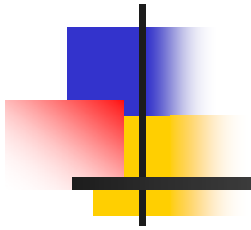


Minería de Datos

Profra. Heidy Marisol Marin Castro
Universidad Politécnica de Victoria

Que es un conjunto de datos?

Es una colección de objetos con sus respectivo atributos.

- Un **atributo** es una propiedad o característica de un objeto.
  Ejemplos: color de ojos de una persona, peso, salario anual, etc.

- Un **atributo** también es conocido como **variable o característica**

Una colección de atributos describe un objeto.

- Un objeto es también llamado registro, caso,muestra, entidad, o instancia.

Atributos

Objetos

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Tipos de datos

¿Que tipos de datos se pueden utilizar en la minería de datos?

- En general podemos encontrar tres clasificaciones dependiendo de como se encuentre almacenada la información: estructurada, semi-estructurada y no estructurada.
  - ➢ Las bases de datos relacionales

EMPLOYEE

Fname	Minit	Lname	Ssn	Bdate	Address	Sex	Salary	Super_ssn	D
John	B	Smith	123456789	1965-01-09	731 Fondren, Houston, TX	M	30000	333445555	
Franklin	T	Wong	333445555	1955-12-08	638 Voss, Houston, TX	M	40000	888665555	
Alicia	J	Zelaya	999887777	1968-01-19	3321 Castle, Spring, TX	F	25000	987654321	
Jennifer	S	Wallace	987654321	1941-06-20	291 Berry, Bellaire, TX	F	43000	888665555	
Ramesh	K	Narayan	666884444	1962-09-15	975 Fire Oak, Humble, TX	M	38000	333445555	
Joyce	A	English	453453453	1972-07-31	5631 Rice, Houston, TX	F	25000	333445555	
Ahmad	V	Jabbar	987987987	1969-03-29	980 Dallas, Houston, TX	M	25000	987654321	
James	E	Borg	888665555	1937-11-10	450 Stone, Houston, TX	M	55000	NULL	

DEPARTMENT

Dname	Dnumber	Mgr_ssn	Mgr_start_date
Research	5	333445555	1988-05-22
Administration	4	987654321	1995-01-01
Headquarters	1	888665555	1981-06-19

DEPT_LOCATIONS

Dnumber	Dlocation
1	Houston
4	Stafford
5	Bellaire
5	

Tipos de datos

- **Bases de datos relacional**: colección de relaciones (tablas), donde cada relación contiene atributos (columnas o campos) y tuplas (registros o filas). Cada tupla representa un objeto que se describe a través de sus atributos y posee un clave única o primaria.enter

Desde las técnicas de minería de datos se manejan 2 tipos de atributos:

- Atributos numéricos: contienen valores enteros o reales. Ejemplos: salario, edad
- Atributos categóricos o nominales: toman valores en un conjunto finito y preestablecido de categorías. Ejemplos: sexo, nombre del depto.(gestión, marketing, ventas)

Tipos de datos

Los siguientes casos entran en la categoría de bases de datos pero cuyo contenido requiere un tratamiento especial.

- **Bases de datos espaciales** mantienen información relacionada a espacios físicos cuyos datos pueden ser geográficos, redes de transporte, información de tráfico, etc., donde la minería de datos podría encontrar patrones que permitan la construcción de nuevos caminos o líneas del metro.
- **Bases de datos temporales** mantienen información relacionada al tiempo ya sea instantes específicos o intervalos temporales. Aquí la minería de datos podría encontrar tendencias climatológicas.

Tipos de datos

- **Bases de datos documentales** que pueden poseer datos de los tres tipos (estructurados, semi-estructurados y no estructurados) y donde la minería de datos podría utilizarse para encontrar asociaciones entre contenidos o clasificación de objetos.
- **Bases de datos multimedia** además de las técnicas de minería de datos se requieren algoritmos de búsqueda eficiente y almacenamiento sobre este tipo de formatos.

Tipos de datos

- La Internet es el repositorio de información más grande donde la información prevaleciente es semi-estructurada y en la mayoría de los casos no estructurada. La minería web generalmente se utiliza para realizar:
  - ***Minería del contenido*** la cual pretende encontrar patrones en el contenido de las paginas web.
  - ***Minería de la estructura*** entendiendo por estructura los hipervínculos y URLs.
  - ***Minería del uso*** para encontrar patrones de preferencias entre los usuarios de un sitio web y poder adecuar el sitio a sus necesidades.

Tipos de modelos

Modelos predictivos

- Un ***modelo predictivo***: se entrena (estima) un modelo usando los datos recolectados para hacerpredicciones futuras. Nunca es 100% precisa y lo que más importa es el rendimiento del modelo cuando es aplicado a nuevos datos..
- Un ***modelo descriptivo*** sirve para identificar patrones que permiten explorar las propiedades de los datos examinados no para predecir sino para describir futuros datos. Este modelo permite descubrir las características más importantes de la BD

Tareas en data mining

- Regresion (Predictiva)
- Classificacion (Predictiva)
- Classificacion No supervisada – Clustering (descriptiva)
- Reglas de Asociacion (descriptiva)
- Deteccion de Outliers (descriptiva)
- Visualizacion (descriptiva)

Minería de datos y el proceso de descubrimiento del conocimiento (KDD)

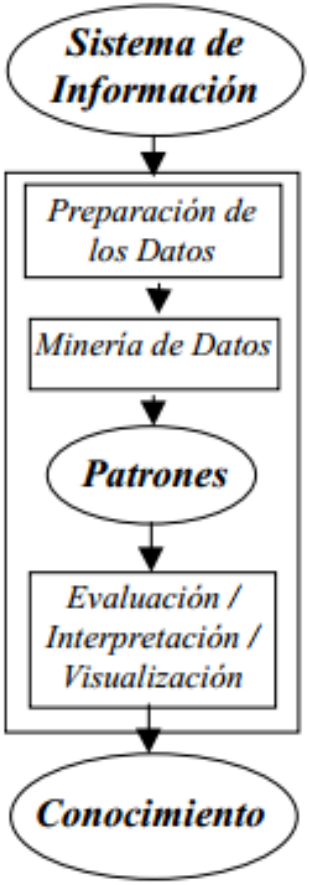
Fayyad et al. 1996, **KDD**: proceso no trivial de identificar patrones validos, novedosos, potencialmente útiles y comprensibles a partir de los datos.

- Valido: los patrones deben seguir siendo precisos para nuevos datos
- Novedoso: que se aporte algo desconocido tanto para el sistema como para el usuario
- Potencialmente útil: las acciones deben reportar algún beneficio
- Comprensible: fácilmente de interpretar y validar para la toma de decisiones

Minería de datos y el proceso de descubrimiento del conocimiento (KDD)

Minería de datos: el corazón del proceso de descubrimiento de conocimiento

Bases de datos

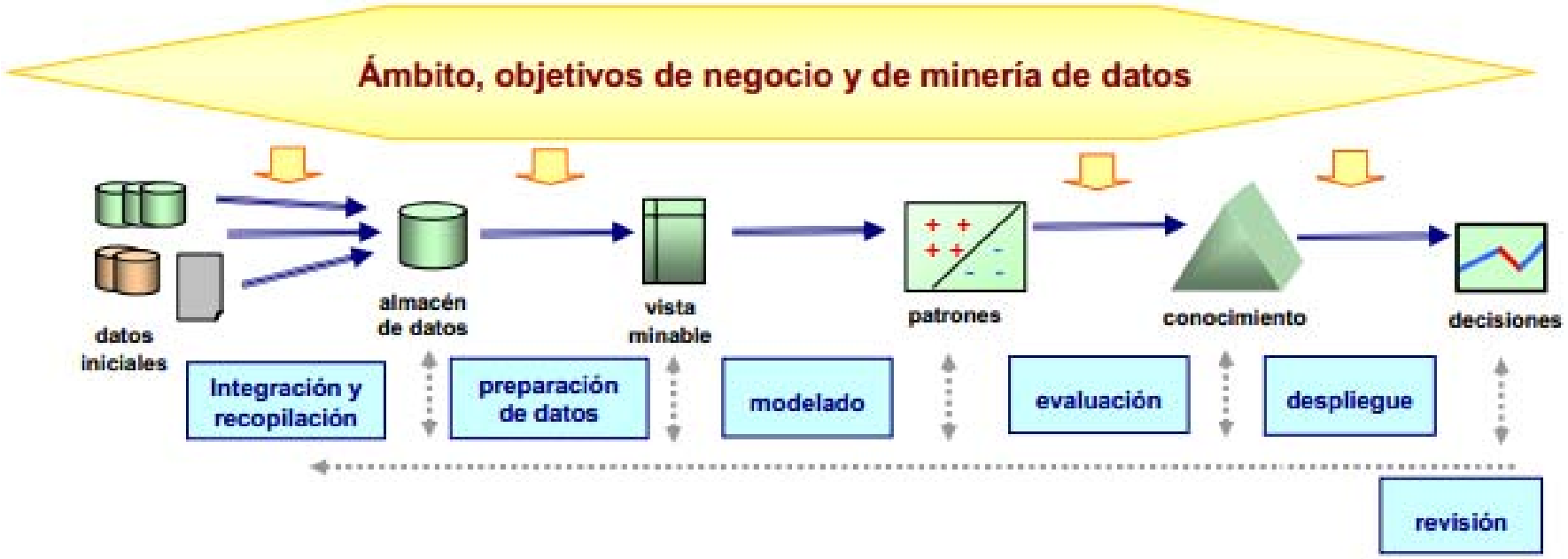


KDD

Data Warehouse (almacén de datos)

1. **Integración de datos**
2. **Selección de datos**
3. **Limpieza de datos**
4. **Transformación de datos**

Proceso KDD


Ámbito, objetivos de negocio y de minería de datos
datos iniciales
almacén de datos
vista minable
patrones
conocimiento
decisiones
Integración y recopilación
preparación de datos
modelado
evaluación
despliegue
revisión

Proceso KDD

1. **Integración de datos** – combinar multiples fuentes
2. **Limpieza de datos** – remover ruido e inconsistencia de datos
3. **Selección de datos** – datos relevantes a la tarea de análisis son recuperados a partir de la BD
4. **Transformación de datos** – los datos son transformados a un formato apropiado para minería de datos
5. **Minería de datos-** proceso escencial donde métodos inteligentes son aplicados para extraer patrones de datos
6. **Evaluación –** identificar los patrones interesantes representando conocimiento o medidas de interes
7. **Presentación del conocimiento**-tecnicas de visualizacion y representacion del conocimiento son usadas para presenter el conocimiento minado

Processo KDD

- Las primeras fases del KDD determinan que las fases sucesivas sean capaces de extraer conocimiento válido y útil a partir de la información original.
- Generalmente, la información que se quiere investigar sobre un cierto dominio de la organización se encuentra:
  - En bases de datos y otras fuentes **muy diversas**
  - El análisis posterior será mucho más sencillo si la fuente es unificada, accesible (interna) y desconectada del trabajo transaccional

Fase de integración y recopilación

- El primer paso : la integración de múltiples bases de datos en almacenes de datos (data warehousing).
- Un **almacén de datos** es un conjunto de datos históricos, internos o externos, que describen un contexto o área de estudio.
- Un **almacén de datos** se encuentra organizado de manera que permite aplicar eficientemente las herramientas para resumir, describir y analizar los datos.
- Un **almacén de datos** generalmente maneja modelos de tipo multidimensional.

Fase de integración y recopilación

- En un **modelo multidimensional** los datos se organizan en torno a hechos que poseen ciertos atributos o medidas que pueden verse con cierto detalle dependiendo de las dimensiones.

  Medidas o atributos → cuánto.

  Dimensiones → cuándo, qué, dónde.
- Un **almacén de datos** es deseable pero no imprescindible.
- Se puede trabajar con formatos heterogéneos.

  Archivos de texto.

  Hojas de cálculo.

  Bases de datos.

Fase de selección, limpieza y tranformación

Proceso de selección

La selección de características reduce el tamaño de los datos eligiendo las variables más influyentes en el problema, sin sacrificar la calidad del modelo de conocimiento obtenido del proceso de minería.

- Métodos para la selección de características :
  - Basados en la elección de los mejores atributos del problema
  - Los que buscan variables independientes mediante tests de sensibilidad, algoritmos de distancia o heurísticos.

Fase de selección, limpieza y tranformación

Proceso de limpieza

- Al obtener la información desde distintas fuentes se debe procurar que datos sobre el mismo objeto se unifiquen.
- Posibles errores:
  - Dos o más datos de diferentes individuos se mezclan → nuevos individuos que pueden ocasionar ruido en el modelo.
  - Dos o más fuentes del mismo individuo se replican → ocasiona menos ruido pero si es recurrente llevara a resultados inesperados.
- Soluciones:
  - Identificar patrones similares durante el proceso de mezclado. Por ejemplo:
  - {hombre, varón, masculino} → hombre

Fase de selección, limpieza y tranformación

Proceso de limpieza

- Identificar posibles formatos de claves y unificarlos. Por ejemplo:

  {55−58−22−35−71, 55−5822−3571, 58223571, +525558223571}→ 55 – 58223571

  Donde 55 es la lada del estado y 582 de la región o zona.

- Detección de valores faltantes.

Posibles problemas si no se aplica limpieza de datos:

- El método de minería de datos puede no tratar correctamente estos datos.
- Los valores faltantes pueden ocasionar malos cálculos de totales, medias.

Fase de selección, limpieza y tranformación

Proceso de limpieza

- Es importante entender la posible causa que origina el dato faltante. Cuestiones a considerar:
  - La falta de valores puede expresar características relevantes. Ej.: La falta de número celular pues no todos poseen.
  - Valores no existentes. Ej.: Quizá el registro -usuario, cliente- es nuevo y aun no tiene historial.
  - Datos incompletos. Ocasionado por el proceso de mezcla.

Limpieza de datos

- Una vez analizado la causa del dato faltante, se pueden dar los siguientes posibles tratamientos:
  - Ignorar tales datos.
  - Eliminar el campo (Solo si la cantidad de valores nulos es muy alta).
  - Filtrar la fila (Puede sesgar los datos).
  - Reemplazar el valor (Ya sea manualmente o automáticamente en
  - funciones de otros objetos similares).
  - Esperar los datos faltantes (Puede retrasar el proyecto).
  - Detección de valores erróneos o anómalos.

Fase de minería de datos

- Entender el problema que se desea resolver.
- Determinar el tipo de modelo a aplicar.
  - Predictivo.
  - Descriptivo.
- Elegir el algoritmo de minería que resuelva la tarea y/o obtenga el modelo.

Fase de minería de datos

Tareas de la minería de datos:

- **Clasificación (Tarea predictiva)**

Cada instancia pertenece a una clase distinguida por un tipo de atributo. Los demás atributos de la instancia se utilizan para predecir la clase de nuevas instancias.

Ejemplo: Un oftalmólogo desea determinar cuáles de sus nuevos clientes son candidatos a una cirugía ocular y cuales no, basado en los resultados de sus clientes anteriores y la evolución de éstos después de la cirugía. Algunos factores a considerar son el tipo de enfermedad ocular y algunos padecimientos que pueden afectar la cirugía.

El modelo final clasificaría a los nuevos pacientes como operables o no.

Fase de minería de datos

Tareas de la minería de datos:

Regresión (Tarea predictiva)
Consiste en aprender una función real que asigna a cada instancia un valor real, de manera que el objetivo es minimizar el error entre el valor predicho y el valor real.
Ejemplo: Una constructora desea determinar el costo adecuado para los departamentos que va a construir y vender en una determinada zona. Para ello la constructora se puede basar en el costo de las viviendas cercanas a la zona.
El modelo ayudaría a predecir el costo factible de sus nuevos deptos.

Tareas de la minería de datos:

Agrupamiento (Tarea descriptiva)

- A diferencia de la clasificación, aquí se busca agrupar a los individuos maximizando el grado de similitud entre las instancias de un mismo grupo y minimizando la similitud entre los distintos grupos. Los grupos pueden ser o no disjuntos.

Ejemplo: Una librería que ofrece sus servicios a través de la red usa el agrupamiento para identificar grupos de clientes en base a sus preferencias de compras que le permita dar un servicio más personalizado. Si un cliente se interesa por un libro, el sistema identifica a que grupo pertenece y le recomienda otros libros.

Correlaciones (Tarea descriptiva)

Ayudan a determinar el grado de similitud de los valores de dos variables numéricas. Para medir la correlación se usa el coeficiente de correlación ***r***, donde $r \in [-1, 1]$, si r = 0 indica que no hay relación entre las variables, ***r*** > 0 indica que las variables están directamente relacionadas y ***r*** < 0 indica relacion inversa. ´

Ejemplo: El departamento de bomberos desea determinar las correlaciones negativas entre el empleo de distintos grosores de protección de material electrico y la frecuencia de incendios.

Reglas de asociación (Tarea descriptiva)

El objetivo es determinar relaciones no evidentes entre atributos categóricos. No necesariamente las reglas son del tipo causa-efecto.

Ejemplo: El ejemplo clásico es el de la cesta de compra para organizar ´ los productos físicamente en el supermercado.

Técnicas de la minería de datos.

Métodos estadísticos

Regresión lineal. La formula general para una regresión lineal es

$$y = c_0 + c_1 x_1 + \ldots c_n x_n$$

Donde x_i son los atributos predictores.
y es la salida o variable dependiente.

Regresión no lineal .

$$y = c_0 + f_1(x_1) + \ldots f_n(x_n)$$

Cuadrados, logaritmos, etc.
Métodos bayesianos Basados en el teorema de bayes.

Técnicas de la minería de datos.

Regla de Bayes: Si tenemos una hipótesis H sustentada para una evidencia E ->

$$p(H|E) = (p(E|H) * p(H))/p(E)$$

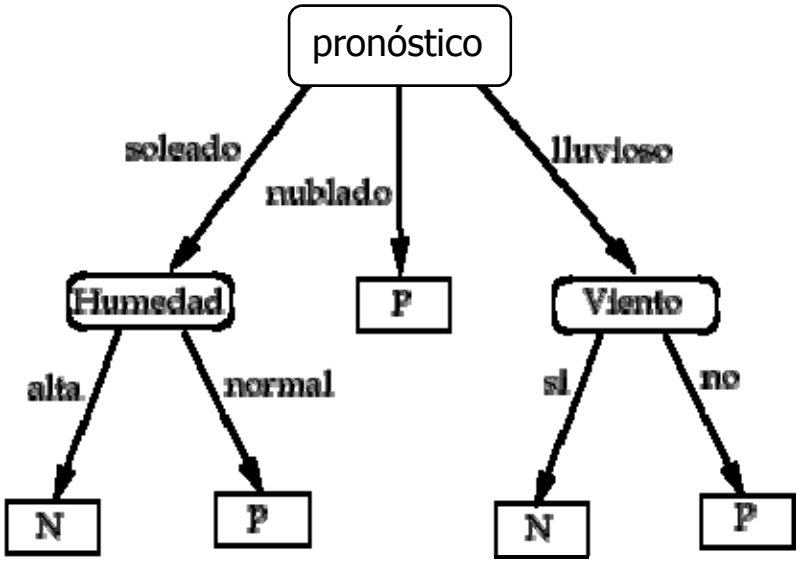
Donde p(A) representa la probabilidad del suceso y p(A|B) la probabilidad del suceso A condicionada al suceso B

Arboles de decisión . Son una serie de condiciones organizadas en forma jerárquica, a modo de árbol. Útiles para problemas que mezclen datos categóricos y numéricos.

- Útiles en Clasificación, Agrupamiento, Regresión

Instancia	pronóstico	humedad	viento	jugar
1	soleado	alta	débil	No
2	nublado	alta	débil	Si
3	lluvioso	alta	débil	Si
4	lluvioso	normal	fuerte	No
5	soleado	baja	débil	Si

- Árbol de decisión



Los arboles de decisión pueden considerarse como una forma de aprendizaje de reglas donde cada rama puede interpretarse como una regla.

La inducción de Reglas es conjunto de métodos para derivar un conjunto de reglas de la forma:

Si cond1 y cond2 y . . . Y condn entonces predicción .

Para nuestro ejemplo tenemos que:

Si pronóstico = soleado **y** humedad = normal **entonces** jugar← sí
Si pronóstico = cubierto **entonces** jugar← sí
Si pronóstico = lluvioso **y** viento = débil **entonces** jugar← sí
en otro caso ← no

Problemas con la inducción de reglas:

- Las reglas no necesariamente forman un árbol.
- Las reglas pueden no cubrir todas las posibilidades.
- Las reglas pueden entrar en conflicto.

Redes neuronales. Trabajan directamente con números y en caso de que se desee trabajar con datos nominales, estos deben numerizarse.

- Se usan en Clasificación, Agrupamiento, Regresión
- Las redes neuronales consisten generalmente de tres capas: de entrada, oculta y de salida.
- Internamente pueden verse como una grafica dirigida.

Algoritmos evolutivos. Son métodos de búsqueda colectiva dentro de un espacio de soluciones. Se siguen patrones de la evolución biológica como el cruce de los genes de los padres para la producción de hijos.

- Se usan en Clasificación, Agrupamiento, Reglas de asociación

Construcción del modelo

- En los modelos predictivos se requiere etapas de entrenamiento y validación para asegurar predicciones robustas y precisas.

- La idea es entrenar el modelo con una porción de datos (training dataset) y luego validarlo con el resto de los datos (test dataset)

Fase de evaluación e interpretación

- Verifica si los resultados son coherentes. El cliente es el que tiene la palabra final.

- Una vez obtenido el modelo, se debe proceder a su validación, comprobando que las conclusiones que arroja son válidas y suficientemente satisfactorias.
- Si se obtienen varios modelos mediante el uso de distintas técnicas, se deben comparar los modelos en busca de aquel que se ajuste mejor al problema.

Fase de evaluación e interpretación

Técnicas de evaluación . Para entrenar y probar un modelo se parten los datos en dos conjuntos: el conjunto de entrenamiento y el conjunto de prueba.

- **Validación simple**. Se utiliza cuando se posee un gran conjunto de datos. De modo que se elige entre el 5 y 50 % de los datos para la parte de las pruebas.
- **Validación cruzada** . Se parte el conjunto de datos en dos conjuntos. Primero se utiliza al primer conjunto para predecir los datos del segundo conjunto y luego se aplica el mismo proceso en modo inverso. Si la cantidad de errores no es muy grande se crea un modelo con ambos conjuntos.
- **Validación cruzada con n pliegues**. En este caso los datos se parten en n conjuntos y solo se reserva un conjunto para pruebas. El proceso se repite para los n – 1 grupos restantes.