OLAP y Minería de Datos: Introducción

Carlos Hurtado L.

churtado@dcc.uchile.cl

Departamento de Ciencias de la Computación

Universidad de Chile

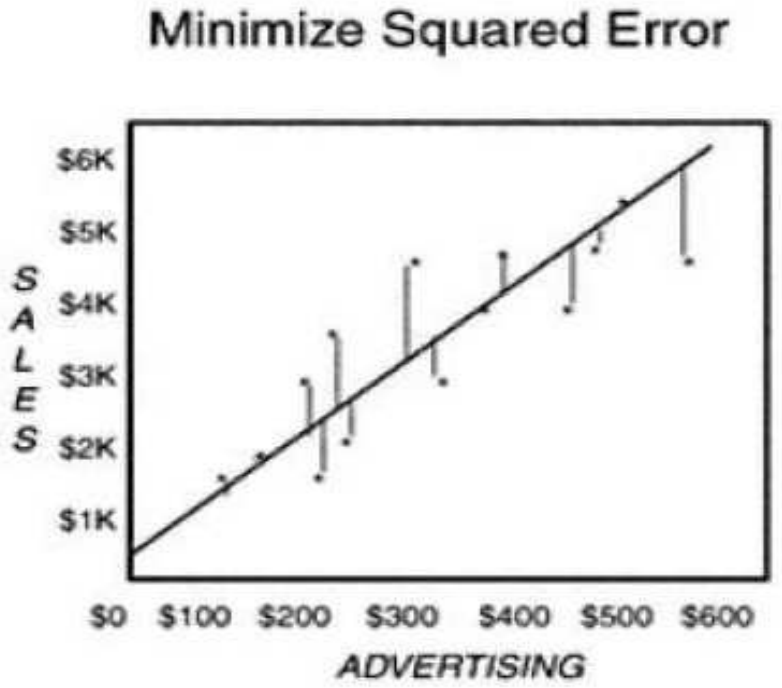
¿Qué es la Minería de Datos?

- “Extracción de modelos y patrones interesantes, potencialmente útiles y no triviales desde bases de datos de gran tamaño”
  Jiawei Han, Micheline Kamber, “Data Mining: Concepts and Techniques”, Morgan Kaufman, 2001.
- **Modelo:** representación abstracta de un conjunto de datos.
- **Patrón:** similar a un modelo pero se enfoca en un subconjunto de los datos.

Noción de Modelo: Regresión

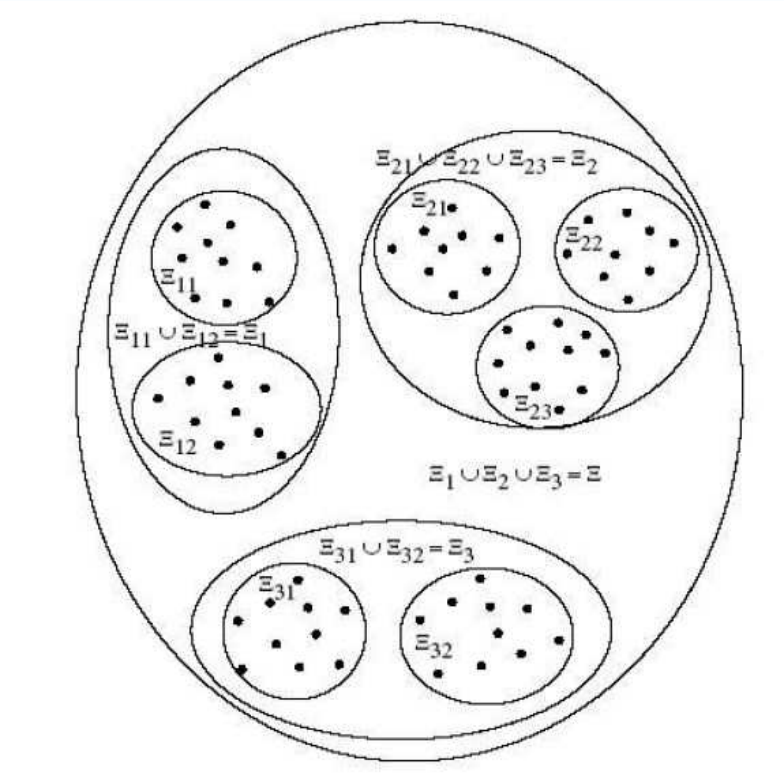
Sales = 17.813 + .0897*Advertising

Advertising	Sales
$120	$1,503
$160	$1,755
$205	$2,971
$210	$1,682
$225	$3,497
$230	$1,998
$290	$4,528
$315	$2,937
$375	$3,622
$390	$4,402
$440	$3,844
$475	$4,470
$490	$5,492
$550	$4,398



En este caso el modelo es una función lineal.

Noción de Modelo: Segmentación

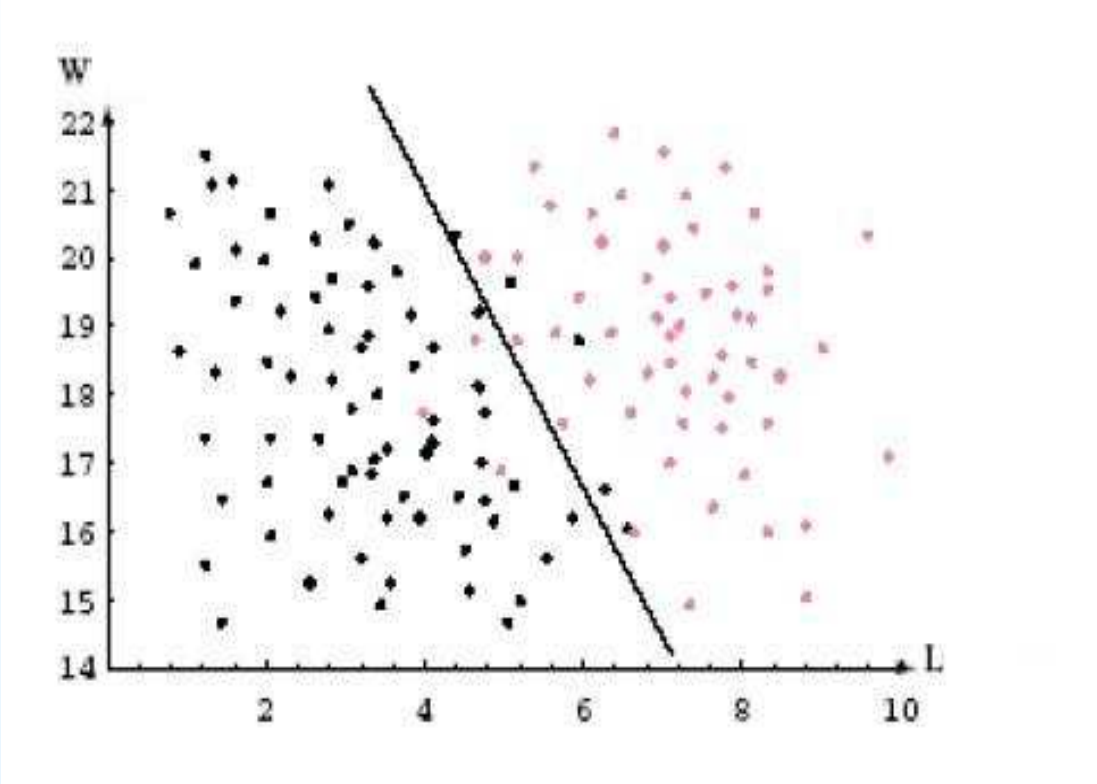


En este caso el modelo es una jerarquía de segmentos (dendograma)
Cada segmento es un par centroide-radio.

Aplicaciones Clustering: Segmentación de Contenido

http://www.marumushi.com/apps/newsmap/

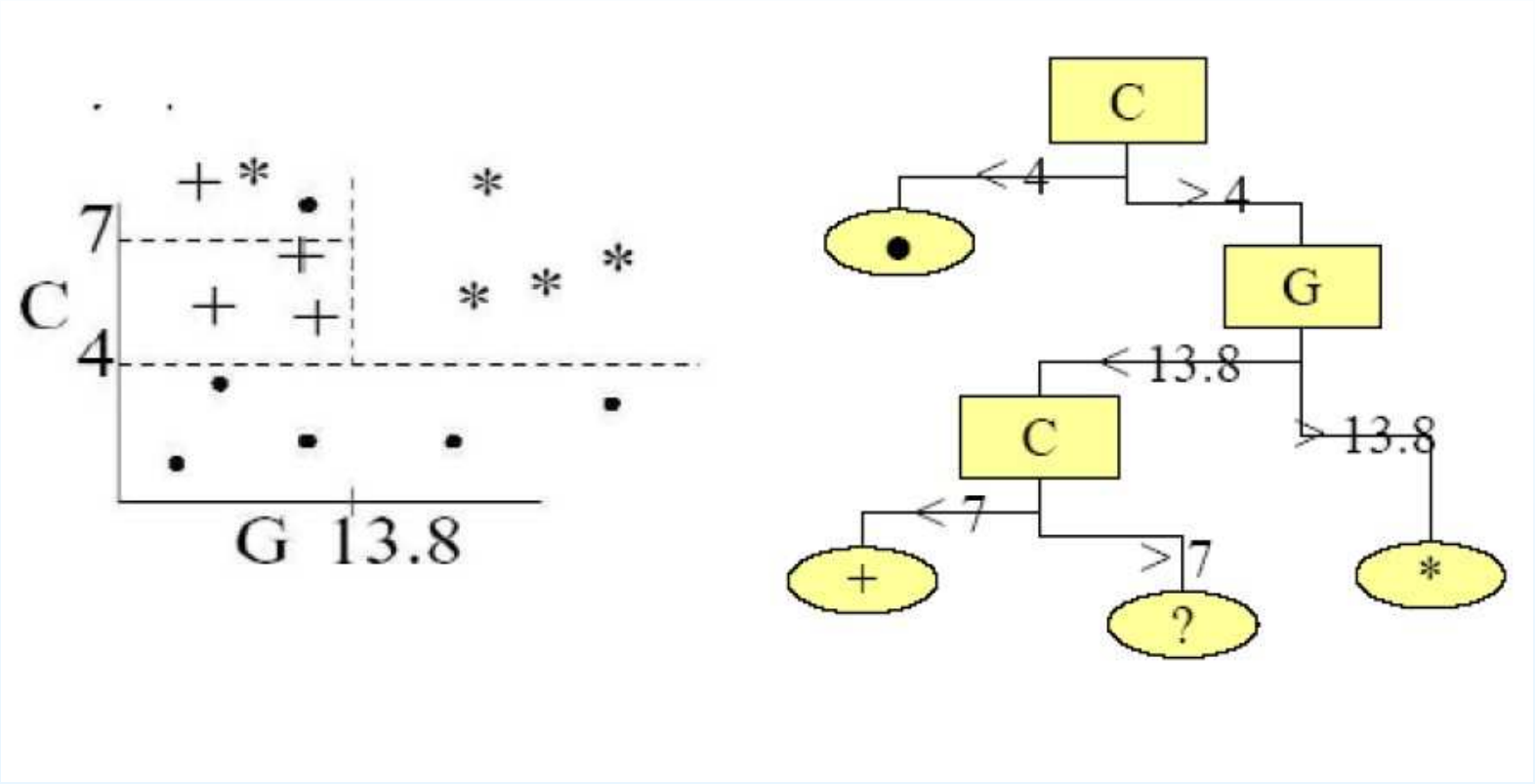
Noción de Modelo: Clasificación



En este caso el modelo es la función binaria:

$$f(X) = \begin{cases} 1 & \text{si } W + 4{,}5\ L - 39 < 0 \\ 0 & \text{si no.} \end{cases}$$

Noción de Modelo: Clasificación



En este caso el modelo es el árbol de decisión que se muestra a la derecha.

Tipos de Modelos

- **Regresión** Predice el valor de una variable numérica.
- **Clasificación** Predice el valor de una variable categórica.
- **Segmentación** Describe los datos como un conjunto de segmentos o grupos. Descube los valores de una variable oculta que no está registrada en los datos.
- **Asociaciones** Describe correlaciones o dependencias entre variables.
- **Otros:** distribuciones de probabilidad, modelos probabilísticos (cadenas de markov), redes semánticas, modelos para análisis de lenguaje natural, cubos de datos OLAP, etc.

¿Qué tan interesante es un modelo?

La minería de datos es un proceso de inducción de modelos. Cualquier modelo puede resultar de este proceso, sin embargo la idea es producir modelos “interesantes”:

- **Validez**: Certeza de que el modelo será verdadero ante nuevos datos. El modelo es generalizable en el futuro.

¿Qué tan interesante es un modelo?

La minería de datos es un proceso de inducción de modelos. Cualquier modelo puede resultar de este proceso, sin embargo la idea es producir modelos “interesantes”:

- **Validez**: Certeza de que el modelo será verdadero ante nuevos datos. El modelo es generalizable en el futuro.
- **Novedad**: El modelo contiene conocimiento novedoso y no trivial.

¿Qué tan interesante es un modelo?

La minería de datos es un proceso de inducción de modelos. Cualquier modelo puede resultar de este proceso, sin embargo la idea es producir modelos “interesantes”:

- **Validez**: Certeza de que el modelo será verdadero ante nuevos datos. El modelo es generalizable en el futuro.
- **Novedad**: El modelo contiene conocimiento novedoso y no trivial.
- **Utilidad**: El modelo nos sirve para algo: toma de decisión, aplicación.

¿Qué tan interesante es un modelo?

La minería de datos es un proceso de inducción de modelos. Cualquier modelo puede resultar de este proceso, sin embargo la idea es producir modelos “interesantes”:

- **Validez**: Certeza de que el modelo será verdadero ante nuevos datos. El modelo es generalizable en el futuro.
- **Novedad**: El modelo contiene conocimiento novedoso y no trivial.
- **Utilidad**: El modelo nos sirve para algo: toma de decisión, aplicación.
- **Simplicidad**: el modelo es fácil de entender y facilita la comprensión de los datos.

Usos de un modelo

- **Análisis exploratorio:** el objetivo es explorar los datos sin tener necesariamente una idea clara de lo que se busca. Ejemplos:
  - Modelos para Visualización.
  - Histogramas.
  - Cubo de datos OLAP.

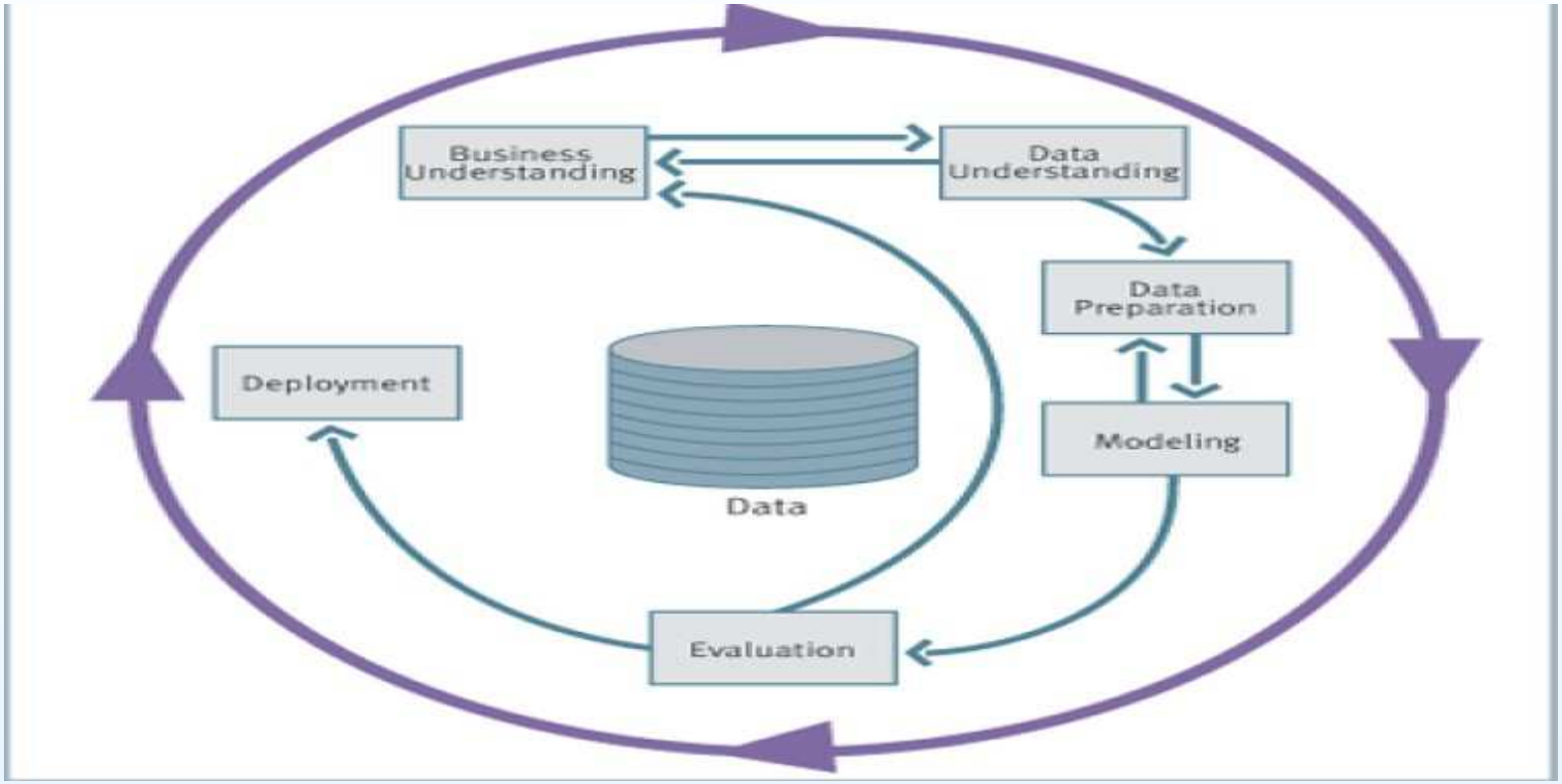
Usos de un modelo

- **Análisis exploratorio:** el objetivo es explorar los datos sin tener necesariamente una idea clara de lo que se busca. Ejemplos:
  - Modelos para Visualización.
  - Histogramas.
  - Cubo de datos OLAP.
- **Descripción de los datos:** el objetivo del modelo es ayudarnos a entender los datos. Ejemplos:
  - Segmentación.
  - Distribuciones de probabilidad.
  - Asociaciones.

Usos de un modelo (cont.)

- **Predicción:** Objetivo es predecir el valor de una variable en el futuro. Ejemplos:
  - Regresión.
  - Clasificación.

Proceso de Minería de Datos



CRISPDM: modelo del proceso de minería de datos neutral a herramientas. Desarrollado a partir de 1996 por un consorcio de más aprox. 300 organizaciones.

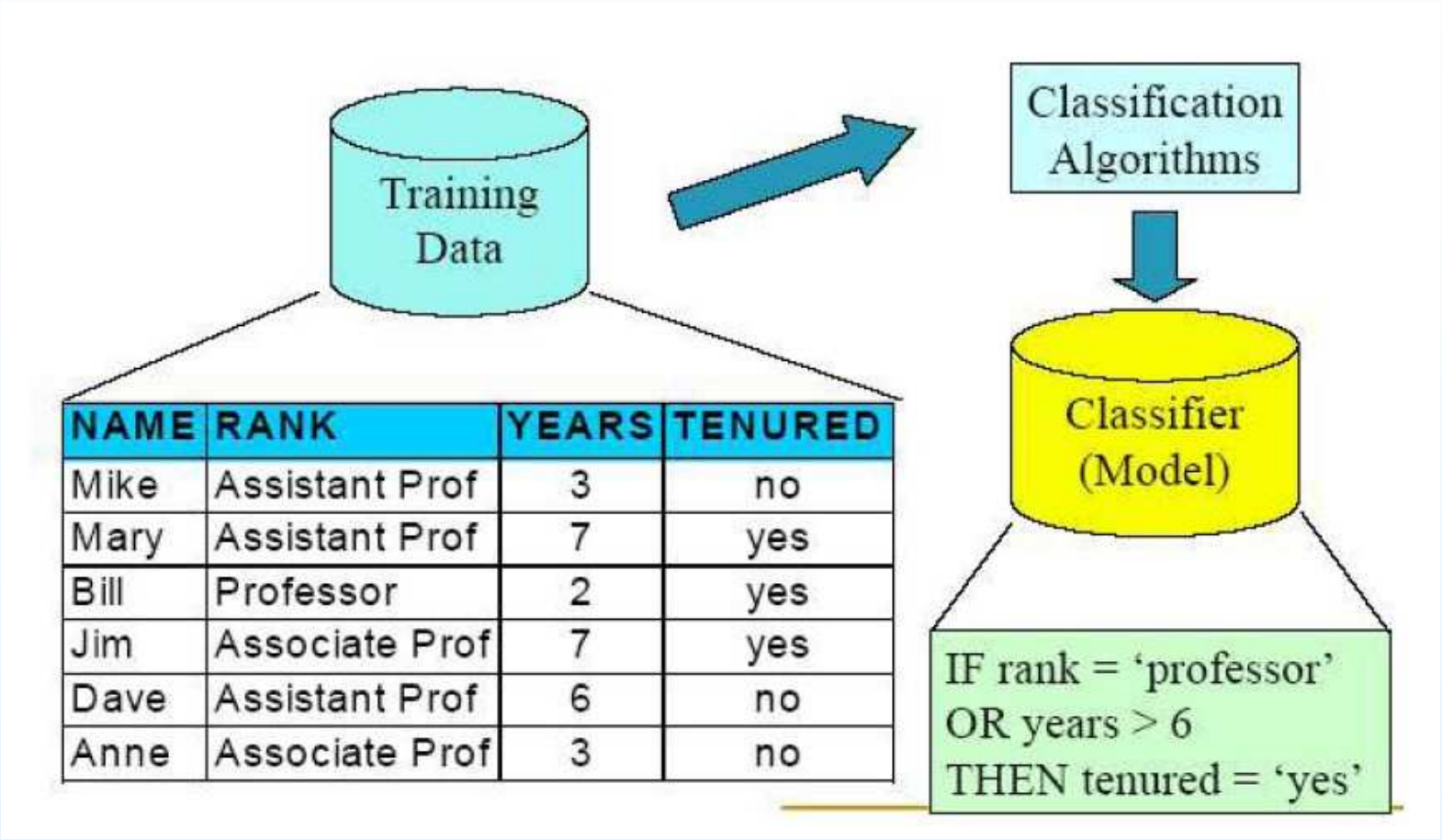
CRISP-DM

- **Entendimiento del Problema**
- **Entendimiento de los Datos**
  - Datos y su estructura
  - Variables relevantes
  - Calidad de los Datos
- **Preparación de los Datos**
  - Extracción de datos
  - Transformación y limpieza

CRISP-DM

- **Modelación**
  - Seleccionar algoritmo
  - Diseño de experimento de evaluación
  - Inducción del modelo
  - Evaluar el modelo
- **Evaluación de resultados**
- **Implantación del modelo**

Algoritmo de Minería de Datos



NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

Algoritmo de Minería de Datos (Etapa de Modelación)

Toma un conjunto de datos e induce un modelo o patrones. Es necesario tomar las siguientes decisiones:

- **Tipo de modelo o patrón:** existen cientos de tipos de modelos.
  - Porejemplo, redes Neuronales, árboles de decisión, regresión logística, segmentación jerárquica, discriminantes lineales, etc. etc.

Algoritmo de Minería de Datos (Etapa de Modelación)

Toma un conjunto de datos e induce un modelo o patrones. Es necesario tomar las siguientes decisiones:

- **Tipo de modelo o patrón:** existen cientos de tipos de modelos.
  - Porejemplo, redes Neuronales, árboles de decisión, regresión logística, segmentación jerárquica, discriminantes lineales, etc. etc.
- **Función de Evaluación:** nos entrega un puntaje que mide la calidad del modelo.
  - Por ejemplo, en regresión la función de evaluación más común es el error mínimo cuadrado.

Algoritmo de Minería de Datos (cont.)

- **Estrategia de Optimización y búsqueda del modelo:** define el proceso de búsqueda de un modelo particular en el espacio de posibles modelos.
  - Por ejemplo, en redes Neuronales se usa descenso por el gradiente para encontrar un modelo que minimice el error estimado.

Algoritmo de Minería de Datos (cont.)

- **Estrategia de Optimización y búsqueda del modelo:** define el proceso de búsqueda de un modelo particular en el espacio de posibles modelos.
  - Por ejemplo, en redes Neuronales se usa descenso por el gradiente para encontrar un modelo que minimice el error estimado.
- **Estrategia de manejo de datos:** define el manejo de datos en el proceso de inducción del modelo.
  - Por ejemplo, estructuras de datos, manejo de memoria secundaria, etc.

Aplicaciones de Clustering

- Sistemas de Recomedación (filtrado colaborativo): Amazon, MovieLens.
- “Database Marketing”. Weden and Kamakura. Market Segmentation: Conceptual and Methodological Foundations.
- Segmentación de Contenido.
- Generación automática de directorios Web.

Aplicaciones de Clustering (cont.)

- Minería de uso en la Web. Ejemplo: creación de la sección “living” en MSNBC.
- Ranking y recomendación de consulta en motores de búsqueda: R. Baeza-Yates, C. Hurtado, M. Mendoza. Improving Search Engines by Query Clustering. JASIST, 2007.
- Detección de tópicos emergentes en la “Web viva”.
- etc. etc. etc.

Aplicaciones de Clasificación

- Detección de fraude.
- Predicción de riesgo.
- Clasificación de objetos celestes.
- Clasificación de texto.
- Diagnóstico de enfermedades.
- etc. etc. etc.

Aplicaciones Clasificación: Clasificación de Texto

- **Tópicos:** deporte, política, tecnología, etc.

Aplicaciones Clasificación: Clasificación de Texto

- **Tópicos:** deporte, política, tecnología, etc.
- **Sentimientos:** connotacion positiva, negativa, neutral, emotiva, ofensiva, etc.

Aplicaciones Clasificación: Clasificación de Texto

- **Tópicos:** deporte, política, tecnología, etc.
- **Sentimientos:** connotacion positiva, negativa, neutral, emotiva, ofensiva, etc.
- **Spam:** entre 3000 - 7000 splogs (blogs falsos o spam) se crean diariamente (fuente: Technorati).

Aplicaciones Clasificación: Clasificación de Texto

- **Tópicos:** deporte, política, tecnología, etc.
- **Sentimientos:** connotacion positiva, negativa, neutral, emotiva, ofensiva, etc.
- **Spam:** entre 3000 - 7000 splogs (blogs falsos o spam) se crean diariamente (fuente: Technorati).
- **Contenido no apto:** pornografía, violencia, etc.

Aplicaciones Clasificación: Clasificación de Texto

- **Tópicos:** deporte, política, tecnología, etc.
- **Sentimientos:** connotacion positiva, negativa, neutral, emotiva, ofensiva, etc.
- **Spam:** entre 3000 - 7000 splogs (blogs falsos o spam) se crean diariamente (fuente: Technorati).
- **Contenido no apto:** pornografía, violencia, etc.
- **Comunitario:** contenido de interés para una comunidad de usuarios.

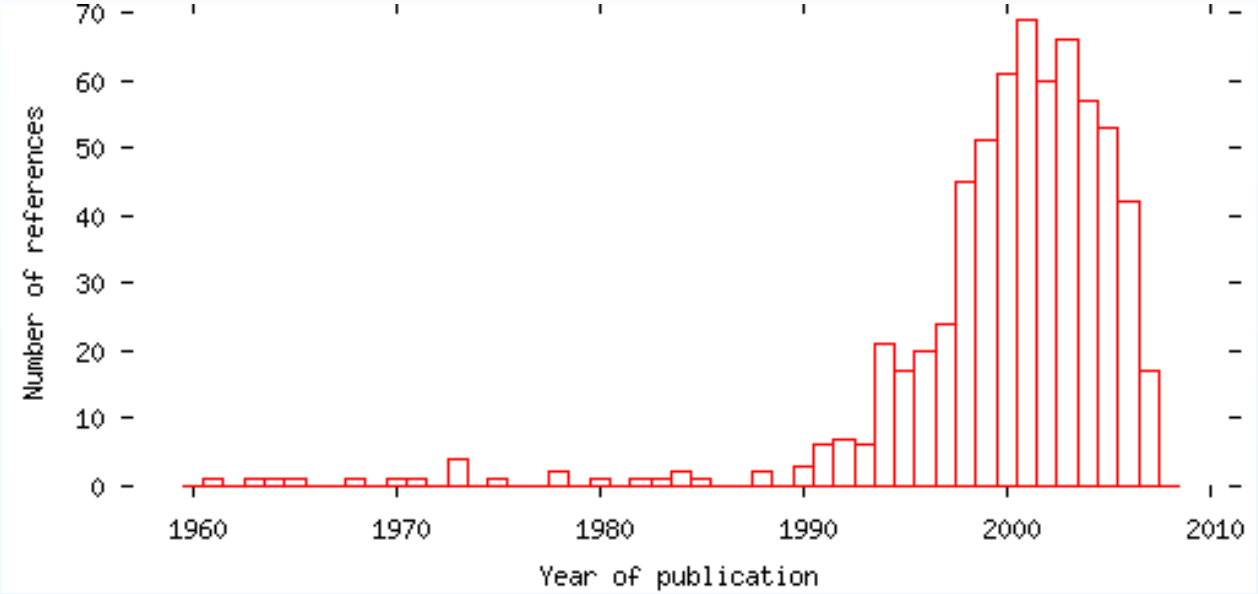
Aplicaciones Clasificación: Clasificación de Texto

- **Tópicos:** deporte, política, tecnología, etc.
- **Sentimientos:** connotacion positiva, negativa, neutral, emotiva, ofensiva, etc.
- **Spam:** entre 3000 - 7000 splogs (blogs falsos o spam) se crean diariamente (fuente: Technorati).
- **Contenido no apto:** pornografía, violencia, etc.
- **Comunitario:** contenido de interés para una comunidad de usuarios.
- **Personalizado:** contenido de interés para un usuario único.

Aplicaciones Clasificación: Clasificación de Texto

- **Tópicos:** deporte, política, tecnología, etc.
- **Sentimientos:** connotacion positiva, negativa, neutral, emotiva, ofensiva, etc.
- **Spam:** entre 3000 - 7000 splogs (blogs falsos o spam) se crean diariamente (fuente: Technorati).
- **Contenido no apto:** pornografía, violencia, etc.
- **Comunitario:** contenido de interés para una comunidad de usuarios.
- **Personalizado:** contenido de interés para un usuario único.
- **Categorías del lenguaje** opiniones vs. hechos.

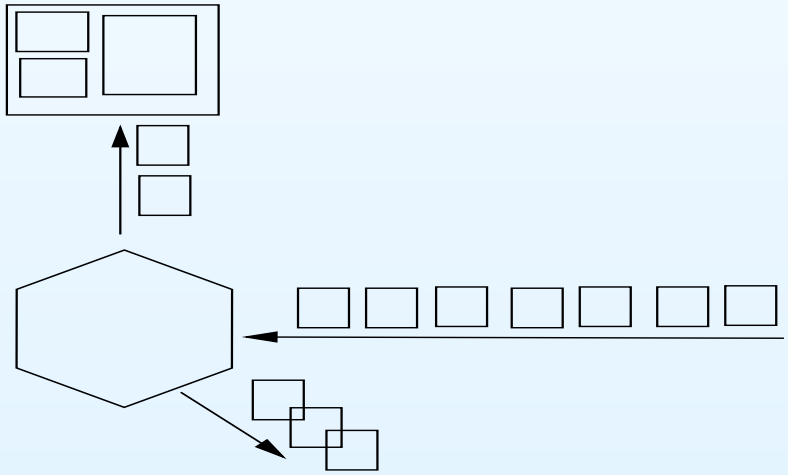
Clasificación Automática de Texto



Bibliografía por Evgeniy Gabrilovich (Technion, Israel).

`http://www.cs.technion.ac.il/~gabr/resources/atc/atcbib.html`

Minería de Datos en Agregadores de Contenido



Software para Minería de Datos

- SAS Enterprise Miner
- SPSS Data Mining
- IBM/DB2 Intelligent Miner
- Microsoft Analysis Services
- Weka (open source), Cluto (libre).

Disciplinas relacionadas a Minería de Datos

- Aprendizaje de Máquinas (Inteligencia Artificial)
- Estadística
- Bases de Datos
- Algoritmos

Minería de Datos vs. Estadística

- La mayoría de las técnicas de minería de datos se desarollaron muchos años atrás en estadística.

Minería de Datos vs. Estadística

- La mayoría de las técnicas de minería de datos se desarollaron muchos años atrás en estadística.
- Una primera diferencia entre las dos áreas es el volúmen de los datos que manejan:
  - Estudios de clasificadores estadísticos (J. Cattlet 1991) consideran como máximo 32000 registros.
  - En minería de datos se procesan bases de datos de millones (miles de millones) de registros.

Minería de Datos vs. Estadística

- Una segunda diferencia es la complejidad de los datos:
  - En estadísticas se consideran datos sobre pequeños conjuntos de variables.
  - En minería de datos se consideran datos sobre miles de variables.
- Otra diferencia está dada por la calidad y origen de la información.
  - En estadística los datos son de alta calidad y prospectivos (se generan con un objetivo determinado).
  - En minería de datos los datos provienen de una base de datos, con todos los problemas que estos significa: valores nulos, ruido, errores, etc.

Volumen de Datos en Minería de Datos

- Muchos algoritmos de minería de datos operan en varias lecturas sobre datos en memoria secundaria.
- Se requieren 50 minutos en leer una base de datos de 30 GB usando un disco duro de 10 MB/seg. de velocidad de transferencia.
- A esta velocidad, no se puede calcular un promedio sobre una base de datos 1 Tera Byte en menos de 28 horas.

Grandes Bases de Datos

- Wallmart maneja aprox. 20 millones de transacciones diarias. Su base de datos de transacciones de venta pesa 11 Tera Bytes.
- AT&T tiene más de 100 millones de clientes y almacena más de 300 millones de llamados diarios.
- El sistema SKYCAT (Fayyad et. al 1996) maneja más de 3 Tera Bytes.
- El log de Yahoo! registra más de 3900 millones de vistas de página diariamente y maneja cuentas de correo para más de 250 millones de usuarios.

Complejidad de los Datos

- **Búsqueda de Asociaciones:** encontrar canastas de productos que se venden con una frecuencia mayor que k en la base de datos de transacciones de un supermercado.
- Algoritmo básico:
  - Tenemos una base de datos de transacciones, cada transacción es un conjunto de productos.
  - Usar una tabla de hash con una entrada para cada canasta posible. La tabla guarda un contador por canasta.
  - Leer la base de datos de transacciones. Para cada transacción, actualizar todas las canastas contenidas en la transacción.

Complejidad de los Datos

- Para n productos, tenemos 2^n canastas frecuentes posibles.
- Para Wallmart $n = 100000$, es decir necesitamos 2^{100000} contadores: impracticable.

Programa del Curso

- Fundamentos:
  - Probabilidad y Estadística.
  - Clasificación
  - Segmentación
  - Asociaciones
  - Regresión
  - Visualización de datos
  - OLAP
- Aplicaciones:
  - Minería de Texto.
  - Extracción de Información.
  - Segmentación y filtrado de contenido.

Evaluación

- Controles y examen ($50\,\%$)
- Proyecto ($50\,\%$)
  - Puede ser aplicado o de investigación.
  - Presentación propuesta y revisión bibliográfica.
  - Presentación final proyecto.
- Hay nota de participación en clases.