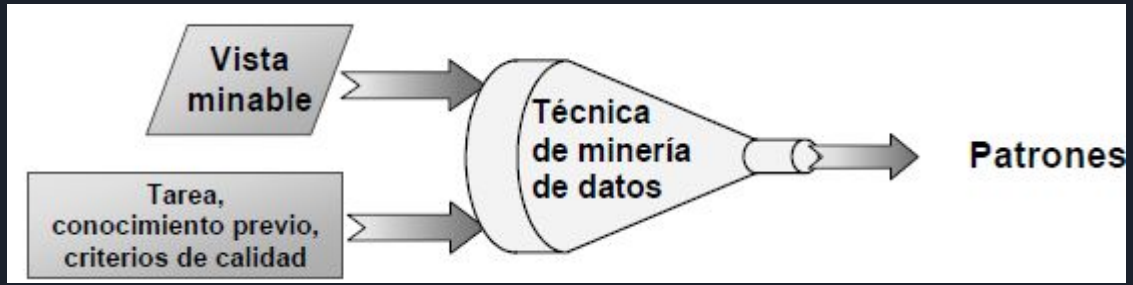
Unidad 6

La extracción de patrones

Modelizado de Minería de Datos - Q22025

Introducción

- **Problema Central**: Descubrir patrones válidos, novedosos e interesantes que, además, sean comprensibles para los humanos.
- **La Dificultad**: El proceso es computacionalmente costoso y depende de si el patrón existe y cuán intrincado es.
- **Dos Aspectos Fundamentales**:
  - **Expresividad**: Qué tan complejo y detallado es un patrón (su capacidad para representar la realidad).
  - **Comprensibilidad**: Qué tan fácil es para nosotros, los humanos, entender ese patrón.



Tareas de la Minería de Datos

Predicción vs. Descripción 🔮📖

¿Qué es una Tarea? El problema a resolver, independiente del método (árboles de decisión, redes neuronales, etc.).

Tareas de la Minería de Datos

Tareas Predictivas: ¿Qué Pasará?

- **Clasificación**: Predecir una categoría (ej., spam vs. no-spam).
- **Clasificación Suave**: Predecir la categoría y la probabilidad de pertenencia (ej., 95% de probabilidad de ser spam).
- **Regresión**: Predecir un valor numérico (ej., el precio de una casa).
- **Preferencias**: Ordenar elementos según un criterio (ej., priorizar un cliente sobre otro).

Tareas de la Minería de Datos

Tareas Descriptivas: ¿Qué Pasó?

- **Agrupamiento** (Clustering): Encontrar grupos de elementos similares sin etiquetas previas (ej., segmentar clientes).
- **Reglas de Asociación**: Descubrir reglas del tipo "Si... entonces..." (ej., si un cliente compra pan, es probable que compre leche).
- **Detección de Anomalías**: Identificar datos o instancias que son significativamente diferentes del resto (ej., detectar fraudes).
- **Correlaciones**: Analizar la relación entre variables (ej., si hay una relación entre la edad y el nivel de ingresos).

Tareas y métodos

De la Tarea al Método: El Problema y la Solución 🛠️

- **¿Cuál es la idea?** Una tarea (el problema) puede resolverse con varios métodos (los algoritmos).
- **La Versatilidad**: La mayoría de los métodos se basan en el aprendizaje inductivo, lo que les permite ser aplicados a diferentes tareas.

Tareas y métodos

Algunos tipos de Métodos (Herramientas):

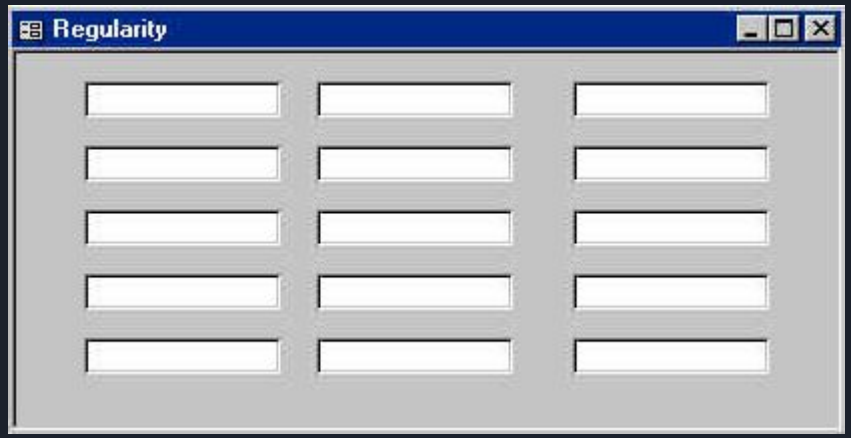
- **Técnicas Estadísticas**: Usan fórmulas matemáticas para encontrar patrones.
  - Uso: Regresión, correlaciones.
- **Técnicas de Frecuencias**: Se basan en el conteo de la co-ocurrencia de eventos.
  - Uso: Reglas de asociación (ej., Apriori).
- **Técnicas de Árboles y Reglas**: Crean modelos fáciles de interpretar en forma de si-entonces o diagramas de flujo.
  - Uso: Clasificación, predicción.

Tareas y métodos

Algunos tipos de Métodos (Herramientas):

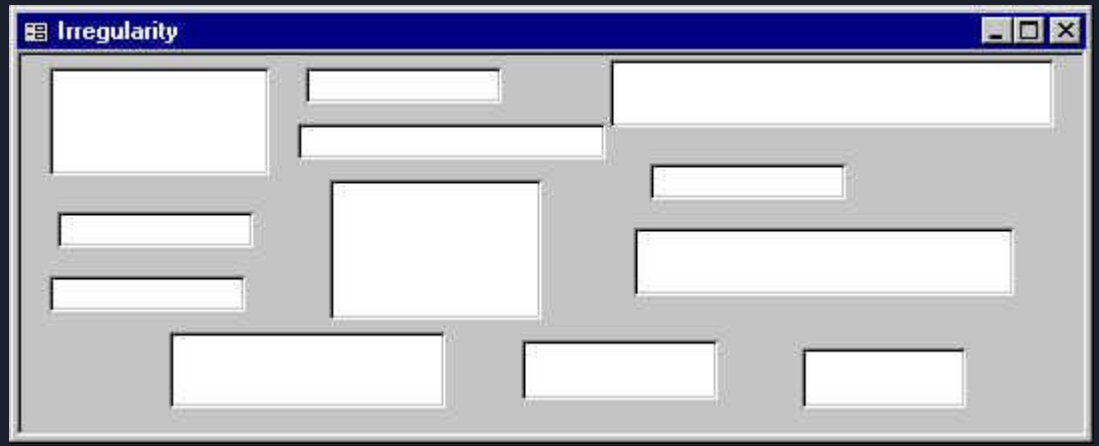
- **Técnicas de Redes Neuronales**: Modelos inspirados en el cerebro. Muy poderosos, pero a menudo difíciles de interpretar.
  - Uso: Clasificación compleja, predicción.
- **Técnicas de Distancia**: Agrupan o clasifican elementos por su similitud o cercanía.
  - Uso: Agrupamiento (ej., K-medias), clasificación (ej., k-vecinos más cercanos).
- **Técnicas de SVM**: Buscan el mejor margen de separación.
  - Uso: SVM para problemas de clasificación complejos
- **Técnicas Bayesianas**: Calculan probabilidad basado en eventos previos.
  - Uso: Clasificador Naive Bayes para detección de SPAM

La Irregularidad en la Visualización de Datos



https://ruby.fccu.edu/courses/cbaise/hci/regularity.htm

La Irregularidad en la Visualización de Datos



https://ruby.fccu.edu/courses/cbaise/hci/regularity.htm

El Valor de la Irregularidad 📈

- ¿Qué es una irregularidad? Una anomalía, una desviación o algo que no encaja en un patrón esperado.
- ¿Por qué es valiosa? A menudo, las irregularidades son más informativas que la normalidad.
  - Un pico de ventas puede ser una oportunidad.
  - Una lectura atípica en datos de salud puede ser un problema médico.
  - Una desviación en el comportamiento puede indicar un fraude.
- Nuestro Superpoder: El cerebro humano es excepcionalmente bueno para detectar estas anomalías en visualizaciones. Por eso, el primer paso en el análisis de datos es la visualización: para encontrar lo que "no cuadra" de un solo vistazo.

Aprendizaje

Aprendizaje deductivo:

- Comprobar la hipótesis para probar la teoría. Va de lo general, a lo específico.

Aprendizaje inductivo:

- Es una forma de razonar, en que la verdad de las premisas apoyan la conclusión, pero no la garantizan. Va de lo específico a lo general

El Aprendizaje: Más Allá de la Definición

En esencia, el aprendizaje es la **mejora del comportamiento a partir de la experiencia**. En minería de datos, esa "experiencia" son los datos, y esa "mejora" es la capacidad de hacer mejores predicciones, descripciones o tomar mejores decisiones.

Existen varias visiones para entenderlo:

- **Visión Práctica**: Aprender es la capacidad de **predecir el futuro** o **explicar el pasado** basándose en la información que tenemos.
- **Visión Teórica:** Aprender es encontrar **patrones** o **regularidades** en los datos. Es una forma de "comprimir" la información, ya que un patrón es una forma más simple de representar una gran cantidad de datos.

Aprendizaje Deductivo: De la Teoría a la Verificación 🧐

- ¿Qué es? Un razonamiento que va de lo general a lo específico.
- El Proceso: Partes de una teoría o regla establecida para probarla en un caso particular.
  - Ejemplo: Teoría: "Toda la publicidad en redes sociales aumenta las ventas."
  - Prueba: Analizas los datos de tu empresa para ver si esa teoría se cumple.
- La Conclusión: Si tus premisas son ciertas, la conclusión es garantizada.
- Enfoque: Es un método de "arriba hacia abajo", ideal para probar hipótesis.

Aprendizaje Deductivo

Aprendizaje deductivo

Aprendizaje Deductivo (Comprobar la hipótesis para probar la teoría):

Teoría (General): Las ofertas de supermercado atraen a más clientes. En general, cuando un supermercado anuncia descuentos significativos, un mayor número de personas acude a la tienda para aprovechar esas ofertas.

Hipótesis (Específica y Comprobable): "Si veo una propaganda de importantes descuentos para el día martes, entonces el día martes el supermercado estará lleno de gente." Esta es una predicción específica basada en la teoría general.

Premisa (Observación Específica): "Veo una propaganda de importantes descuentos para el día martes." Esta es la evidencia o la información inicial que se utiliza para probar la hipótesis.

Aprendizaje deductivo

Conclusión (Resultado Esperado): "Por lo tanto, el martes el supermercado estará lleno de gente." Esta es la predicción que se espera observar si la hipótesis es correcta y la teoría general se aplica en este caso específico.

En este caso deductivo, la lógica va de lo general (la teoría sobre las ofertas) a lo específico (la predicción sobre el martes). Si al ir al supermercado el martes, efectivamente está lleno de gente, esto apoya (pero no prueba de forma absoluta) la teoría general.

Aprendizaje Inductivo: De lo Específico a lo General 🧠

- **¿Qué es?** La base de la minería de datos. Un proceso que va de lo particular a lo general.
- **El Objetivo**: Un algoritmo analiza ejemplos específicos para encontrar patrones y crear un modelo que sea capaz de predecir o clasificar nuevos datos que nunca ha visto.

Visión del Proceso

1. **Observación**: Analizar datos históricos (ejemplos).
2. **Abstracción**: Identificar regularidades o redundancias.
3. **Generalización**: Crear un modelo o regla que explique el patrón.

Aprendizaje inductivo

Aprendizaje Inductivo

Aprendizaje Inductivo (La verdad de las premisas apoyan la conclusión, pero no la garantizan):

Premisa (Observación Específica): "Voy al supermercado el martes y veo mucha gente." Esta es la evidencia o la información inicial que se observa.

Posible Hipótesis (Explicación Tentativa): "Podría existir una oferta hoy martes." Esta es una posible explicación para la observación de mucha gente en el supermercado. Podría haber otras razones, pero una oferta es una explicación plausible.

Conclusión (Generalización Probable): "Por lo tanto, es probable que haya una oferta en el supermercado hoy martes." Esta es la inferencia o la regla general que se induce a partir de la observación específica.

Aprendizaje Inductivo

Teoría (Potencial Generalización Futura): Si esta situación se repite en varios martes diferentes en los que el supermercado está lleno de gente sin que el observador conozca la razón inicial, podría llevar a una teoría más general: "Los martes, cuando el supermercado está lleno de gente, es probable que haya ofertas."

En este caso inductivo, la lógica va de lo específico (ver mucha gente un martes) a lo general (inferir la probable existencia de una oferta). La conclusión es probable pero no necesariamente cierta. Podría haber otras razones para la multitud (un día de pago, un evento especial, etc.). La repetición de esta observación en diferentes instancias podría fortalecer la creencia en la teoría general.

Tipos de Aprendizaje según las Circunstancias

Tipos de Aprendizaje

- **No Incremental (Batch)**: Se aprende de todos los datos al mismo tiempo. Es el más común en minería de datos.
- **Incremental**: El modelo se actualiza continuamente a medida que llegan nuevos datos.
- **Interactivo**: El algoritmo puede "hacer preguntas" para obtener nuevos datos que mejoren su aprendizaje.
- **Por Refuerzo**: Aprende a base de recompensas y castigos por sus acciones.

Evaluación: La Prueba de Fuego del Aprendizaje 🔥

El Juicio Final del Descubrimiento ✅

- **¿Qué es**? La etapa crucial donde se valida la utilidad y la fiabilidad de los patrones encontrados.
- **¿Por qué es vital?** Evita que un modelo solo "memorice" los datos y asegura que realmente ha aprendido una regla general.
- **El Proceso:**
  - Entrenamiento: El modelo aprende de una parte de los datos.
  - Prueba: Se valida usando datos nuevos y no vistos.
- **El Objetivo:** Asegurar que los modelos pueden predecir el futuro y elegir el mejor método para cada problema.

Los Patrones son Hipótesis: La Prueba de la Realidad 🔬

- **¿Qué es?** Todo patrón o modelo es una hipótesis que debe ser validada contra evidencia futura.
- **Evaluar la Predicción:** Se usa la división de datos en conjuntos de entrenamiento y prueba para estimar el rendimiento del modelo en datos que nunca ha visto.
    - Ejemplo: Medir el porcentaje de acierto de un modelo de clasificación con datos nuevos.
- **Evaluar la Compresión:** Un buen modelo es el más simple y conciso, ya que es menos probable que sea una coincidencia.

La Evaluación de los Métodos: ¿Qué Algoritmo Funciona Mejor? 🤔

- La Premisa: No existe un algoritmo universalmente superior.
  - "No hay almuerzo gratis": Un método que funciona bien para un problema podría ser ineficaz para otro.
- Más allá de la Precisión: Se evalúan otros factores cruciales:
  - Estabilidad: La consistencia del rendimiento del método con diferentes muestras de datos.
  - Robustez: La capacidad de manejar datos con ruido o valores faltantes sin que su rendimiento se desplome.
  - La Solución: La experimentación es clave para encontrar el mejor método para una familia específica de problemas.

Comprensibilidad

- **Definición**: La capacidad de un humano para entender y explicar un modelo de minería de datos.
- Método vs. Modelo:
  - **Método**: El algoritmo o la técnica usada para aprender (ej., Regresión Lineal).
  - **Modelo**: El resultado concreto del aprendizaje (ej., y = 2.5x + 1.2).
- **Propósito**: Que el conocimiento extraído sea útil y confiable.

La importancia de la comprensibilidad

- **Transparencia**: Nos permite entender cómo un modelo toma sus decisiones.
- **Validación**: Nos deja validar si la lógica aprendida por el modelo tiene sentido en el mundo real.
- **Confianza**: La transparencia del modelo aumenta la confianza en los resultados.
- **Comunicación**: Facilita la explicación de los hallazgos a audiencias no técnicas.

Comprensibles vs. cajas negras

- **Modelos Comprensibles:**
  - **Ventaja**: Fáciles de entender y explicar.
  - **Ejemplos**: Árboles de Decisión (Si... entonces...), Regresión Lineal.

- **Modelos de Caja Negra:**
  - **Ventaja**: A menudo, son más precisos.
  - **Ejemplos**: Redes Neuronales Profundas, Máquinas de Soporte Vectorial.

La Decisión: ¿Qué Elegir?

La Decisión, depende del objetivo: Comprensibilidad para explicación y confianza vs. Precisión para rendimiento puro.

Eficiencia

Definición: Es la rapidez con la que un método de aprendizaje encuentra patrones.

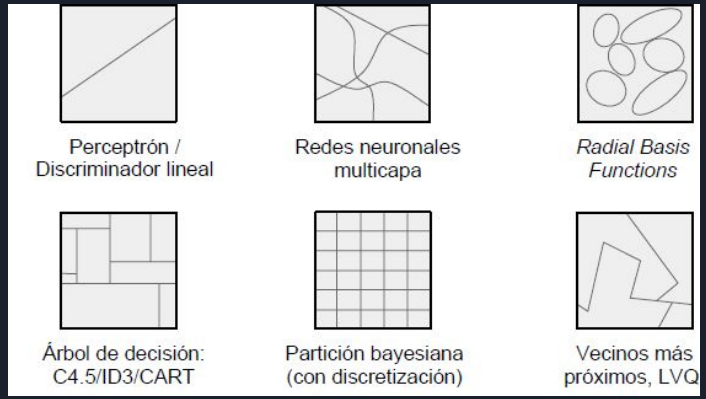
- **Factores Clave**:
  - Tamaño de los Datos: El tiempo de aprendizaje escala de forma no lineal con el número de ejemplos y atributos, llevando a la "maldición de la dimensionalidad".
  - Espacio de Hipótesis: Los métodos más expresivos (como las redes neuronales) tardan más en encontrar patrones porque buscan en un espacio de soluciones más amplio.

La Eficiencia en la Práctica

- **Soluciones para la Ineficiencia:**
  - **Reducción de Datos**: Se usan técnicas como el muestreo y la selección de atributos para trabajar con menos datos y mantener el tiempo de aprendizaje aceptable.
  - **Calidad de los Datos**: Los datos limpios y el conocimiento previo mejoran la eficiencia.

Expresividad: El Lenguaje de los Patrones 🗣️

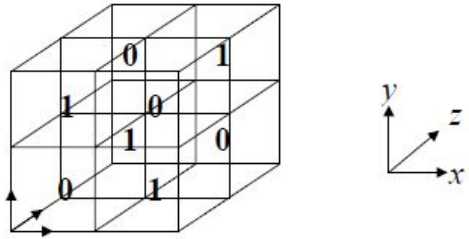
- **¿Qué es?** La capacidad de un método para representar los patrones que encuentra en los datos. Es lo que lo hace más adecuado para un problema que para otro.
- **Fronteras y Patrones**: Muchos métodos funcionan creando fronteras en el espacio de los datos (métodos de "cerca y rellena") para separar las clases.
- **Sobreajuste**: Una alta expresividad puede llevar al sobreajuste, donde el modelo se ajusta demasiado al ruido en lugar de al patrón real.



El Desafío de los Patrones Relacionales

- Limitación: Los métodos de "cerca y rellena" luchan por capturar patrones relacionales que no dependen de la proximidad geométrica de los datos.
  - Ejemplo: La función de la paridad (o "o exclusivo") no puede ser resuelta por estos métodos.
- Soluciones:
  - Aumento de Dimensionalidad: Aumentar las dimensiones de los datos puede hacer que un patrón sea más fácil de separar.
  - Lenguajes Universales: Se utilizan para capturar patrones relacionales complejos.
  - Minería de Datos Relacional: Utiliza la lógica (ej. en la Programación Lógica Inductiva) para aprender y expresar patrones complejos.

x	y	z	→	clase
0	0	0	→	**0**
0	0	1	→	**1**
0	1	0	→	**1**
0	1	1	→	**0**
1	0	0	→	**0**
1	0	1	→	**0**
1	1	0	→	**0**
1	1	1	→	**1**

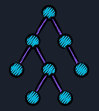


Comparación de Métodos: La Caja de Herramientas del Analista 🛠️

- La Premisa: No hay un "mejor" algoritmo. La elección depende del problema y las características de los datos.
- Factores Clave: Cada método tiene fortalezas y debilidades en términos de:
  - Comprensibilidad: ¿Qué tan fácil es entender el modelo?
  - Expresividad: ¿Qué tan complejos son los patrones que puede capturar?
  - Eficiencia y Robustez: ¿Qué tan rápido y confiable es el algoritmo?

Comparación de Métodos: La Caja de Herramientas del Analista 🛠️

Tipos de Herramientas Principales



- **Árboles de Decisión y Reglas**:
  - Ventaja: Muy comprensibles. Se leen como reglas simples (si-entonces).
  - Desventaja: Expresividad limitada.

- **Redes Neuronales:**
  - Ventaja: Alta expresividad y precisión. Ideales para problemas complejos.
  - Desventaja: Son "cajas negras" (incomprensibles) y sensibles a los datos atípicos.

- **Modelos Estadísticos (Regresión):**
  - Ventaja: Muy eficientes y comprensibles.
  - Desventaja: Su expresividad es limitada para patrones no lineales.

- **Métodos Basados en Distancia (K-Medias, K-Vecinos):**
  - Ventaja: Fáciles de usar y eficientes en datasets pequeños.
  - Desventaja: La presencia de atributos irrelevantes puede afectar su precisión.
- **Métodos Avanzados (SVM, Relacionales):**
  - Ventaja: Capturan patrones complejos (alta expresividad).
  - Desventaja: Menos comprensibles y, a menudo, más lentos.