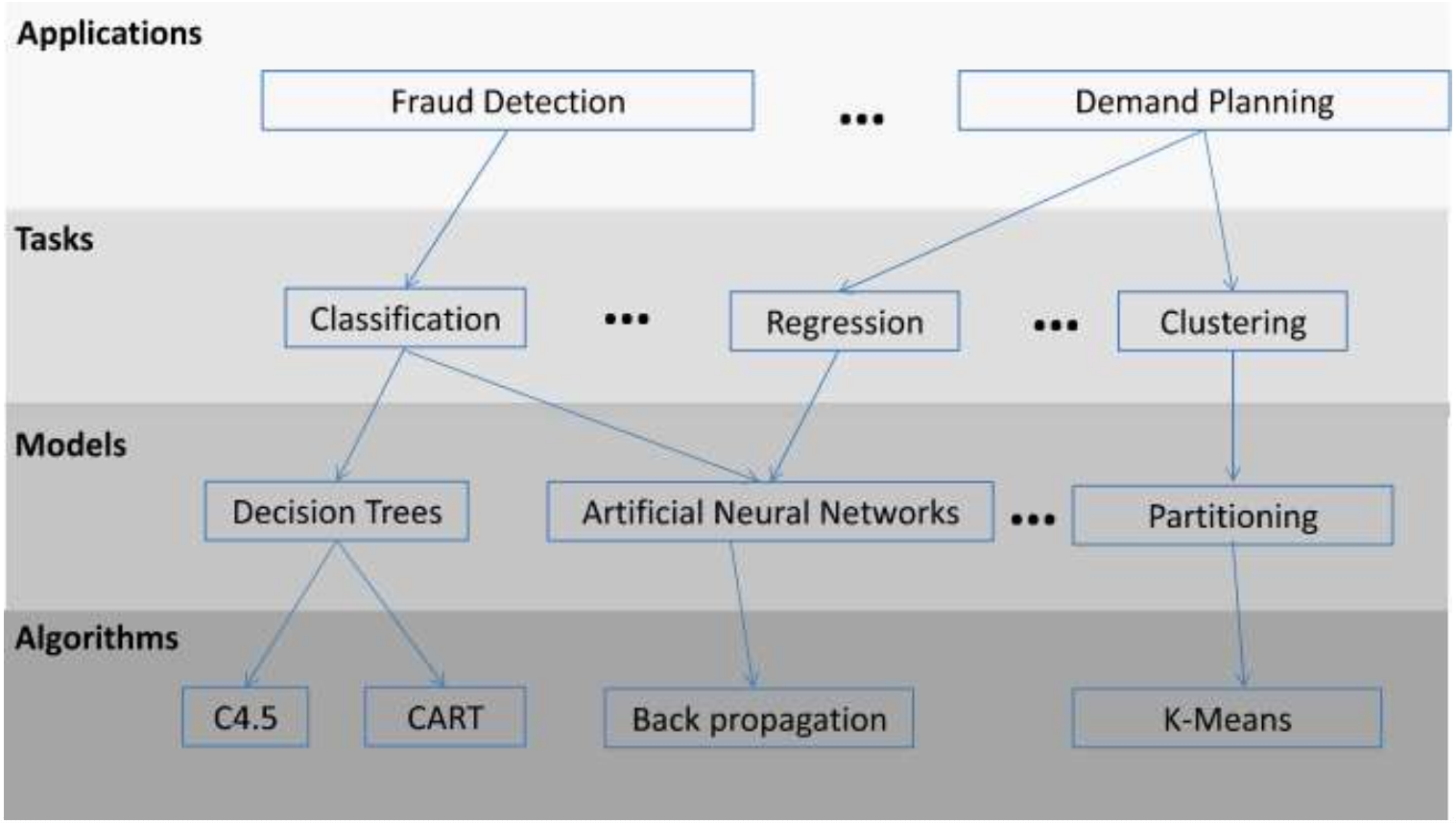
UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
FACULTAD DE CIENCIAS
ALMACENES Y MINERÍA DE DATOS

Minería de datos: Clasificación

Gerardo Avilés Rosas
gar@ciencias.unam.mx

Introducción

- La minería de datos agrupa seis actividades: ***Clasificación, Estimación, Predicción, Asociación, Agrupación, Descripción y Visualización***.
- Las tres primeras tareas son ejemplos de la ***minería de datos dirigida*** o ***aprendizaje supervisado***.



...Introducción

- Las BD contienen una buena cantidad de información escondida que puede ser usada para tomar **decisiones inteligentes**.
- **Clasificación** y **Predicción** son dos formas de análisis de datos que se utilizan para **extraer modelos** que describan importantes clases de datos o predigan tendencias futuras en los mismos.
- Los **modelos de clasificación *predicen*** etiquetas categóricas (*discretas y sin ordenar*):

Préstamo bancario

…Introducción

- Los **modelos de predicción** trabajan con funciones continuas:

- La mayoría de estos algoritmos han sido propuestos en los campos de **máquinas de aprendizaje**, **reconocimiento de patrones** y **estadística**; muchos de ellos **residen en memoria** (*asumen un tamaño pequeño de los datos*).

¿Qué es la clasificación?

- Consiste en **predecir** un **resultado determinado** con base en una **entrada dada**.
- **Asignar** a un objeto una **cierta clase** en función de la **similitud** con ejemplos previos de otros objetos.

¿Qué es la clasificación?

- En cualquiera de estos ejemplos, la tarea de análisis es la **clasificación**, donde un modelo o ***clasificador*** se construye para predecir etiquetas categóricas:

seguro, riesgo, si, no, tratamiento A, tratamiento B, tratamiento C

- Estas categorías pueden se pueden representar por **valores discretos**, donde el ordenamiento entre los mismos no tiene ningún significado.

¿Cómo trabaja la clasificación?

Se trata de un proceso de dos pasos, en el **primero:**

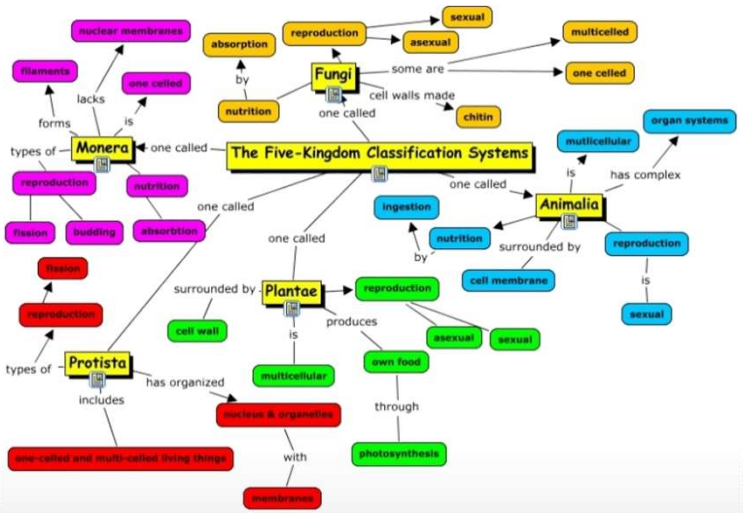
- Se construye un ***clasificador*** que permita describir un número predeterminado de ***clases o conceptos*** (*se conoce como* ***aprendizaje*** o ***fase de entrenamiento***).
- Un ***algoritmo de clasificación*** construye el *clasificador* a través de *analizar o aprender* de un **conjunto de entrenamiento** hecho a partir de tuplas de una BD y sus etiquetas de clase asociadas.
- Una ***tupla X*** representa un ***vector de atributos n-dimensional*** que representa ***n mediciones*** hechas sobre la tupla de **n atributos**.
- Cada ***tupla X*** se supone que pertenece a una clase predefinida y determinada por un atributo de la BD llamado ***etiqueta de clase*** (*atributo categórico en el que cada valor sirve como una categoría o clase*).
- Las tuplas que componen el conjunto de entrenamiento se conocen como ***tuplas de entrenamiento*** (*se seleccionan de la BD a través de un análisis*).

…¿Cómo trabaja la clasificación?

- Dado que las etiquetas de clase son proporcionadas para cada entrenamiento, este paso es también conocido como ***aprendizaje supervisado***.
- Este primer paso puede verse como un ***mapeo o función y = f(X)***, que puede predecir las etiquetas de clase asociadas ***y*** de una ***tupla X*** dada:

  El objetivo del mapeo debe permitir **separar las clases de datos**.

- Típicamente este mapeo se representa a través de:
  - ***Reglas de clasificación***
  - ***Árboles de decisión***
  - ***Fórmulas matemáticas***



…¿Cómo trabaja la clasificación?

Algoritmo de clasificación

nombre	edad	ingreso	decisión_prestamo
Sandy Jones	youth	low	risk
Bill Lee	youth	low	risk
Caroline Fox	middle_aged	high	safe
Rick Field	middle_aged	low	risk
Susan Lake	senior	low	safe
Claire Phips	senior	medium	safe
Joe Smith	middle_aged	high	safe
…	…	…	…

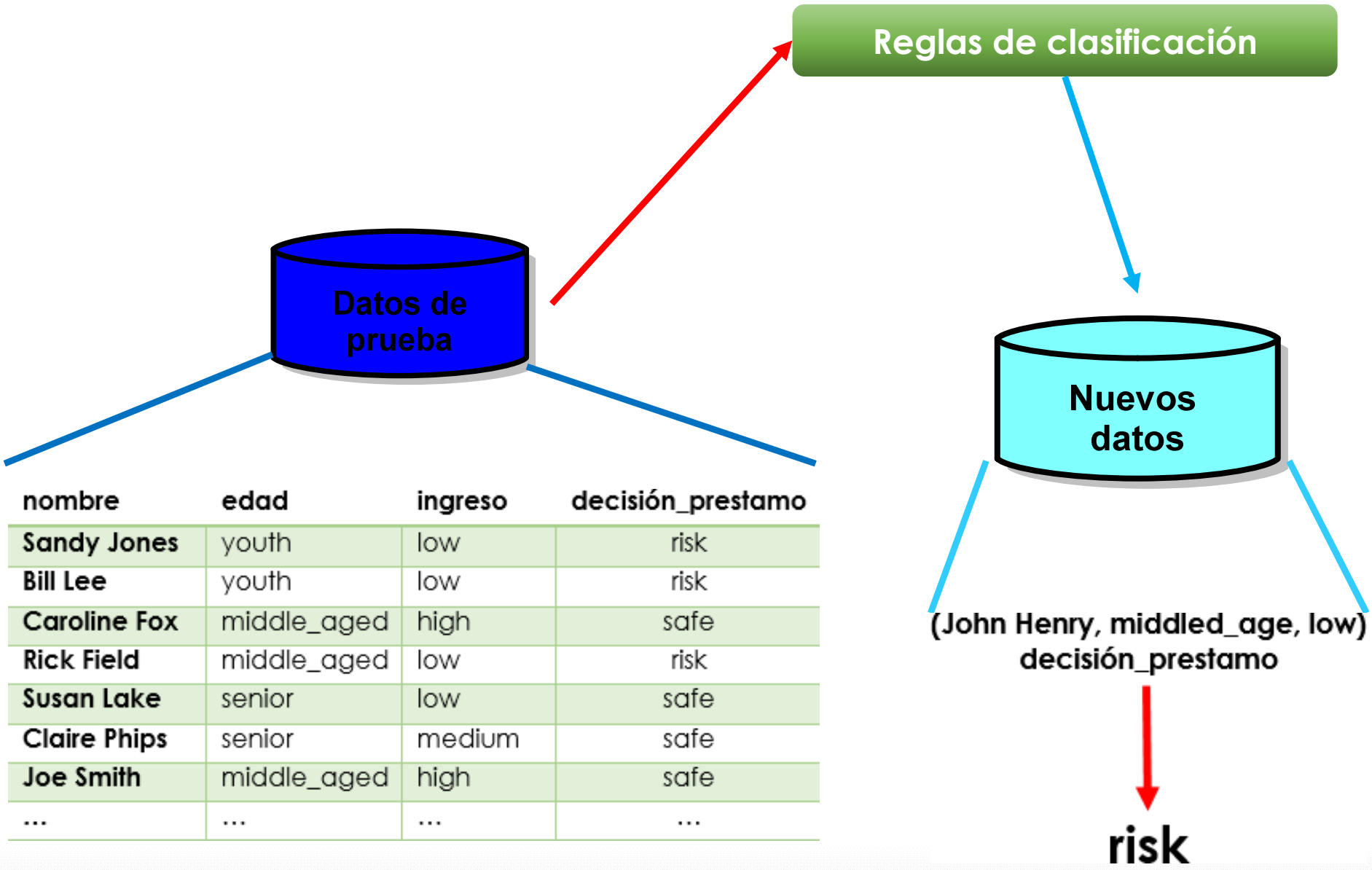
IF edad = *youth* **THEN** decisión_prestamo = *risk*
IF ingreso = *high* **THEN** decisión_prestamo = *safe*
IF edad = *middle_aged* **AND** ingreso = *low*
THEN decisión_prestamo = *risk*
…

…¿Cómo trabaja la clasificación?

En el **segundo paso**:

- El **modelo** se utiliza para clasificar (*se debe estimar la exactitud predictiva del clasificador*).
- Si se utiliza el conjunto de entrenamiento para medir la exactitud, se obtiene una estimación bastante optimista, debido a que el clasificador tiende a ***sobreajustar*** los datos.
- Por esta razón se utiliza un ***conjunto de prueba*** (*tuplas de prueba y sus etiquetas de clase asociadas*). Las tuplas se seleccionan de manera aleatoria y son **independientes** de conjunto de tuplas de entrenamiento.
- La ***exactitud del clasificador*** en un conjunto de prueba dado es el porcentaje de tuplas que son correctamente clasificadas por el **clasificador**. Las etiquetas de clase asociadas de cada tupla son comparadas con las clases que predijo el clasificador en la fase de aprendizaje.
- ***Si la exactitud es aceptable, se puede utilizar para clasificar futuras tuplas***.

…¿Cómo trabaja la clasificación?



nombre	edad	ingreso	decisión_prestamo
Sandy Jones	youth	low	risk
Bill Lee	youth	low	risk
Caroline Fox	middle_aged	high	safe
Rick Field	middle_aged	low	risk
Susan Lake	senior	low	safe
Claire Phips	senior	medium	safe
Joe Smith	middle_aged	high	safe
…	…	…	…

Preprocesamiento de datos

- **Limpieza de datos:**

  Es necesario ***preprocesar*** los datos a fin **de remover o reducir el ruido** y tratar los **valores perdidos** (*missing values*), pues aunque la mayoría de los métodos de clasificación disponen de algunos mecanismos para para manejar este tipo de datos, esto puede ayudar a **reducir la confusión** durante el aprendizaje.

- **Análisis de relevancia:**

  Muchos de los atributos en los datos pueden ser **redundantes** o bien irrelevantes, de manera que es importante detectar a aquellos que no contribuyen con la tarea de clasificación. Este tipo de análisis nos puede ayudar a **mejorar la eficiencia y escalabilidad**.

- **Transformación y reducción de datos:**

  Los datos se pueden **normalizar** (*para datos que involucran mediciones de distancia*) o bien **generalizar** (*principalmente utilizado para atributos que poseen valores continuos*).

Comparación y evaluación

- **Exactitud:**

  Habilidad de predecir las etiquetas de clase de datos nuevos (previamente invisibles). Se estimada usando uno o más conjuntos de prueba que son independientes del conjunto de entrenamiento.

- **Velocidad:**

  Costo computacional involucrado en la generación y uso del clasificador.

- **Robustez:**

  Habilidad del clasificador de realizar predicciones correctas dados datos con ruido o con valores perdidos.

- **Escalabilidad:**

  Habilidad de construir un clasificador que pueda trabajar con grandes cantidades de datos.

- **Interpretabilidad:**

  Nivel de entendimiento y de visión que es proporcionado por el clasificador. Se trata de un aspecto subjetivo y por lo tanto es más difícil de asegurar.