

Introducción a la Minería de Datos

Profesores: Bárbara Poblete y Felipe Bravo

Presentación

- Equipo/Alumnos
- ¿Motivaciones?
- ¿Expectativas?
- Expectativas sobre participación de los alumnos
- Reglas del curso: Laboratorios, Proyectos
- suscribirse al calendario del curso

Metodología

1. Clases de cátedra
2. Laboratorios prácticos: grupos de 2 personas
3. Proyecto (en grupos): evaluado en **3 hitos** (o etapas)
 - **Hito 1:** presentación + informe parcial 1
 - **Hito 2:** presentación + informe parcial 2
 - **Hito 3:** presentación + informe final
4. El informe es incremental y se presenta en formato de una página Web

Evaluación

- Hay **Nota Proyecto** y **Nota Laboratorios** (No hay Controles)
- **Nota Final** = $NP \cdot 0,4 + NL \cdot 0,6$
- **Nota Laboratorios** = Promedio Laboratorios (incluye nota de preguntas individuales teóricas)
- **Nota Proyecto** = Promedio (H1, H2, H3)
- **Requisitos de aprobación** del curso (se deben cumplir TODOS):
 - $NP \geq 4.0 \text{ && } NL \geq 4.0$
 - Cumplir con un **mínimo de 70%** asistencia a sesiones **Labs** (5/7), pero todos los Labs deben ser realizados y enviados en 2da fecha, excepto faltas justificadas en las cuales se debe acordar una fecha nueva de entrega.
 - Cumplir con un **mínimo 70%** de asistencia a sesiones de **Presentaciones Proyecto** (5/7), pero sólo en días que no le toque presentar a su grupo.
 - **No se tolerarán copias** en Laboratorios ni Proyectos, es causal de Nota Final = 1.0 (R)

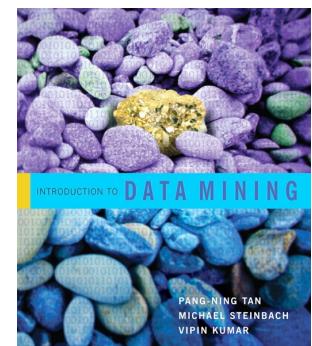
Leer en detalle

Para la próxima clase traer leídos (en u-cursos)

- Reglas de los labs
- Reglas de los proyectos

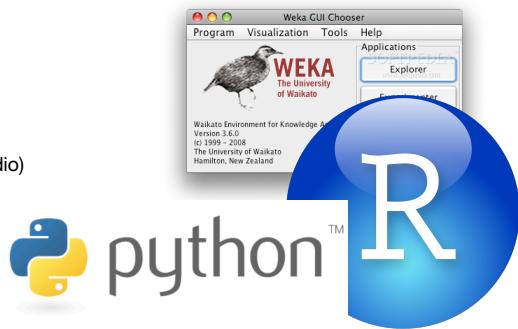
Libro del Curso

- Introduction to Data Mining
- Autores: Pang-Ning Tan, Michael Steinbach, Vipin Kumar



Herramientas del curso

- WEKA
- R (R Studio)
- Python



Objetivos del curso

- Curso **introductorio**
- Aprender a aplicar el proceso de DM a datos reales
- Conocer, seleccionar y utilizar las técnicas básicas de DM
- Aprender a interpretar los resultados de estos procesos
- Proveer la base para adquirir conocimiento más avanzado

¿Qué es la Minería de Datos?

- Descubrir automáticamente información útil en grandes repositorios de datos

¿Qué es la Minería de Datos?

- Descubrir **automáticamente** información útil en grandes repositorios de datos

10

¿Qué es la Minería de Datos?

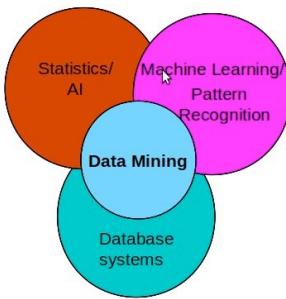
- Descubrir automáticamente información **útil** en grandes repositorios de datos

¿Qué es la Minería de Datos?

- Descubrir automáticamente información útil en **grandes repositorios** de datos

Orígenes de la MD

- Une ideas de ML/AI, reconocimiento de patrones, estadística y BD
- Enfoques tradicionales fallan con datos masivos (alta dim., datos heterogéneos y distribuidos)



¿Cuál es la diferencia entre Data Science, Machine Learning e Inteligencia Artificial?

- Están de moda, pero no son lo mismo, ni son intercambiables
- **Data Science** es el nombre reciente para algo mucho más antiguo: **Data Mining (90's)**
- Definición (sobre) simplista:
 - **Data mining** genera **entendimiento**.
 - **Machine learning** genera **predicciones**.
 - **Artificial intelligence** genera **acciones**.

14 <http://varianceexplained.org/r/ds-ml-ai/>

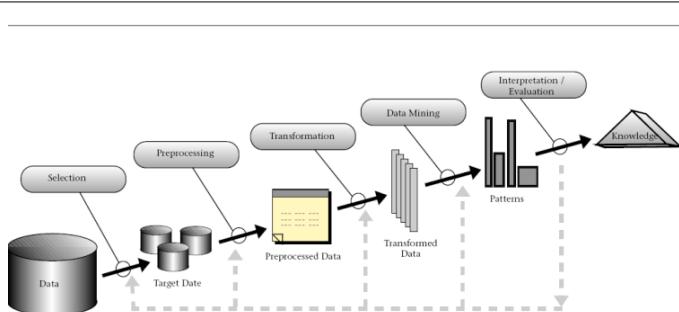
¿Cuál es la diferencia entre Data Science, Machine Learning e Inteligencia Artificial?

- Las definiciones tampoco sirven para describir el trabajo de alguien:
 - "Yo soy *Data Scientist*" no depende de lo que uno haga, sino que de experiencia que se tenga y enfoque principal del trabajo que se hace.
- El hecho que alguien escriba no lo convierte en escritor.

¿Por qué es importante entender estas diferencias?

- Porque este **no es un curso de Machine Learning**, es un **curso de Minería de Datos**.
- **ML: Estudio, diseño y desarrollo de algoritmos** que permiten a los computadores aprender sin ser explícitamente programados (Arthur Samuel). Técnicas genéricas, aplicables a varios dominios.
- **Minería de Datos:** El enfoque está en **extraer conocimiento**, o patrones previamente desconocidos, a partir de (grandes) volúmenes de datos (en su mayoría no estructurados). Para esto se pueden utilizar técnicas de ML, entre otras. Requiere conocimiento de los datos mismos y su dominio.

16

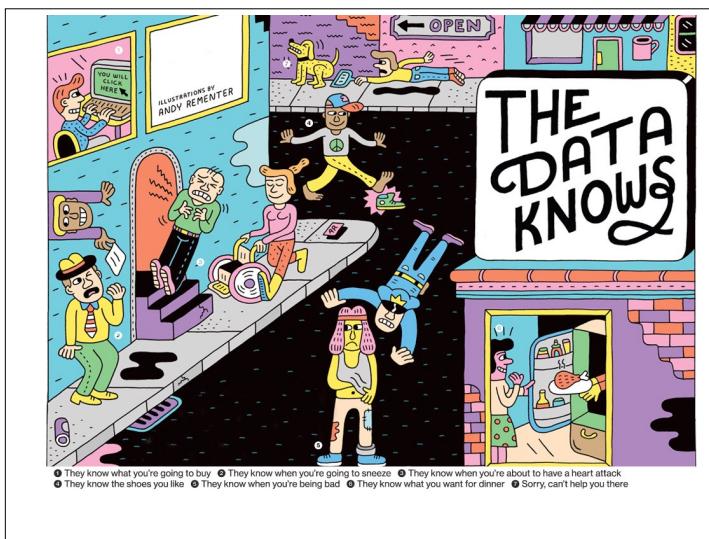


Knowledge Discovery
in Databases (KDD)

BIG BANG

- 2006 Hadoop
- Análisis de datos masivos al alcance de todos (cientos de start-ups)





¿Por qué hacer minería de datos?

- Aspecto comercial
- Aspecto científico



¿Por qué hacer minería de datos?

¿Motivación Comercial?

- Recolección de MUCHOS datos comerciales:
- Datos Web, e-commerce
- Compras en tiendas
- Transacciones en Bancos/ Tarjetas de Crédito

amazon

Your Amazon.com Account Your Shopping History Today's Deals Gift Cards Sell Help

Barbara's Amazon ON ORDER 0 items AMAZON PRIME AMAZON APPS & SERVICES YOUR RECOMMENDATIONS YOUR PROFILE LEARN MORE

Barbara's Amazon Try Prime Try Audible View Benefits

AMAZON PRIME MEMBER SINCE 2008

Recommended for you, Barbara

Literature & Fiction 100+ items	Science Fiction & Fantasy Books 61+ items	Prime Video – Unlimited Streaming for Prime Members 17+ items	Mystery, Thriller & Suspense Books 50+ items
Personal Care Products 100+ items	Recommended Based On Sketching User Experiences: Getting the Design Right 16+ items	Office & School Supplies 100+ items	Cell Phones & Accessories 100+ items

1 X Disney Frozen Pencil Case

by American Greetings, LLC

Price: \$22.55

In Stock.

This item ships to Santiago, Chile. Want it Friday, March 11? Order within 9 hrs 49 mins and choose Amazon's Priority Shipping at checkout. Learn more

Shipping rates and delivery times are subject to change. Gift wrap available

Purchase Quantity: 1

Style Name: Purple

1 Disney Frozen Pencil Case

15 new from \$1.50

Up to 60% off Select Post-it products. Shop now

Roll over image to zoom in.

Frequently Bought Together

Total price: \$22.55

Add to Cart

Add both to Cart

This Item: 1 X Disney Frozen Pencil Case \$22.55

Thermos 12-Ounce Funtainer Bottle, Frozen Purple \$17.25

Customers Who Bought This Item Also Bought

Disney Frozen Light Blue Backpack with Case (15 Pint) \$14.99

Disney Frozen Rolling 16" Backpack Bag Lunchbox 2pc \$14.99

Disney Frozen 1 Subject Wide Ruled Notebook Set (12 Count) \$14.97

Disney Frozen Elsa and Anna Stationery Set with Pencils \$14.99

American Greetings Frozen Party Favor Assortment, Purple \$12.99

Thermos 12-Ounce Funtainer Bottles \$17.25

NETFLIX

Browse Kids

TV Thrillers & Mysteries

Romantic Movies

Continue Watching for Barbara

Watch It Again

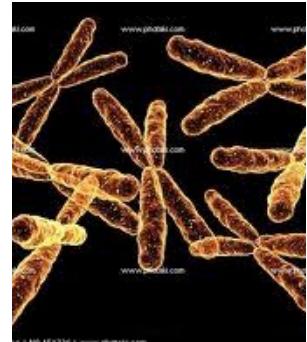
Top Picks for Barbara

The screenshot shows the Netflix homepage with a banner for 'Top Picks for Barbara'. Below it, a section for 'House, M.D.' is displayed, followed by a grid of other TV show thumbnails. At the bottom, there's a 'Because you watched' section with recommendations like 'The Mind of a Chef' and 'Chef's Table'.

¿Por qué hacer minería de datos?

¿Motivación Científica?

- Datos (observaciones) recolectadas a gran velocidad (GB/hr, Tb/día)
- Telescopios, Satélites, Requerimientos Web, ADN, etc ([Google Flu Trends](#))



The screenshot shows the Google Flu Trends website. It features a world map where countries are colored according to flu activity levels: United States (High), Canada (High), Mexico (High), Russia (High), and others in shades of yellow and green. A legend on the left defines activity levels from Intense (dark blue) to Minimal (light green). The URL is google.org/FluTrends.

Métodos utilizados en DM

- **Métodos predictivos:** Usar variables para predecir variables desconocidas o valores futuros de otras variables
- **Métodos descriptivos:** Encontrar patrones interpretables por humanos que permitan describir los datos



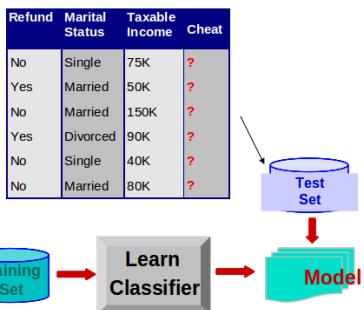
Métodos utilizados en DM

- Clasificación (Predictivo)
- Clustering (Descriptivo)
- Descubrimiento de Reglas de Asociación (Descriptivo)
- Descubrimiento de Patrones Secuenciales (Descriptivo)
- Regresión (Predictivo)
- Detección de Desviación (Predictivo)

Clasificación

- Set de Entrenamiento (atributos incluyendo clase)
- Busca modelar en atributo clase
- Objetivo: asignar la clase más correcta a records nuevos
- Set de Evaluación

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



Clasificación: Aplicación 1

- Marketing directo
- Meta: Reducir costos de publicidad apuntando directamente a potenciales compradores.
- ¿Cómo?

Clasificación: Aplicación 2

- Detección de Fraude
- Meta: Predecir transacciones fraudulentas en el uso de tarjetas de crédito
- ¿Cómo?

Clasificación: Aplicación 3

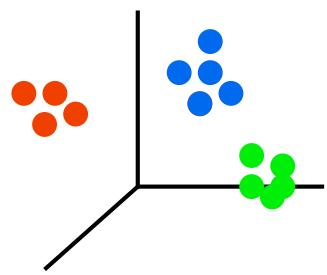
- Fidelidad de Clientes
- Meta: Predecir si es posible perder a un cliente a la competencia
- ¿Cómo?

CLUSTERING

- Conjunto de puntos (datos), cada uno con un set de atributos y una medida de similitud
- Encontrar conjuntos tales que:
 - Puntos en un *cluster* sean más similares entre sí
 - Puntos en conjuntos diferentes sean menos similares entre sí

Visualización de clustering

- Clustering 3D basado basado en distancia Euclidiana
- Distancia intra-cluster es minimizada
- Distancia inter-cluster es maximizada



Clustering Aplicación 1

- Segmentación de mercado
- Meta: Subdividir un mercado en subconjuntos de clientes en donde cualquier conjunto es un potencial objetivo de marketing (ej: Netflix, Amazon)
- ¿Cómo?

Clustering Aplicación 2

- Clustering de documentos
- Meta: Encontrar grupos de documentos que son similares entre sí, basándose en las palabras más importantes que contienen. (Directorios, Wikipedia)
- ¿Cómo?

Ejemplo

- Clustering de puntos: 3204 artículos del L.A. Times
- Medida de similitud: cuántas palabras tienen en común estos documentos (después de filtrar algunas palabras).

Category	Total Articles	Correctly Placed
Financial	555	364
Foreign	341	260
National	273	36
Metro	943	746
Sports	738	573
Entertainment	354	278

Reglas de Asociación

- Dado un conjunto de records, cada uno contiene un número de elementos de una colección determinada
- Objetivo: Producir reglas de dependencia que predecirán la ocurrencia de un elemento (ítem) basándose en ocurrencias de otros ítems.

Reglas de Asociación

TID	Items
1	Pan, Coca-cola, Pañales, Leche
2	Cerveza, Pan
3	Cerveza, Coca-cola, Pañales, Leche
4	Cerveza, Pan, Pañales, Leche
5	Coca-cola, Pañales, Leche

Reglas de Asociación Aplicación 1

- Promoción de Marketing y Ventas
 - Sea la regla encontrada del tipo
 $\{Queso, \dots\} \rightarrow \{PapasFritas\}$

Patrones secuenciales

- Dado un set de objetos asociados a una línea de tiempo de eventos, encontrar los elementos que tengan fuertes dependencias secuenciales entre ellos
- Se forman reglas descubriendo patrones y luego se aplican restricciones de tiempo

Regresión

- Predecir el valor de una variable continua, en base a valores de otras variables, asumiendo modelo de dependencia lineal o no-lineal.
- Estadística y redes neuronales

Detección de desviación/anomalía

- Detectar desviaciones significativas de los valores normales

Desafíos de DM

- Escalabilidad
- Dimensionalidad
- Datos complejos y heterogéneos
- Calidad de los datos
- Distribución de los datos y propiedad
- Privacidad
- Streaming

Próxima Clase

- Leer reglas del curso: ver que no haya problemas con los requisitos de asistencia, se entiende que Ud. puede cumplirlos si sigue en el curso.
- Hacer el tutorial que publicarán los auxiliares.
- Bonus track ver el video de [Hans Rosling](#).