# Aprendizaje Automático y Data Mining

*Bloque IV*

# DATA MINING

# Índice

- Definición y aplicaciones.
- Grupos de técnicas:
  - Visualización.
  - Verificación.
  - Descubrimiento.
- Eficiencia computacional.
- Búsqueda de patrones temporales.
- Terminología.

# DEFINICIÓN Y APLICACIONES

# Definición y aplicaciones (I)

- Data Mining (minería de datos): técnicas para la extracción de información oculta en grandes bases de datos.
  - Grandes cantidades de información recopiladas en los últimos años (ámbitos financiero, metereológico, telefonía, medicina, investigación, supermercados, etc.).
  - Fácil y barato recopilar información.
  - Se piensa que la información puede ser útil.
  - Pero el gran volumen la hace inmanejable, es imposible extraer la información útil y descartar la irrelevante.

# Definición y aplicaciones (II)

- Dos posibles enfoques para el problema:
  - Tradicional:
    - Análisis manual realizado por un estadístico o un programador.
    - Se requiere personal muy experimentado.
  - Actual:
    - Análisis automático o semi-automático mediante herramientas de fácil uso.
    - No es necesario personal experto.
    - DATA MINING.

- Origen del término DATA MINING
  - Minería de datos: es necesario remover muchos datos (tierra) para extraer algo de información (metal).

# Definición y aplicaciones (III)

- Relación con el aprendizaje automático:
  - Se busca un modelo que explique o se ajuste a los ejemplos recopilados, igual que en aprendizaje automático.
  - Se utilizan modelos similares:
    - Árboles de decisión.
    - Listas de reglas.
    - Métodos bayesianos.
    - Redes neuronales.
  - Principal diferencia: los algoritmos están adaptados para poder trabajar sobre grandes bases de datos.

# Definición y aplicaciones (IV)

- Principales aplicaciones:
  - Marketing: estudio del comportamiento de consumidores a partir de datos recopilados (compra con tarjetas de crédito).
  - Finanzas: estudio de mercados, de productos, de clientes, de préstamos, etc.
  - Medicina: diagnóstico automático a partir de bases de datos con historias clínicas de pacientes.
  - Distribución de energía: pervisiones de demanda a partir de datos históricos.
  - Redes de telefonía o datos: previsiones de demanda, de ocupación de líneas, de anchos de banda utilizados a lo largo del día, etc.
  - Detección de fallos: en cadenas de producción, en centrales de producción de energía, etc.

# GRUPOS DE TÉCNICAS

# Grupos de técnicas (I)

- 3 grupos de técnicas principales:
  - Visualización.
    - Ayudas para el descubrimiento manual de información.
    - Se muestran tendencias, agrupamientos de datos, etc.
    - Funcionamiento semi-automático.
  - Verificación.
    - Se conoce de antemano un modelo y se desea saber si los datos disponibles se ajustan a él.
    - Se establecen medidas de ajuste al modelo.
  - Descubrimiento.
    - Se busca un modelo desconocido de antemano.
    - Descubrimiento descriptivo: se busca modelo legible.
    - Descubrimiento predictivo: no importa que el modelo no sea legible.

# Grupos de técnicas (II)

- Técnicas de visualización:
  - Visualización en 2D de datos multidimensionales.
  - Problema con múltiples atributos:
    - Vibraciones.
    - Temperatura.
    - Horas funcionamiento.
    - Meses desde revisión…
  - Se calculan las distancias entre cada 2 instancias de entrenamiento.

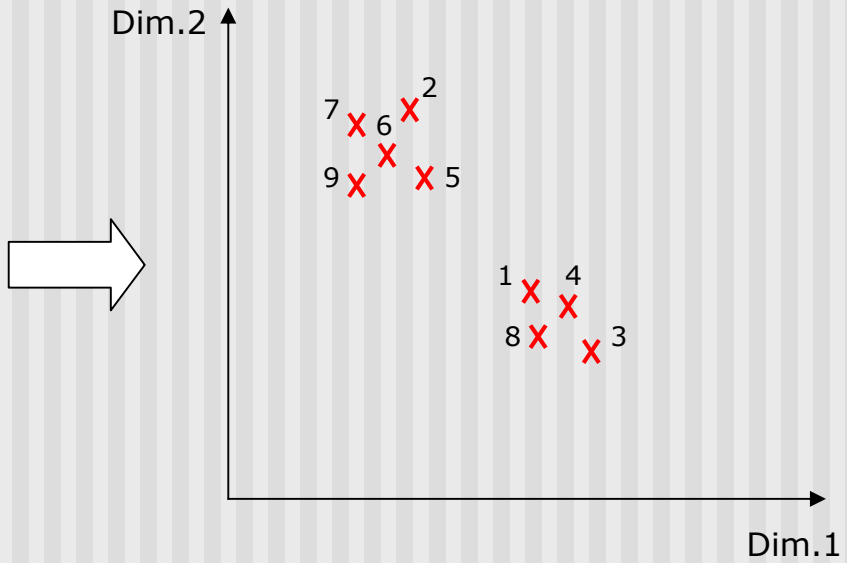
$$d(x_1, x_2) = \sqrt{(temp_1 - temp_2)^2 + (vib_1 - vib_2)^2 + (horas_1 - horas_2)^2 + (rev_1 - rev_2)^2}$$

  - Se crea un gráfico 2D en el que cada instancia corresponde a un punto y en el que las distancias entre instancias son lo más parecidas posible a las distancias reales.

# Grupos de técnicas (III)

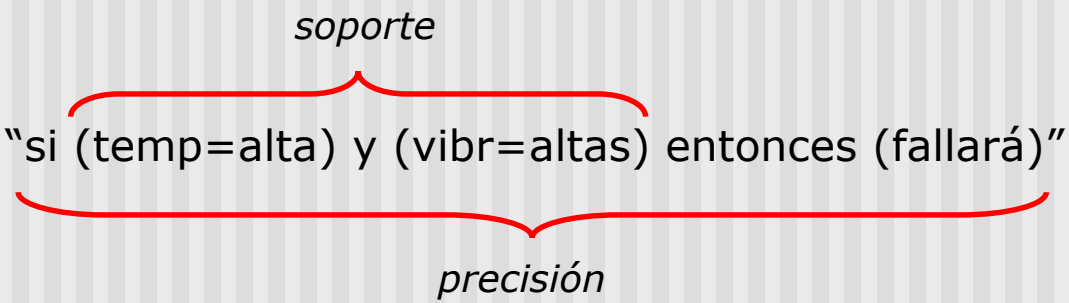
- Ejemplo de visualización:

| | At1 | At2 | At3 | At4 | At5 | At6 |
|---|---|---|---|---|---|---|
| Ej. 1 | - | - | - | - | - | - |
| Ej. 2 | - | - | - | - | - | - |
| Ej. 3 | - | - | - | - | - | - |
| Ej. 4 | - | - | - | - | - | - |
| Ej. 5 | - | - | - | - | - | - |
| Ej. 6 | - | - | - | - | - | - |
| Ej. 7 | - | - | - | - | - | - |
| Ej. 8 | - | - | - | - | - | - |
| Ej. 9 | - | - | - | - | - | - |



# Grupos de técnicas (IV)

- Técnicas de verificación:
  - Se conoce de antemano un modelo y se desea verificar si es un buen modelo para el conjunto de instancias disponible.
  - Medidas utilizadas: soporte y precisión.
    - Soporte: dada una regla, porcentaje de instancias que cumplen sus condiciones.
    - Precisión: dada una regla, porcentaje de casos en los que la regla se cumple.




  - Interesan soporte y precisión altos.

# Grupos de técnicas (V)

- Técnicas de descubrimiento.
  - Descriptivo: se busca un modelo legible.
    - Clustering (agrupamiento).
    - Resumen.
    - Generación de reglas en cualquier formato.
      - Justifican la clase (ej. árboles de decisión).
      - Justifican cualquier relación (ej. reglas de asociación).
  - Predictivo: no importa si el modelo no es legible.
    - Clasificación.
    - Análisis de series temporales (predicción).
    - Regresión.

# EFICIENCIA COMPUTACIONAL

# Eficiencia computacional (I)

- Ejemplo: procesamiento paralelo para crear un árbol de decisión.
- Objetivo: elegir el atributo más apropiado para cada nodo, pero con un gran número de datos.
- Dos posibles estrategias:
  - Con movimiento de datos entre procesadores.
  - Sin movimiento de datos entre procesadores.

# Eficiencia computacional (II)

- Con movimiento de datos entre procesadores.
  1. Reparto aleatorio de ejemplos (instancias) entre los procesadores.
  2. En cada procesador, ajuste de una función de distribución de probabilidad de los valores de los atributos.
  3. Recopilación de resultados (funciones) y envío a un único procesador.
  4. En ese procesador se elige el atributo a utilizar en el nodo correspondiente del árbol.
  5. Se repite el proceso para todos los nodos.

# Eficiencia computacional (III)

- Sin movimiento de datos entre procesadores.
  1. Nodo raíz: se elige el atributo como en el caso anterior.
  2. Los ejemplos correspondientes a cada rama (desde el nodo anterior) se llevan a un conjunto distinto de procesadores.
  3. Cada grupo de procesadores trabaja independientemente sobre su rama del árbol.
  4. El reparto continúa hasta que el número de ramas es igual al número de procesadores.
  5. Cada procesador trabaja independientemente hasta completar su rama del árbol.

# BÚSQUEDA DE PATRONES TEMPORALES

# Búsqueda patrones temporales (I)

- Ejemplo de la dificultad que puede alcanzar la extracción de información en bases de datos.

- Trabajaremos sobre un problema médico:
  - Se dispone de historias clínicas de múltiples pacientes.
  - Se desean extraer secuencias de comportamientos que se repitan con frecuencia.
  - Los atributos son los síntomas detectados o las mediciones tomadas en cada revisión:
    - Temperatura.
    - Presión.
    - Medidas en análisis (porcentajes).
    - Etc.

# Búsqueda patrones temporales (II)

- Algoritmo:
  1. Ordenación temporal de los registros de cada paciente.

| PACIENTE 1 | | | | | | | |
|---|---|---|---|---|---|---|---|
| | ATRIBUTOS | | | | | | |
| DÍAS | - | - | - | - | - | - | - |
| | - | - | - | - | - | - | - |
| | - | - | - | - | - | - | - |
| | - | - | - | - | - | - | - |

| PACIENTE 2 | | | | | | | |
|---|---|---|---|---|---|---|---|
| | ATRIBUTOS | | | | | | |
| DÍAS | - | - | - | - | - | - | - |
| | - | - | - | - | - | - | - |
| | - | - | - | - | - | - | - |
| | - | - | - | - | - | - | - |

| PACIENTE 3 | | | | | | | |
|---|---|---|---|---|---|---|---|
| | ATRIBUTOS | | | | | | |
| DÍAS | - | - | - | - | - | - | - |
| | - | - | - | - | - | - | - |
| | - | - | - | - | - | - | - |
| | - | - | - | - | - | - | - |

| PACIENTE 4 | | | | | | | |
|---|---|---|---|---|---|---|---|
| | ATRIBUTOS | | | | | | |
| DÍAS | - | - | - | - | - | - | - |
| | - | - | - | - | - | - | - |
| | - | - | - | - | - | - | - |
| | - | - | - | - | - | - | - |

| PACIENTE 5 | | | | | | | |
|---|---|---|---|---|---|---|---|
| | ATRIBUTOS | | | | | | |
| DÍAS | - | - | - | - | - | - | - |
| | - | - | - | - | - | - | - |
| | - | - | - | - | - | - | - |
| | - | - | - | - | - | - | - |

| PACIENTE 6 | | | | | | | |
|---|---|---|---|---|---|---|---|
| | ATRIBUTOS | | | | | | |
| DÍAS | - | - | - | - | - | - | - |
| | - | - | - | - | - | - | - |
| | - | - | - | - | - | - | - |
| | - | - | - | - | - | - | - |

# Búsqueda patrones temporales (III)

2. Búsqueda de combinaciones de atributos (síntomas) que se repiten en un mismo día simultáneamente con alta frecuencia (entre todos los pacientes).

3. Se descartan todos los restantes días (combinaciones de atributos que no se repiten con frecuencia).

4. Sobre la lista ordenada restante, se buscan secuencias repetidas.

# Búsqueda patrones temporales (IV)

4. Búsqueda de secuencias repetidas:
   - Múltiples pasadas sobre la base de datos.
   - Se comienza con secuencias de un elemento (las de la etapa anterior).
   - En cada pasada se añade un nuevo elemento para crear nuevas secuencias candidatas.
   - Las secuencias candidatas se evalúan, sólo se mantienen aquellas cuya frecuencia de repetición supera un cierto umbral.

5. Poda de las secuencias:
   - Se eliminan todas las secuencias que están contenidas en otra secuencia más larga.

# TERMINOLOGÍA

# Terminología (I)

- KDD: Knowledge discovery in databases.
- Engloba más aspectos que Data Mining:
  - Preprocesados de los datos:
    - Eliminación de ruidos.
    - Cambios de variable y transformaciones
    - (extracción de características mediante PCA, ICA, etc. Se verá en otra asignatura).
  - Data mining o extracción de información.
  - Procesos posteriores:
    - Interpretación de resultados.
    - Generación de informes.

# Terminología (II)

- EDD: Exploratory Data Analysis.
- Engloba más aspectos que Data Mining:
  - Preprocesados de los datos:
    - Eliminación de ruidos.
    - Cambios de variable y transformaciones
    - (extracción de características mediante PCA, ICA, etc. Se verá en otra asignatura).
  - Data mining o extracción de información.
  - Procesos posteriores:
    - Interpretación de resultados.
    - Generación de informes.

# Terminología (III)

- Text mining.
- Búsqueda de patrones en textos.
  - Búsqueda de documentos similares en bases de datos.
  - Asociación automática de palabras clave (keywords) a documentos.
  - Búsqueda de datos concretos (en tablas, por ejemplo) en grandes bases de datos de documentos.
- En ningún caso son procesos triviales.

# Terminología (IV)

- Web mining.
- Búsqueda de datos en internet.
- Múltiples buscadores: Google, etc.
- No se trata de simples búsquedas en bases de datos.
- El indexado es muy complejo:
  - Por cada palabra (o grupo de palabras), se crean índices indicando el número de ocurrencias en cada documento.
  - Se comprime la información mediante PCA, ICA o RP (se verán métodos en otra asignatura).

*Bloque IV*

# DATA MINING