Descubrimiento de Conocimiento a partir de Datos

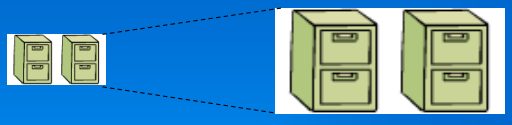
Prof. Dra. Silvia Schiaffino
ISISTAN – UNCPBA
sschia@exa.unicen.edu.ar

http://www.exa.unicen.edu.ar/catedras/dbdiscov/

Contenido del Curso

- Introducción al KDD
- Etapas
- Pre-procesamiento de datos
- Data Mining
  - Reglas de Asociación
  - Redes de Bayes
  - Clasificación
  - Modelos de Markov
  - Clustering
- Web Mining
- Social Mining

Introducción

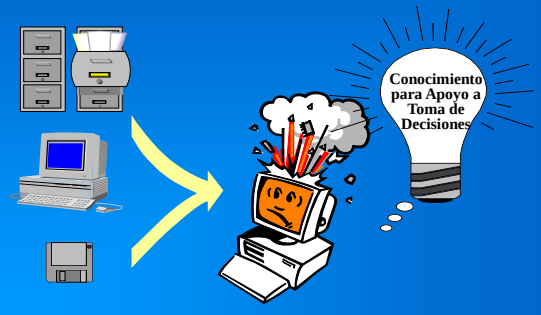


- Estamos en la era de la información
- Somos ricos en datos, pero pobres en información
- Las bases de datos son demasiado grandes
- Data Mining puede ayudar a descubrir conocimiento

Motivación

- Hay tecnología disponible para ayudarnos a juntar datos
  - Códigos de barra, lectores de tarjetas de débito y crédito, satélites, cámaras, celulares, redes sociales, etc.
- Hay tecnología disponible para ayudarnos a almacenar datos
  - Bases de datos, data warehouses, la Web, variedad de repositorios
- Necesitamos conocimiento: interpretar los datos en búsqueda de conocimiento

Conocimiento para Apoyo a Toma de Decisiones



¿Qué es un dato?

- Hechos, imágenes, sonidos...
- Los datos son la estructura fundamental sobre la cual está construida cualquier sistema de información.
- Ej.: 500

¿Qué es información?

- Datos cuya forma o formato es útil para ser usado en el proceso de toma de decisiones
- Ej: 500 mm de lluvia caída

¿Qué es conocimiento?

- Nos da la capacidad de resolver problemas, innovar y aprender basándonos en experiencias previas

- Una combinación de instintos, ideas, reglas y procedimientos que guían las acciones y decisiones

Ej: si lluvia > 200 mm → inundación

De datos a conocimiento

¿Qué es KDD?

- El Descubrimiento de Conocimiento a partir de Bases de Datos es el proceso no trivial de extraer información implícita, previamente desconocida, y potencialmente útil a partir de grandes volúmenes de datos.

¿Qué es Data Mining?

- En teoría, Data Mining (minería de datos) es un paso en el proceso de KDD.
- Es el proceso de identificación de patrones válidos, innovadores, potencialmente útiles y comprensibles de un conjunto de datos [Fayyad et al 96]
- En la práctica, data mining y KDD se han vuelto sinónimos
- Términos similares a KDD: extracción de conocimiento, descubrimiento de patrones, arqueología de datos, business intelligence, ...

¿Qué tipo de datos se almacenan?

- Transacciones de negocios
- Datos científicos
- Datos personales
- Videos e imágenes de vigilancia
- Imágenes satelitales
- Deportes
- Información digital y digitalizada
- Software
- WWW
- Informes y documentos
- Datos médicos y genéticos





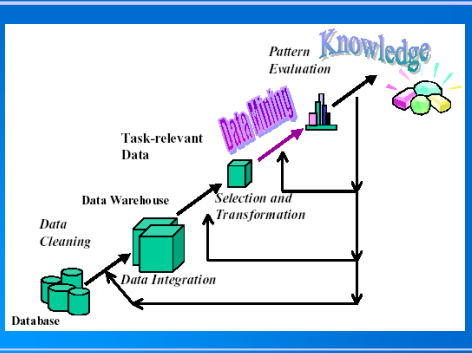
Etapas del KDD

- Recolectar los datos y agruparlos
- Limpiar los datos y juntarlos de manera que encajen
- Seleccionar los datos necesarios
- Trabajar sobre los datos para extraer la esencia de ellos
- Evaluar la salida y usarla





Etapas del KDD



Etapas del KDD

- Aprender acerca del dominio de aplicación (conocimiento previo relevante y objetivos de la aplicación)
- Recolectar e integrar los datos
- Limpiar y preprocesar los datos
- Reducir y proyectar los datos (encontrar características útiles, reducción de dimensionalidad/variable)
- Elegir las funciones de data mining (clasificación, regresión, asociación, clustering)
- Elegir los algoritmos
- Data Mining: buscar patrones de interés
- Evaluar los resultados
- Interpretar y analizar los resultados (visualización, eliminación de patrones redundantes)
- Usar el conocimiento descubierto

Identificación del Problema

- Estudio del dominio de aplicación: obtener conocimiento inicial del dominio
- Definición de los objetivos y metas a ser alcanzados
- Identificación y selección de conjuntos de datos
- Definir la relación entre simplicidad y precisión del conocimiento extraído

Identificación del Problema: Ejemplo

Se quieren analizar los datos de compras en un supermercado para identificar patrones de compras de los clientes, particularmente grupos de productos que se adquieren juntos.

Se trabajará con las compras almacenadas de 1 mes.

Se quieren encontrar patrones simples y precisos.

Data Mining: ¿sobre qué tipos de datos?

- Archivos planos (texto, binarios)
- Bases de datos heterogéneas
- Bases de datos relacionales
- Bases de datos orientadas a objetos
- Data warehouses
- Bases de datos transaccionales
- Bases de datos espaciales
- Bases de datos multimedia
- Datos temporales
- Documentos de texto
- WWW: contenido, estructura, uso

Sobre que tipo de datos? Bases de Datos relacionales

Borrow

customerID	date	itemID	#	…
C1234	99/09/06	98765	1	..
. . .				

Customer

customerID	name	address	password	birthdate	family_income	group	…
C1234	John Smith	120 main street	Marty	1965/10/10	$45000	A	…
. . .							

Items

itemID	type	title	media	category	Value	#	…
98765	Video	Titanic	DVD	Drama	$15.00	2	..
. . .							

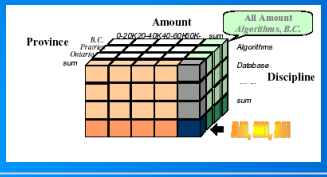
Sobre que tipo de datos? Bases de Datos transaccionales

Rentals

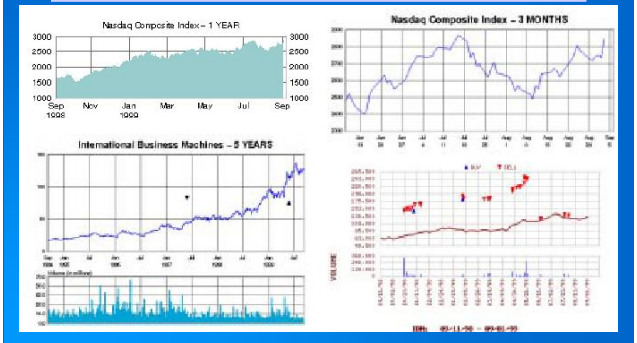
transactionID	date	time	customerID	itemList
T12345	99/09/06	19:38	C1234	I2, I6, I10, I45 …{
. . .				

Sobre qué tipo de datos? Data Warehouses

Un data warehouse es un repositorio de datos obtenido a partir de múltiples fuentes de datos (a menudo heterogéneas) y su propósito es ser utilizado como un todo bajo un mismo esquema unificado para toma de decisiones.



Sobre qué tipos de datos? Datos en series temporales



Pre-procesamiento

- Generalmente, los datos utilizados en el proceso de KDD no son adecuados para ser usados en la etapa de Extracción de Patrones
- Los datos pueden presentar diversos problemas:
  - Ruido
  - Datos incompletos
  - Formato inadecuado
  - Grandes volúmenes
- El pre-procesamiento consiste en la aplicación de técnicas con el objetivo de adecuar los datos para ser utilizados en la etapa de Extracción de Patrones

Pre-procesamiento

- Obtención y unificación de datos
- Limpieza de datos

- Reducción del volumen de datos
  - Reducción del número de ejemplos
  - Reducción del número de atributos
  - Reducción del número de valores de un atributo

Extracción de patrones

Puede ser ejecutada varias veces para ajustar los parámetros de los algoritmos y de esa forma obtener un resultado más adecuado

Sub-etapas:

- Selección de una función
  - Descriptiva o predictiva
- Selección de un algoritmo
  - Algoritmo y parámetros
- Transformación de los datos
- Obtención de patrones
  - Aplicación del algoritmo a los datos

Selección de la función

- Tareas Descriptivas
  - Describen propiedades generales de los datos existentes
    - Asociación
    - Clustering
- Tareas Predictivas
  - Predicciones basándose en inferencias a partir de los datos disponibles
    - Clasificación
    - Regresión

Funcionalidad

- Asociación:
  - Estudia la frecuencia de ocurrencia de elementos que aparecen juntos en bases de datos transaccionales. Ej. compra(x, leche)→compra(x, pan)
- Predicción:
  - Predice algún atributo desconocido o faltante basándose en otra información; o predice la clase de un objeto. Ej. predecir el valor de venta para la próxima semana de un cereal basándose en datos actuales

Funcionalidad

- Clasificación:
  - Organiza los datos en clases dadas basándose en los atributos de los objetos a clasificar. Ej: clasificar a los alumnos según su estilo de aprendizaje
- Clustering:
  - Organiza los datos en grupos basándose en sus atributos (clasificación no supervisada) Ej. agrupar lugares donde se producen crímenes para encontrar patrones de distribución.

Funcionalidad

- Análisis de Excepciones (Outliers):
  - Identifica y explica excepciones, elementos que no cumplen con el modelo de datos (detección de fraude)
- Análisis de series de tiempo:
  - Analiza tendencias y desviaciones, secuencias similares, patrones secuenciales

Selección del Algoritmo

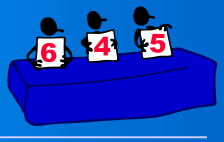
- Hay varios algoritmos disponibles para cada actividad
- En esta etapa debe escogerse el algoritmo a ser utilizado así como también los parámetros del mismo.
- Resultados experimentales muestran que no existe un único algoritmo bueno para todas las tareas. Pueden elegirse varios algoritmos para la extracción de patrones

Transformación de los Datos

- Una vez elegida la función y el/los algoritmos, los datos deben ser adecuados al formato de entrada de los algoritmos de extracción de patrones para que éstos puedan ser utilizados

Pos-procesamiento

- Se necesitan métodos para investigar la precisión de los algoritmos, la representación del modelo, la complejidad y dificultad del conocimiento extraído.
- El conocimiento extraído puede ser simplificado, evaluado, visualizado o simplemente documentado para el usuario final.



Motivación del pos-procesamiento

- Una tarea de data mining puede generar miles de patrones, los cuales no son todos interesantes o relevantes para el usuario.
- Darle al usuario una gran cantidad de patrones no es productivo
- El usuario quiere una cantidad pequeña de patrones interesantes
- Si el conocimiento descubierto es útil o interesante depende de la aplicación y del usuario (es subjetivo)

Tareas de pos-procesamiento

- Validación
  - Precisión
  - Comprensibilidad
  - Interés
- Interpretación y explicación
  - Documentación
  - Visualización
  - Modificación
  - Comparación
- Filtrado de Conocimiento
  - Restricción de atributos
  - Ordenamiento por métricas

Filtrado de Conocimiento

- Mecanismos de poda (árboles de decisión) y truncado para reglas de decisión
- Aplicado en casos en que los algoritmos generan árboles de decisión con muchas hojas o reglas de decisión muy específicas cubriendo pocos ejemplos (overfitting)

Interpretación y Explicación

- El conocimiento puede ser resumido, documentado, visualizado o modificado de manera de tornarlo comprensible para el usuario
- El conocimiento se puede documentar o modificar, posibilitando su uso en sistemas de soporte a toma de decisiones
- El conocimiento puede ser comparado con conocimiento pre-existente para verificar conflictos y conformidades.

Evaluación del Conocimiento

- Métricas objetivas
  - Basadas en estadísticas o en la estructura de los patrones. Ej: validez (soporte), certeza (confianza)

- Métricas subjetivas
  - Basadas en las creencias del usuario acerca de los datos. Ej: comprensibilidad, novedad, utilidad, *unexpectedness*

Medidas de interés

- Un patrón es interesante si es:
  - Fácil de entender para los usuarios
  - Válido sobre nuevos datos con algún grado de certeza
  - Potencialmente útil
  - Novedoso, o valida alguna hipótesis que el usuario busca confirmar

Nombres para tipos de datos complejos

- Text Mining
  - Correo electrónico, librerías, bibliotecas, páginas Web
- Spatial Mining
  - GIS, imágenes médicas
- Multimedia Mining
  - Bases de datos que contienen audio y video
- Web Mining
  - Análisis de patrones de acceso
  - Datos estructurados y semi-estructurados
- OLAP Mining
  - Data mining y data warehousing
- Opinion Mining

Aplicaciones de KDD

- Toma de decisiones y Análisis de Datos de Negocios
  - Marketing
    - Reconocer segmentos de mercado específicos que respondan a características particulares
    - Campañas de publicidad por mail

  - Perfiles de clientes

    - Segmentación de los clientes para estrategias de marketing y/u ofertas de productos
    - Comprensión del comportamiento de los clientes
    - Retención y lealtad de los clientes

Aplicaciones de KDD

- Detección de fraude

  - Detección de fraude telefónico
    - Modelo de llamada telefónica: destino de la llamada, duración, día de la semana. Analizar patrones que se desvían de la norma esperada.
  - Detección de fraude en obras sociales
  - Detección de fraude en tarjetas de crédito
  - Detección de lavado de dinero

Aplicaciones de KDD

- Text Mining
  - Filtrado de mensajes (correo electrónico, newsgroups, etc.)
  - Análisis de artículos de diarios

- Medicina
  - Asociación entre patologías y síntomas
  - Datos genéticos
  - Imágenes médicas
  - Bioinformática

Aplicaciones de KDD

- Deportes

  - El IBM Advanced Scout se usa para analizar las estadísticas de juego de la NBA para obtener ventaja competitiva (tiros, bloqueos, asistencias, faltas)

http://www.springerlink.com/content/r32q737858160r97/

- Astronomía

  - Identificación de volcanes en Júpiter
  - Identificación de quásares

Aplicaciones de KDD

- Cámaras de vigilancia
  - Uso de cámaras y análisis de excepciones para detectar actividades o individuos sospechosos

- Web surfing y Web Mining
  - IBM Surf-Aid: usa algoritmos de data mining para descubrir preferencias de los clientes en e-commerce (Web access logs)
  - Sitios Web adaptativos
  - Pre-fetching y caching de páginas Web

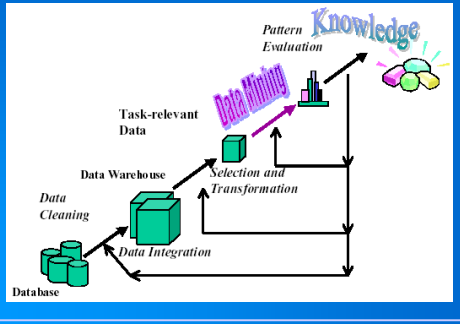
Ejercicio

- Para el viernes traer para comentar ejemplos de empresas en Tandil que utilicen KDD.
- Indicar quién, para qué y qué técnicas usan.

Contenido del Curso

- Introducción al KDD
- Etapas
- Pre-procesamiento de datos
- Data Mining
  - Reglas de Asociación
  - Clasificación y Predicción
  - Clustering
  - Modelos de Markov
- Web Mining
- Social Web Mining

Etapas del KDD



Pre-procesamiento: Motivación

En el mundo real los datos pueden ser inconsistentes, incompletos y/o tener ruido.

Pueden producirse errores por:

- Fallas en los instrumentos de recolección de datos
- Problemas en el ingreso de datos
- Error o mala interpretación humana al cargar los datos
- Problemas de transmisión de datos
- Limitaciones tecnológicas
- Discrepancia en convenciones de nombres

Resultados:

- Registros duplicados
- Datos incompletos
- Inconsistencias en los datos
- Ruido

Pre-procesamiento de Datos

- Limpieza de Datos

- Integración de Datos
- Transformación de Datos

- Reducción de Datos

Limpieza de Datos

En el mundo real los datos pueden ser inconsistentes, incompletos y/o tener ruido.

Valores faltantes en algunos atributos
Valores no considerados durante la carga de datos
Errores aleatorios
...

La limpieza de datos intenta:

- Completar los valores faltantes
- Suavizar los datos con ruido
- Corregir las inconsistencias
- Identificar y eliminar excepciones (outliers)

Datos faltantes

- Ignorar las tuplas que tienen elementos faltantes (ej. si lo que falta es la clave)
- Completar los valores faltantes manualmente
- Usar una constante para completar los valores faltantes (null, unknown, ?). No recomendado
- Usar el valor medio del atributo para completar los faltantes
- Usar el valor medio para todas las muestras que pertenecen a una misma clase
- Inferir el valor más probable para completar el valor faltante (regresión, árbol de decisión)

Suavizado de datos con ruido

- El propósito del suavizado es eliminar el ruido. Puede hacerse mediante:
  - Binning (colocar en compartimientos)
  - Clustering
  - Regresión

Binning

Este método suaviza los datos consultando el valor de su vecino.

1. Los datos son ordenados para obtener los valores “en sus vecindades”.
2. Los datos son distribuidos en compartimientos de iguales tamaños.
3. Se aplica suavizado local

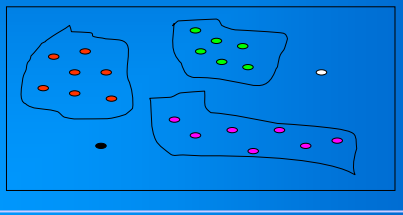
Ejemplo (precios): 4, 8, 15, 21, 21, 24, 25, 28, 34 Profundidad: 3

Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Suavizado por media:
Bin 1: 9, 9, 9 Bin 2: 22, 22, 22 Bin 3: 29, 29, 29
Suavizado por límites:
Bin 1: 4, 4, 15 Bin 2: 21, 21, 24 Bin 3: 25, 25, 34

Clustering

- Los datos son organizados en grupos de valores similares. Los valores raros que caen afuera de estos grupos son considerados excepciones y son descartados

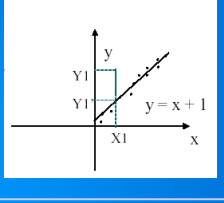


Inspección combinada humano-computadora

- Ejemplo: detección de caracteres
  - usar métricas de teoría de la información para identificar automáticamente *outliers* en escritura manuscrita.
  - Cada carácter predicho tiene un nivel de sorpresa
  - Si este nivel supera un umbral dado, es un *outlier*
  - Un usuario analiza los *outliers*

Regresión

- La regresión consiste en hacer que los datos encajen en una función. Una regresión lineal, por ejemplo, encuentra la recta que encaje con 2 variables de manera que una variable pueda predecir la otra.



Datos inconsistentes

- Corregir las inconsistencias manualmente usando referencias externas; o utilizando algún programita o rutina.
- Usar las restricciones funcionales entre atributos
- Inconsistencias debido a la integración de datos

Integración de Datos

El análisis de datos puede requerir una combinación de datos de múltiples fuentes (MS Excel, BD relacionales, data warehouse) en un almacenamiento de datos coherente

Problemas:

- Esquema de integración: CID = número de cliente = Cust-id
- Heterogeneidad semántica
- Conflictos de valores de datos (diferentes escalas, diferentes representaciones)
- Registros redundantes
- Atributos redundantes (análisis de correlación)

METADATOS

Transformación de datos

Los datos a veces se encuentran en una forma no apropiada. Puede ser que el algoritmo a usar no pueda manejarlos, que la forma de los datos no sea regular o que los datos no sean lo suficientemente específicos.

- Normalización (comparar peras con peras)
- Suavizado (elimina el ruido)
- Agregación (operación de resumen aplicada a datos)
- Generalización (jerarquías de conceptos)
- Construcción de atributos (área con altura y ancho)

Normalización

- Normalización min-max: transformación lineal de v a v', v'=v-min/(max-min)(newmax-newmin)+newmin

Ej: transformar 30000 entre [10000..45000] en [0..1] → 30000 - 10000/(45000 – 10000) (1-0) + 0=0.514

- Normalización zscore: normalizar v en v' basándose en valor medio del atributo y desvío estándar

v'=v-Media/DesvíoEstándar

- Normalización por escala decimal: mueve el punto decimal de v en j posiciones tal que j es el mínimo número de posiciones movidas para hacer que el valor caiga en [0..1] - v'=v/10**j, donde j max(|v'|)<1

Ej: si v está entre –56 y 9976, j=4 → v' está entre –0.0056 y 0.9976

Construcción de atributo

Se construyen nuevos atributos a partir de otros para mejorar la exactitud y entendimiento de los datos

Ej. área a partir de la altura y el ancho

Reducción de datos

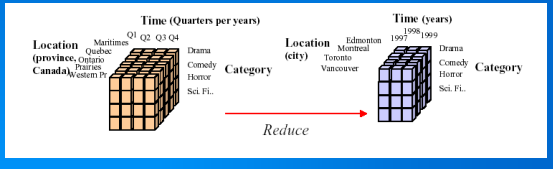
Los datos a menudo son demasiado grandes. Reduciendo la cantidad de datos se puede mejorar la *performance*. La reducción de datos consiste en reducir la representación del conjunto de datos siempre y cuando se produzcan los mismos (o casi los mismos) resultados.

La reducción de datos incluye:

- Agregación del Cubo de Datos
- Reducción de dimensionalidad
- Compresión de datos (codificación)
- Discretización y generación de jerarquías
- Reducción de numerosidad
  - Regresión
  - Histogramas
  - Clustering
  - Muestreo

Agregación del Cubo de Datos

- Reduce los datos al nivel de concepto que se necesita en el análisis.



- Las consultas acerca de información agregada deben ser respondidas usando el cubo de datos cuando sea posible.

Reducción de dimensionalidad

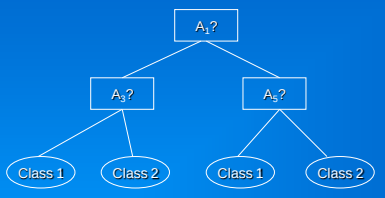
- Selección de atributos
  - Seleccionar solamente los atributos necesarios (eliminar teléfono, mail, etc.)
  - El objetivo es encontrar el mínimo número de atributos tal que la distribución de probabilidad resultante de las clases sea lo más cercana posible a la distribución de probabilidad usando todos los atributos
  - Número exponencial de posibilidades

Reducción de dimensionalidad

- Uso de heurísticas para hallar el atributo “local best” (el más pertinente)

  Conjunto Inicial $\{A_1\ A_2\ A_3\ A_4\ A_5\}$
  - Step-wise forward selection $\{\}$ $\{A_1\}$ $\{A_1, A_3\}$ $\{A_1\ A3_3\ A_5\}$
  - Step-wise backward elimination $\{A_1\ A_2\ A_3\ A_4\ A_5\}$ $\{A_1\ A_3\ A_4\ A_5\}$ $\{A_1\ A_3\ A_5\}$
  - Combinación de las 2 anteriores
  - Inducción de árbol de decisión

Inducción de árbol de decisión



Conjunto inicial de atributos: $\{A_1\ A_2\ A_3\ A_4\ A_5\}$

Conjunto reducido de atributos: $\{A_1\ A3_3\ A_5\}$

Compresión de datos

- La compresión de datos reduce el tamaño de los datos
  - Reduce espacio de almacenamiento
  - Ahorra tiempo de comunicación
- Hay compresión con pérdida y sin pérdida. Se usa para todo tipo de datos. Algunos métodos son específicos de los datos, otros son versátiles.
- Para data mining, la compresión de datos es beneficiosa si los algoritmos de data mining pueden manipular los datos comprimidos directamente sin descomprimirlos.

Reducción de numerosidad

- El volumen de datos puede ser reducido eligiendo formas alternativas de representaciones de datos

Paramétrico
- Regresión: un modelo o función estima la distribución de los datos. En lugar de los datos guardo los parámetros (ej. modelo lineal).

No paramétrico
- Histogramas
- Clustering
- Muestreo

Regresión

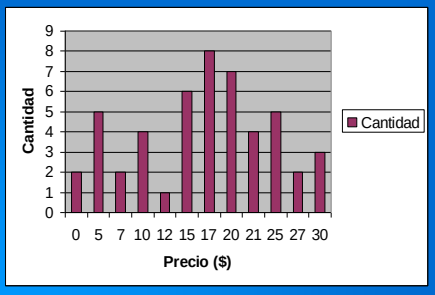
Regresión lineal

Los datos deben ajustarse a una línea recta.

Una variable aleatoria Y puede modelarse como una función dependiente de otra variable aleatoria X con: $Y = \alpha + \beta X$

donde α y β son coeficientes de regresión (cuadrados mínimos)

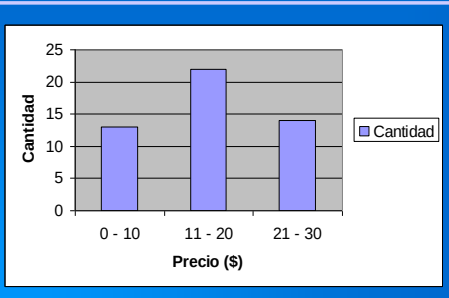
Reducción con histogramas



Reducción con histogramas

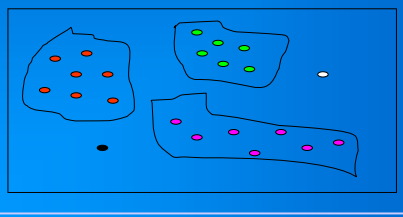
- Dividir los datos en grupos y guardar una representación de cada grupo (suma, cantidad, etc)
- Equi-width (histograma con barras que tienen el mismo ancho)
- Equi-depth (histograma con barras que tienen la misma altura)
- V-optimal (histograma con la menor varianza Σ $count_b$ * $value_b$)
- MaxDiff (balde con límites definidos por umbral definido por el usuario)

Reducción por histogramas



Reducción con clustering

- Particionar los datos en clusters basándose en la cercanía en el espacio. Retener representantes de los clusters (centroides) y *outliers*. La efectividad depende de la distribución de los datos. El clustering jerárquico es posible.



Reducción con muestreo

- Permite que un gran conjunto de datos sea representado por muestras aleatorias de los datos mucho más pequeñas
- Cómo se selecciona una muestra aleatoria?
- Representan los patrones en la muestra los patrones en los datos?

- Muestra aleatoria simple sin reposición (SRSWOR)
- Muestra aleatoria simple con reposición (SRSWR)
- Muestra con clusters (SRSWOR o SRSWR sobre clusters)
- Muestra estratificada (estrato: grupo basado en valores de atributos)

Discretización

- La discretización se usa para reducir el número de valores para un atributo continuo, dividiendo el rango del atributo en intervalos. Las etiquetas de los intervalos se usan para reemplazar los valores reales de los datos.
- Algunos algoritmos de data mining solamente aceptan atributos categóricos y no pueden manejar un rango de valores continuos.
- La discretización puede reducir el conjunto de datos, y puede usarse para generar jerarquías de conceptos automáticamente.

Jerarquías de conceptos: Datos numéricos

Para datos numéricos hay gran diversidad de rangos de valores posibles y frecuentes actualizaciones.
Es difícil construir jerarquías de conceptos para atributos numéricos
La generación automática de jerarquía de conceptos se basa en el análisis de la distribución de los datos

Binning: representante de bin (media, mediana) → binning recursivo
Análisis de Histogramas: recursivo con mínimo tamaño de intervalo
Clustering: clustering recursivo
Basados en Entropía: particionamiento binario con evaluación de ganancia de información
Segmentación 3-4-5: intervalos uniformes con límites redondeados

Jerarquías de Conceptos: Datos categóricos

Atributos categóricos: número finito (pero posiblemente grande) de valores diferentes, sin ordenamiento entre ellos.

- Especificación de ordenamiento parcial de los atributos por usuarios o expertos
- Especificación de una porción de la jerarquía por agrupamiento de datos explícito
- Especificación de un conjunto de atributos, pero no del ordenamiento parcial
- Especificación de conjunto parcial de atributos

Jerarquías de Conceptos: Datos categóricos

- Especificación de ordenamiento parcial de los atributos por usuarios o expertos

  calle < ciudad < provincia < país

- Especificación manual

{san juan, mendoza, san luis} < región de cuyo; {entre ríos, corrientes, misiones} < región del litoral

Jerarquías de Conceptos: Datos categóricos

- Conjunto de atributos que forman la jerarquía, sin orden, que puede intentar generarse automáticamente.

  conjunto calle, país, provincia, ciudad. Se ordenan: país (15), provincia (365), ciudad (3567), calle (674.339).

- Especificación de conjunto parcial de atributos
  Especificación manual

  calle y ciudad. Se usa información semántica para completar los restantes

¿Preguntas?

Forma de Evaluación del curso

GRADO

- Trabajos en grupos de 2 personas (cursada y final)
- Trabajos que muestren el proceso entero de KDD en un dominio particular
- Presentación de 15 min.
- Fecha de presentaciones: Lunes 21 de Noviembre
- Consultas: Viernes de 14 a 16 hs. en ISISTAN

POSGRADO

- A definir con la cátedra

Bibliografía básica

- Data Mining: Concept and Techniques – Jiawei Han and Micheline Kamber – Morgan Kaufmann Publishers – Second Edition – 2006
- Data Mining: Practical Machine Learnig Tools with Java Implementations – Ian Witten and Eibe Frank – Morgan Kaufmann Publishers – Second Edition - 2005