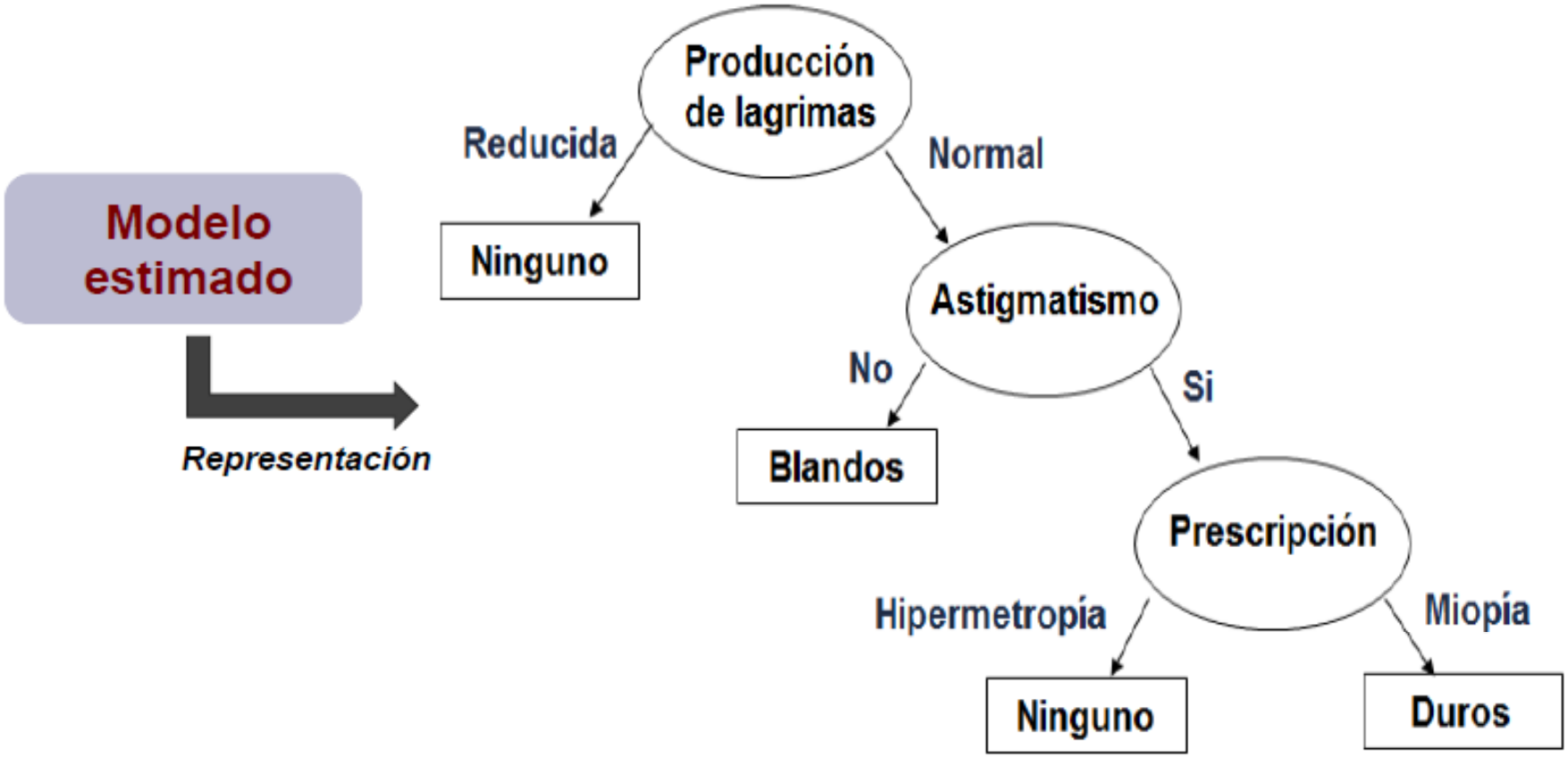
Modelo estimado

Representación

```
If astigmatismo = no
   and produccion de lágrimas= normal
   and prescripción= hipermetropía then blandos
If astigmatismo = no
   and produccion de lágrimas= normal
   and edad = joven then blandos
If edad = pre-presbicia
   and astigmatismo = no
   and produccion de lágrimas= normal then blandos
If astigmatismo = si
   and produccion de lágrimas= normal
   and prescripción= myope then duros
If edad = joven
   and astigmatismo = si
   and produccion de lágrimas= normal then duros
If produccion de lágrimas= reducida then ninguno
If edad = presbicia
   and produccion de lágrimas= normal
   and prescripción= miopía
   and astigmatismo = no then ninguno
If prescripción= hipermetropía
   and astigmatismo = si
   and edad = pre-presbicia then ninguno
If edad = presbicia
   and prescripción= hipermetropía
   and astigmatismo = si then ninguno
```

Tipos de Patrones

- *Árboles de decisión:*



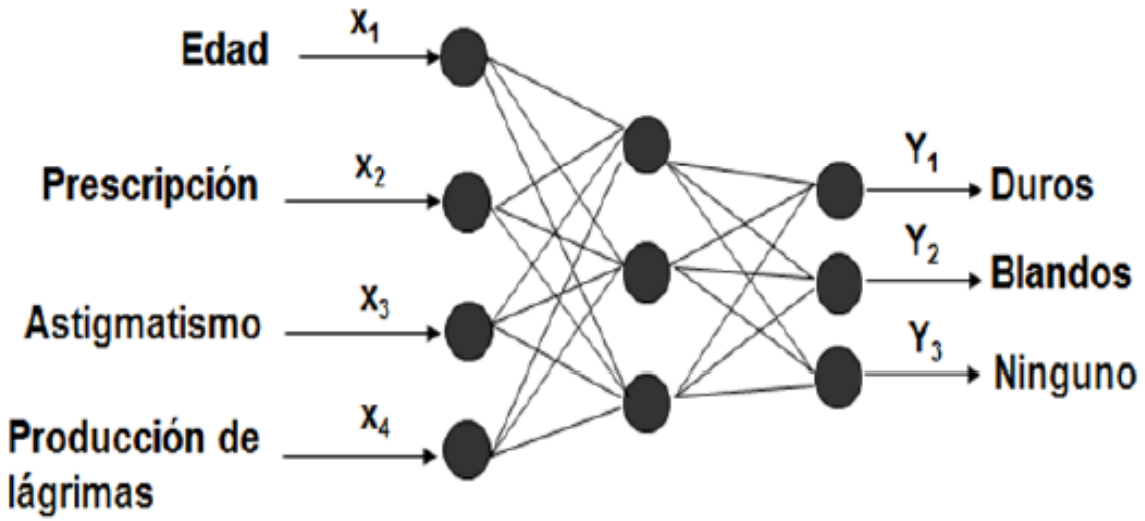
Tipos de Patrones

Descripciones no estructurales:

- Redes neuronales:

Modelo estimado

Representación



$$Y_1 = f(z_1w_{11} + z_2w_{12} + z_3w_{13} + b_1)$$

$$z_1 = f(x_1w_{11} + x_2w_{12} + x_3w_{13} + x_4w_{14} + b_1)$$

$$z_2 = f(x_1w_{21} + x_2w_{22} + x_3w_{23} + x_4w_{24} + b_2)$$

$$z_3 = f(x_1w_{31} + x_2w_{32} + x_3w_{33} + x_4w_{34} + b_3)$$

Tipos de Patrones

El conocimiento expresado por medio de cualquier lenguaje de representación debe ser:

- Válido: El modelo debe predecir el tipo de planta para una muestra desconocida con exactitud
- Novedoso: Aporta conocimiento desconocido para el usuario.
- Potencialmente útil: Resuelve el objetivo o problema planteado.
- Comprensible: Es posible su interpretación

Tarea 1.

1. Para los siguientes problemas, indique la tarea de minería de datos más adecuada que se le pueda aplicar:
   a) Identificar patrones de comportamiento de usuarios en la web.
   b) Determinar la calidad de agua de un repositorio basado en indicadores ecológicos.
   c) Identificar brotes de gripe a partir de los mensajes de una red social (twitter).
   d) Determinar las características físicas de la tierra para un determinado cultivo.
   e) Determinar las preferencias de compra de los clientes en una tienda virtual (comercio electrónico).
   f) Indicar la cantidad de lluvia a corto plazo a partir de datos climatológicos.
   g) Determinar la categoría de un sitio web.
   h) Determinar las fallas de un servicio de telefonía a partir de registros de los usuarios.
   i) Determinar el costo de un nuevo contrato en una compañía a partir de los costos correspondientes a contratos anteriores.
   j) Determinar el rendimiento promedio de un estudiante promedio a partir de datos socio económicos.
2. Proponga un ejemplo para cada caso de las tareas de minería de datos vistas en clase.