INTRODUCCIÓN

MINERÍA DE DATOS
Grado en Ingeniería Informática de Gestión y Sistemas de Información
Escuela de Ingeniería de Bilbao (EIB)

Alicia Pérez
alicia.perez@ehu.eus
Departamento de Lenguajes y Sistemas Informáticos

Universidad del País Vasco
Euskal Herriko Unibertsitatea

Índice

Motivación
Definiciones: DM, KDD, BI

Ejercicio

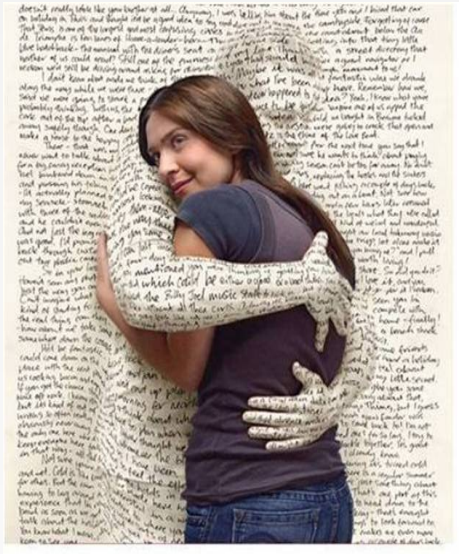
Buscar en distintas fuentes de la bibliografía proporcionada las siguientes definiciones (incluir la fuente).

- Data Mining:
- Knowledge Discovery from Data:
- Business Intelligence:

Motivación
Objetivo: datos → conocimiento

"We are drowning in data, but starving for knowledge!"

(John Naisbitt)



Fuente de la figura:
http://kamafig.wordpress.com

Motivación
Medio: aprendizaje automático

Solución: Aprendizaje automático (*machine learning*)

Clave: aprender automáticamente a partir de ejemplos (datos)

- generar conocimiento automáticamente
- diseñar un modelo (matemático, abstracto) que describa y generalice los datos
- datos particulares de una tarea $\rightarrow$ modelo compacto y general de la tarea

Ejemplos de aplicación en *business intelligence*(BI):

- Física: $\vec{F} = m\vec{a}$
- Google search:
  - *I'm feeling lucky*
  - *Did you mean:* ("Quizás quiso decir...")
- Amazon:
  - *Customers Who Bought This Item Also Bought*
  - *Customers Viewing This Page May Be Interested in These Links*
- Más ejemplos... [Bielza-Larrañaga (Chap. 1)]

Motivación
Campos de aplicación

Ejercicio: ¿dónde se utiliza la minería de datos?

Buscar en distintas fuentes aplicaciones de la minería de datos (incluir la fuente). Debajo aparecen ejemplos descritos en [Alpaydin, 2010], [Bielza-Larrañaga (Chap. 1, pg. 43-60)]

- Ventas (*retail*):
  - Distribución de productos y análisis de afinidad
  - La administración basada en la relación con los clientes (CRM)
- Finanzas:
  - *Credit scoring*
  - Detección de fraude
- Medicina:
  - Diagnósticos médicos
  - Biometría
- Web mining:
  - Ataque a bases de datos
- ...:

Contexto de aplicación: Business Intelligence

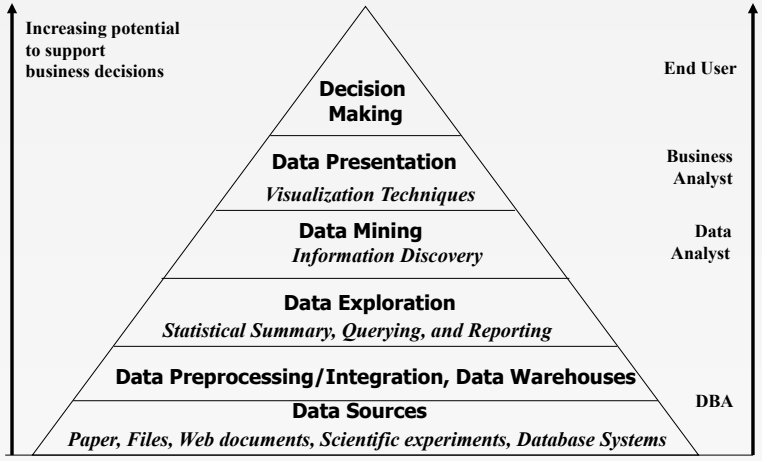
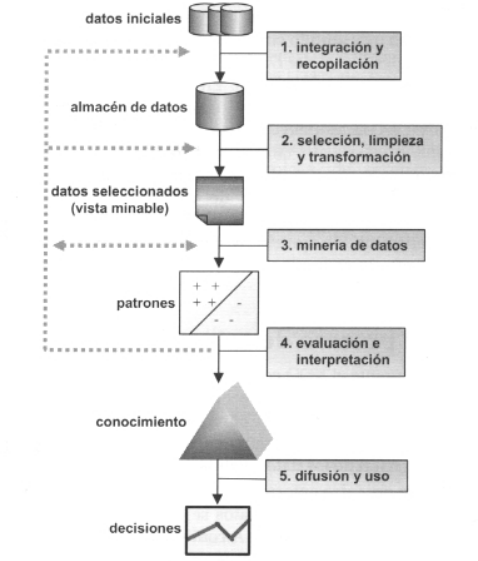


Figura: Contexto de aplicación de la Minería de Datos en Inteligencia Empresarial (BI). Fuente de la figura: [Han et al., 2011]

Contexto de aplicación: Business Intelligence
Proceso KDD



1. Recopilación de datos
2. Selección, limpieza y transformación
3. Minería
4. Evaluación
5. Uso

Paradigmas de aprendizaje [Witten et al., 2011, Han et al., 2011]

1. Generalización
2. Asociación / Correlación
3. Clasificación (clasificación supervisada)
4. Clustering (clasificación no-supervisada)
5. Detección de intrusos (Outlier Analysis)

Clasificación Supervisada

- Disponemos de un conjunto de instancias (objetos, ejemplos . . .),
- estas instancias
  - vienen caracterizadas mediante una serie de n variables predictoras, es decir, una serie de n atributos o características (*feature vector* $\in \mathcal{F}^n$)
  - y cada una de ellas tiene asociada una etiqueta variable o clase (*class* $\in \mathcal{C}$)
- y deseamos estimar un modelo de clasificación capaz de predecir la clase para una nueva instancia dada (la nueva instancia vendrá dada en en términos de las variables predictoras)
- modelo: aplicación sobre el dominio de las variables predictoras y con imagen en el dominio de las clases

$$\mathcal{M} : \mathcal{F}^n \longrightarrow \mathcal{C} \tag{1}$$

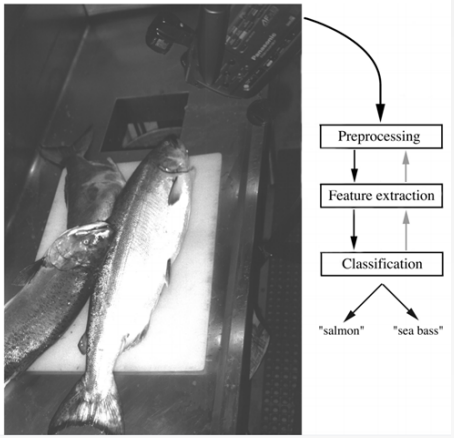


Figura: Preprocessing.
Feature extraction (width, lightness).
Classification(salmon or sea bass?).
Fuente: [Duda et al., 2000]

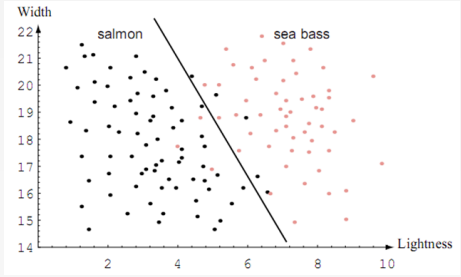


Figura: Classification
Aprox: $\mathcal{M}(w, l)$: *salmon* $\Leftrightarrow w(l) < w_0 - m \cdot l$
Fuente: [Duda et al., 2000]

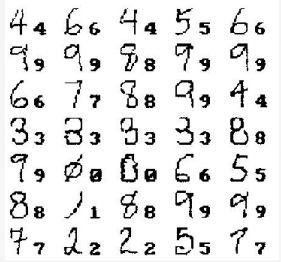


Figura: Clasificación. OCR
[Bielza-Larrañaga (Chap. 1)]

Figura: Clasificación.
Biometría. Fuente: TAS

Figura: Regresión.
Predicción. Finanzas.
Fuente: Wikimedia Commons

Reconocimiento de formas
Paradigmas de aprendizaje

Clasificación No-supervisada

- Disponemos de un conjunto de instancias (objetos, ejemplos . . .),
- estas instancias
  - vienen caracterizadas mediante una serie de n variables predictoras, es decir, una serie de n atributos o características (*feature vector* $\in \mathcal{F}^n$)
- y deseamos generar grupos (*clusters*) de instancias de modo que
  - las instancias del mismo grupo tengan gran similitud entre sí (estén a poca distancia)
  - las instancias de distintos grupos tengan poca similitud (estén a mucha distancia)
- Observación: se requiere una métrica de similitud para las instancias (definición de distancia en el espacio de los atributos)

Reconocimiento de formas
Paradigmas de aprendizaje

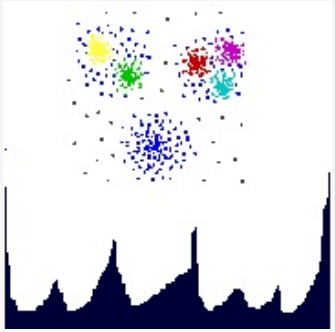


Figura: Clustering (estadístico). Detección de especies y variedades de flores

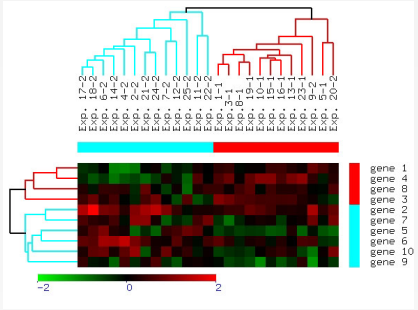


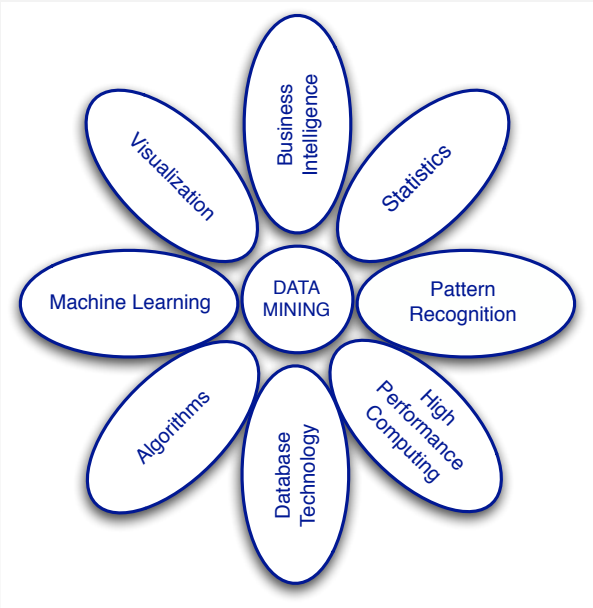
Figura: Clustering (microarrays).Genética

Reconocimiento de formas
Técnicas

[Bielza-Larrañaga (Chap. 1, pg. 30-42)]

- Árboles de clasificación
- Clasificadores k-NN
- Regresión logística
- Redes Bayesianas
- Redes neuronales
- Máquinas de soporte vectorial (SVM)
- Clustering: particional; jerárquico

Relaciones con otras disciplinas



Relaciones con otras disciplinas

Historia

Evolution of Sciences: New Data Science Era [Han et al., 2011]

- Before 1600: Empirical science
- 1600-1950s: Theoretical science
  - Each discipline has grown a theoretical component.
  - Theoretical models often motivate experiments and generalize our understanding.
- 1950s-1990s: Computational science
  - Over the last 50 years, most disciplines have grown a third, computational branch (e.g. empirical, theoretical, and computational ecology, or physics, or linguistics.)
  - Computational Science traditionally meant simulation. It grew out of our inability to find closed-form solutions for complex mathematical models.
- 1990-. . . : Data science
  - The flood of data from new scientific instruments and simulations
  - The ability to economically store and manage petabytes of data online
  - The Internet and computing Grid that makes all these archives universally accessible
  - Scientific info. management, acquisition, organization, query, and visualization tasks scale almost linearly with data volumes
- Now: data mining is a major new challenge!

Bibliografia I


Alpaydin, E. (2010).
Introduction to Machine Learning.
MIT Press.


Duda, R. O., Hart, P. E., and Stork, D. G. (2000).
Pattern Classification.
Wiley-Interscience.


Han, J., Kamber, M., and Pei, J. (2011).
Data Mining: Concepts and Techniques.
Morgan Kaufmann, 3rd edition.


Witten, I. H., Frank, E., and Hall, M. A. (2011).
Data Mining: Practical Machine Learning Tools and Techniques.
The Morgan Kaufmann Series in Data Management Systems, 3rd edition.