Seminario Permanente de Formación en Inteligencia Artificial Aplicada a la Defensa

Principios Básicos de Machine Learning

Salvador García

Instituto Andaluz de Investigación en Data Science and Computational Intelligence (DaSCI)

Dpto. Ciencias de la Computación e I.A.

Universidad de Granada

salvagl@decsai.ugr.es

http://sci2s.ugr.es

DaSCI

UNIVERSIDAD DE GRANADA

Principios básicos de *machine learning*

- Conceptos básicos. Ciencia de Datos, Minería de Datos, Big Data, Machine Learning
- Proceso de Minería de Datos
- Técnicas de Minería de Datos: Clasificación, Regresión, Agrupamiento, Asociación
- Clasificación y Regresión. Predicción por similitud: K-Nearest Neighbour (KNN).
- Validación de Clasificadores
- Clasificación con Árboles de Decisión

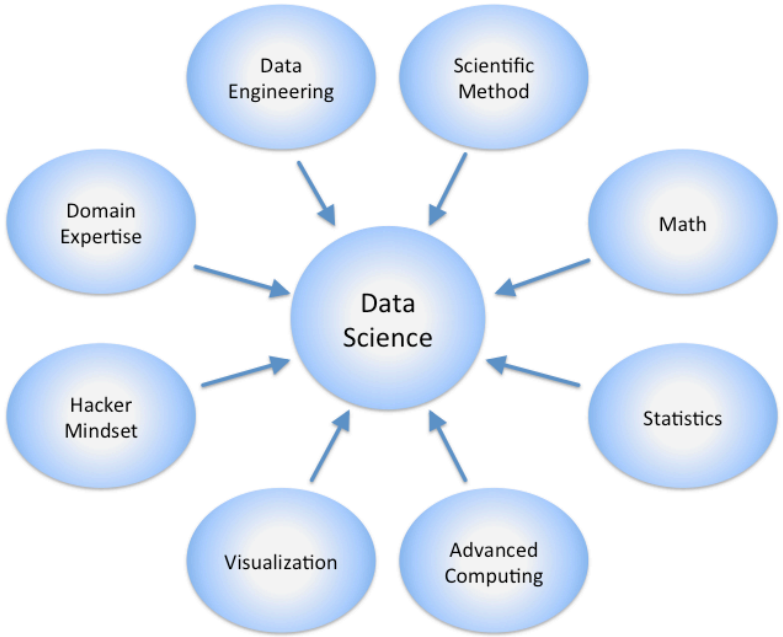
Principios básicos de *machine learning*

- Conceptos básicos. Ciencia de Datos, Minería de Datos, Big Data, Machine Learning
- Proceso de Minería de Datos
- Técnicas de Minería de Datos: Clasificación, Regresión, Agrupamiento, Asociación
- Clasificación y Regresión. Predicción por similitud: K-Nearest Neighbour (KNN).
- Validación de Clasificadores
- Clasificación con Árboles de Decisión

Conceptos básicos

Data Science

Ciencia de Datos es el ámbito de conocimiento que engloba las habilidades asociadas a la extracción de conocimiento de datos, incluyendo Big Data



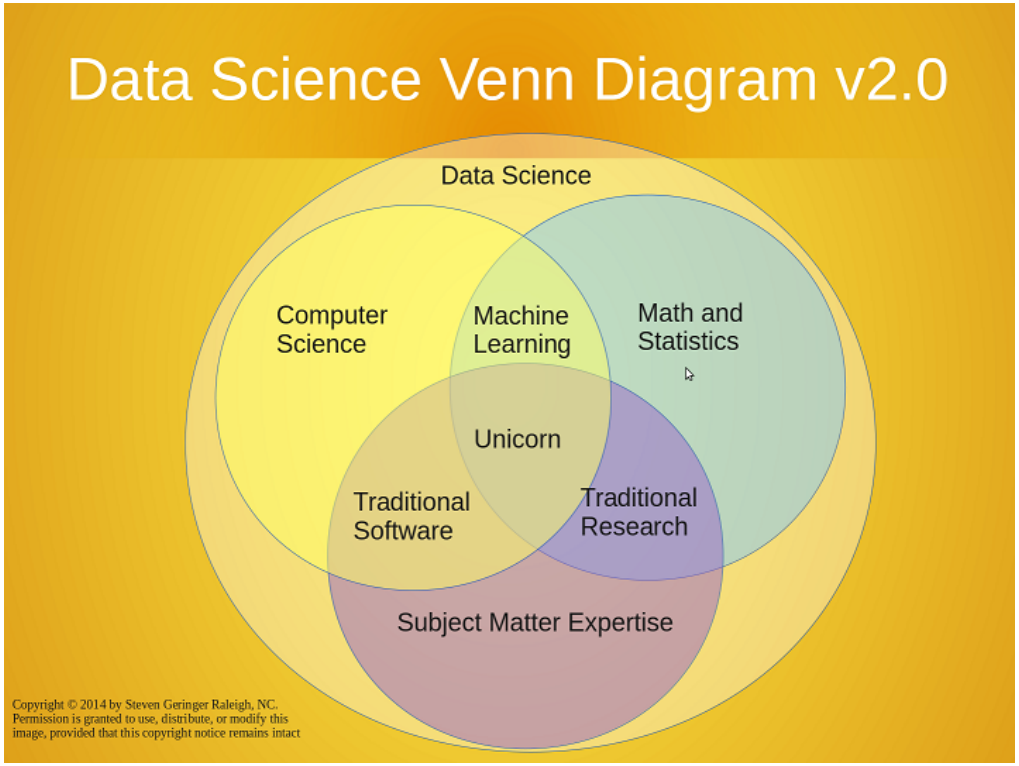
Machine Learning es una disciplina científica del ámbito de la Inteligencia Artificial que crea sistemas que aprenden automáticamente. *Aprender* en este contexto quiere decir identificar patrones complejos en datos.

La máquina que realmente aprende es un algoritmo que revisa los datos y es capaz de predecir comportamientos futuros.

Conceptos básicos

¿Qué es un Científico de Datos?

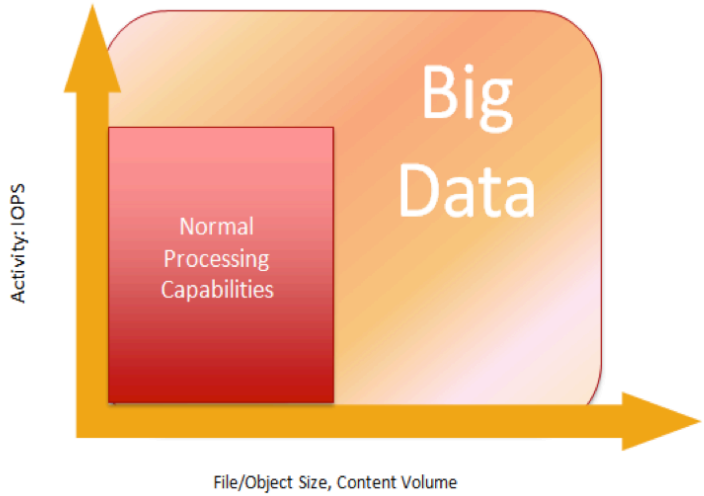
Un científico de datos es un profesional que debe dominar las ciencias matemáticas y la estadística, acabados conocimientos de programación (y sus múltiples lenguajes), ciencias de la computación y analítica.

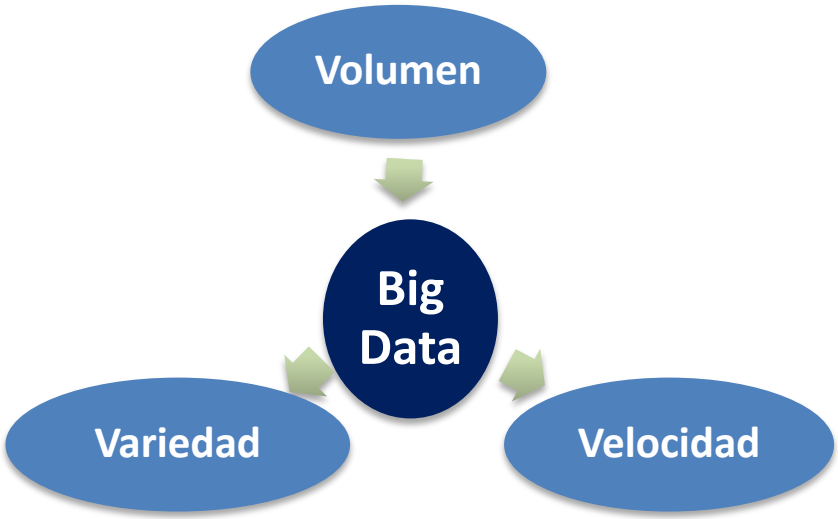


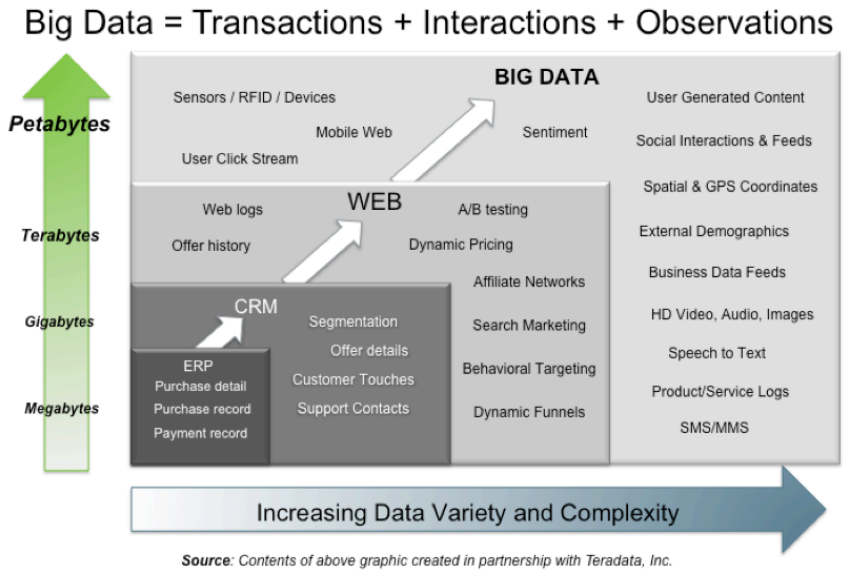
Conceptos básicos

Big Data

"***Big Data***" son datos cuyo volumen, diversidad y complejidad **requieren nueva arquitectura, técnicas, algoritmos y análisis** para gestionar y extraer valor y conocimiento oculto en ellos ...







Source: Contents of above graphic created in partnership with Teradata, Inc.

Conceptos básicos

¿Qué es la Minería de Datos?

La Minería de datos (MD) es el proceso de extracción de patrones de información (implícitos, no triviales, desconocidos y potencialmente útiles) a partir de grandes cantidades de datos

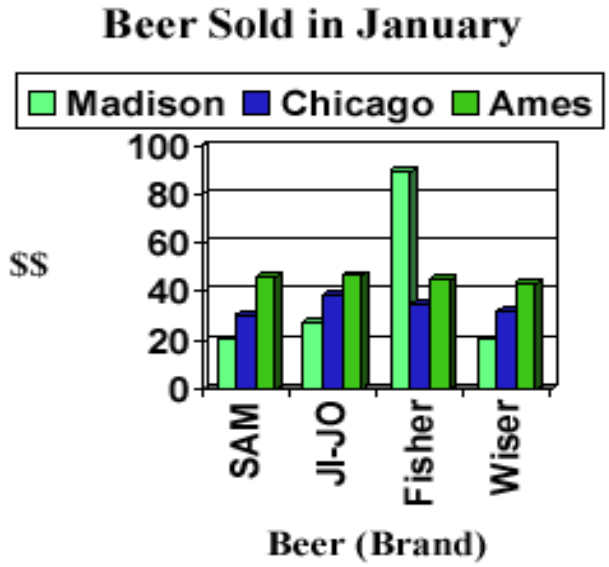
También se conoce como:

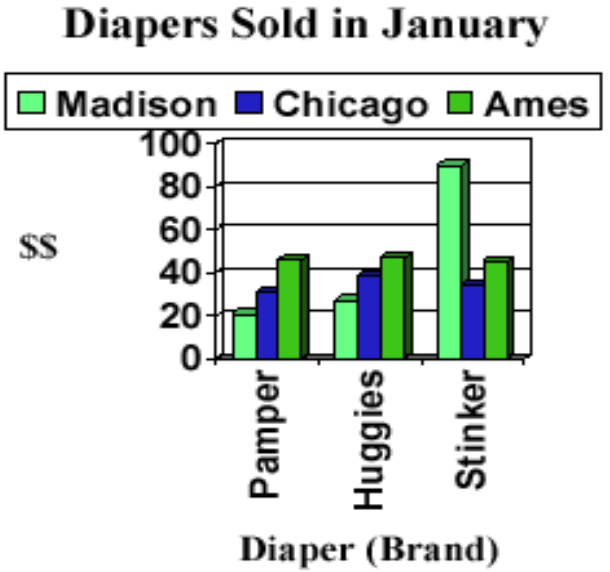
- **Descubrimiento de conocimiento en bases de datos (KDD),**
- **extracción del conocimiento,**
- **análisis inteligente de datos /patrones,**
- **…**

Minería de Datos. Caso de estudio

Marketing y ventas (asociaciones)

- Si se realiza sólo toma de decisión en función de los informes (datos), por ejemplo para dos productos, cerveza y pañales



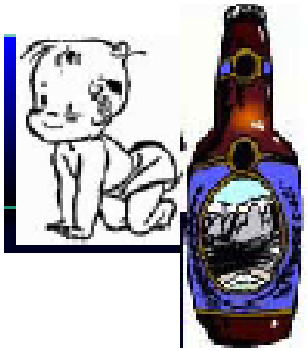


¿Qué información aporta?

Minería de Datos. Caso de estudio

Marketing y ventas (asociaciones)

- Objetivo: determinar grupos de items que tienden a ocurrir juntos en transacciones (=tickets de compra pagados con o sin tarjeta)
- Se utilizan técnicas de asociación, que pueden descubrir información como:
  - Los clientes que compran cerveza también compran patatas **¡Para eso no es necesario el uso de técnicas de DM!**
  - Los viernes por la tarde, con frecuencia, quienes compran pañales, compran también cerveza.



- ✓ ¿Qué significa?
- ✓ ¿A qué se debe?
- ✓ Acciones a realizar

Minería de Datos. Caso de estudio

Marketing y ventas (asociaciones)

Explicación más probable

- Se acerca el fin de semana
- Hay un bebé en casa
- No quedan pañales
- El padre/madre compra pañales al salir del trabajo
- ¡No pueden salir!
- Comprar cervezas para el fin de semana (y un partido/película PPV)

- **Se acerca el fin de semana**
- **Hay un bebé en casa luego nada de ir fuera**
- **Hay que comprar pañales**
- **Quedarse en casa → ver partido/película**
- **Comprar cervezas para el partido/película**

Pañales → Cerveza

Minería de Datos. Caso de estudio

Marketing y ventas (asociaciones)

Acciones a realizar:

- Planificar disposiciones alternativas en el almacén
- Limitar descuentos especiales a sólo uno de los dos productos que tienden a comprarse juntos
- Poner los aperitivos que más margen dejan entre los pañales y las cervezas
- Poner productos de bebé en oferta cerca de las cervezas
- Ofrecer cupones descuento para el producto "complementario", cuando uno de los productos se venda por separado…

La profileración de "tarjetas de lealtad" se debe al interés por identificar el historial de ventas individual del cliente…

Principios básicos de *machine learning*

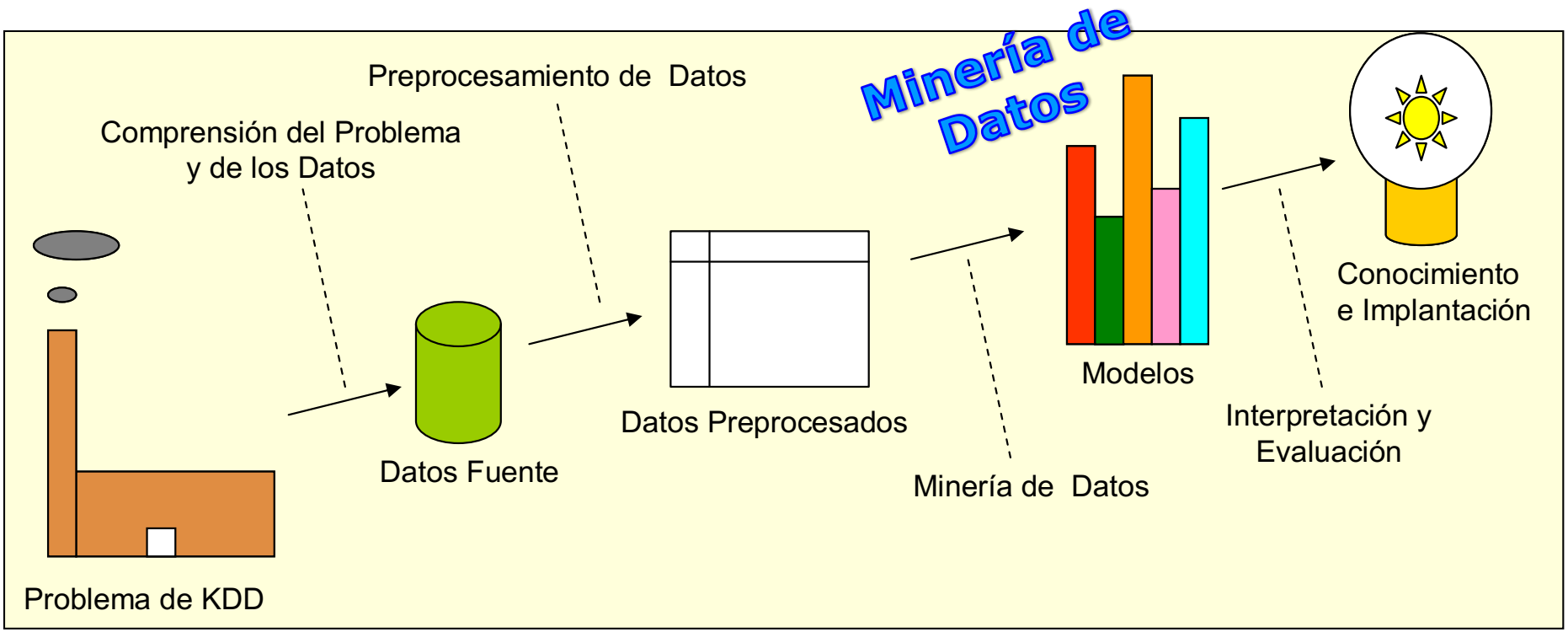
- Conceptos básicos. Ciencia de Datos, Minería de Datos, Big Data, Machine Learning.
- Proceso de Minería de Datos
- Técnicas de Minería de Datos: Clasificación, Regresión, Agrupamiento, Asociación
- Clasificación y Regresión. Predicción por similitud: K-Nearest Neighbour (KNN).
- Validación de Clasificadores
- Clasificación con Árboles de Decisión

Proceso de Minería de Datos

- KDD = *Knowledge Discovery from Databases*
- El KDD es el proceso completo de extracción de conocimiento a partir de bases de datos
- El término se acuñó en 1989 para enfatizar que el conocimiento es el producto final de un proceso de descubrimiento guiado por los datos
- La Minería de Datos es sólo una etapa en el proceso de KDD
- Informalmente se asocia Minería de Datos con KDD

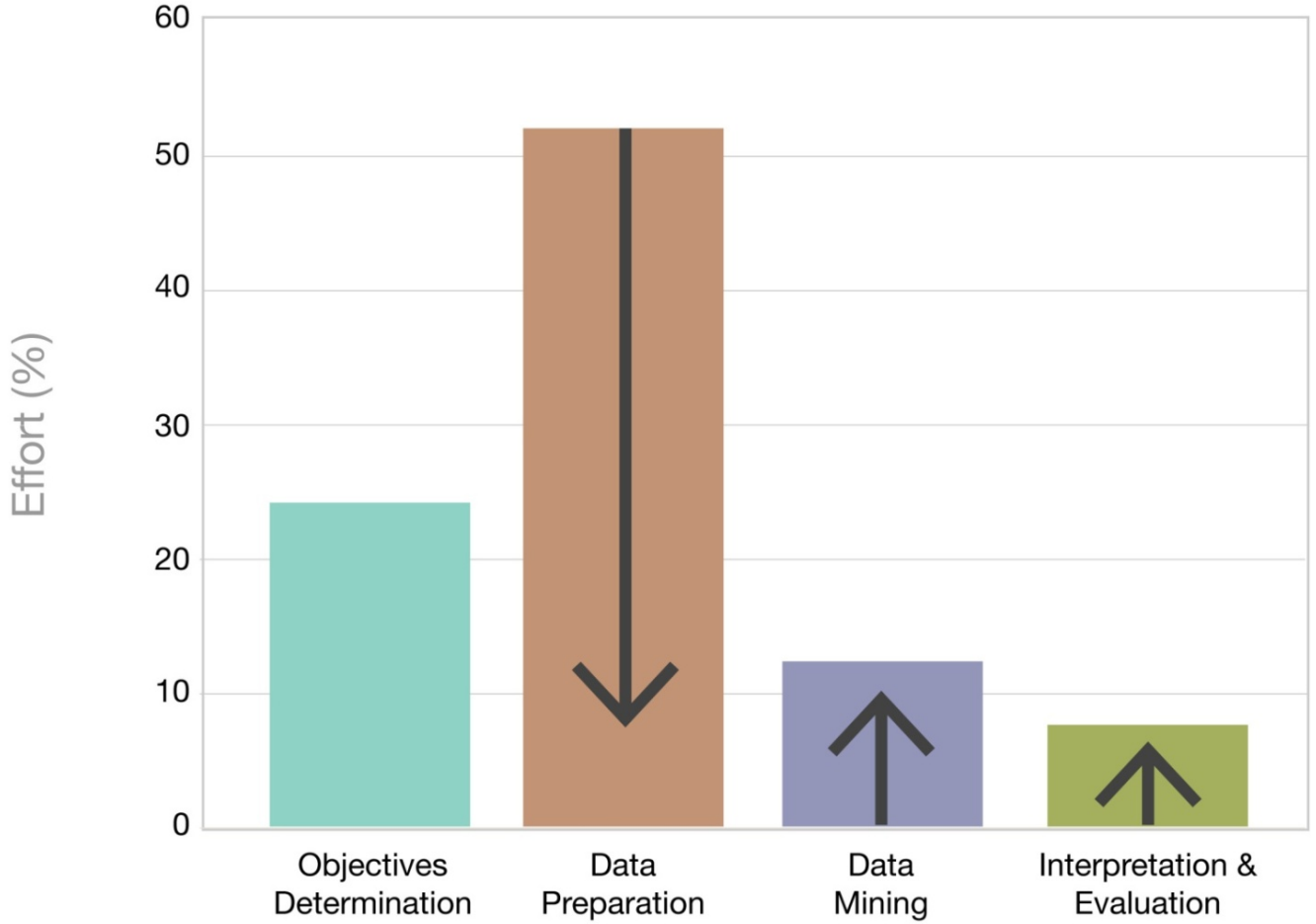
Proceso de Minería de Datos

Etapas en un proceso de KDD



Informalmente se asocia Minería de Datos con KDD

Etapas en el proceso de KDD

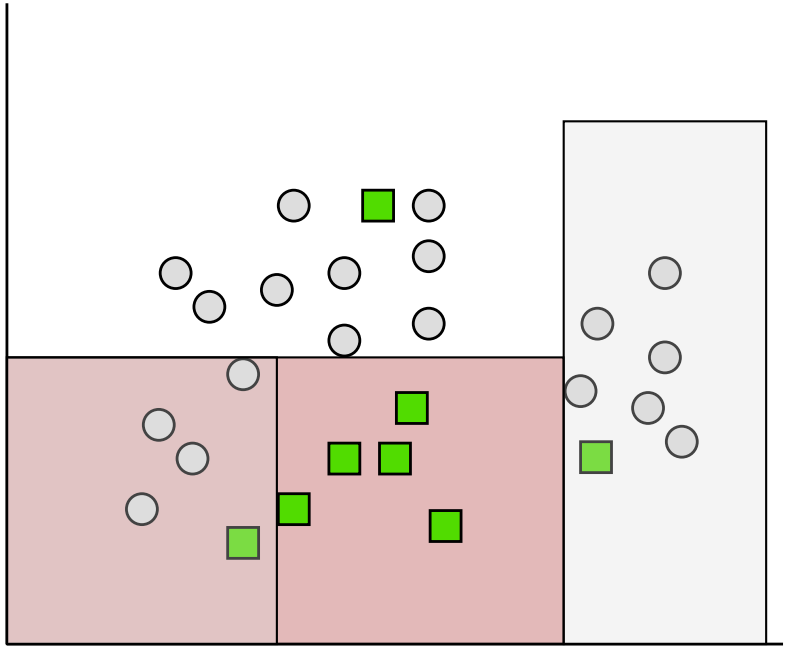


Tiempos estimados en el análisis de un problema mediante técnicas de minería de datos

Principios básicos de *machine learning*

- Conceptos básicos. Ciencia de Datos, Minería de Datos, Big Data, Machine Learning
- Proceso de Minería de Datos
- Técnicas de Minería de Datos: Clasificación, Regresión, Agrupamiento, Asociación
- Clasificación y Regresión. Predicción por similitud: K-Nearest Neighbour (KNN).
- Regresión. Medidas de evaluación, similitud y regresión lineal
- Validación de Clasificadores
- Clasificación con Árboles de Decisión

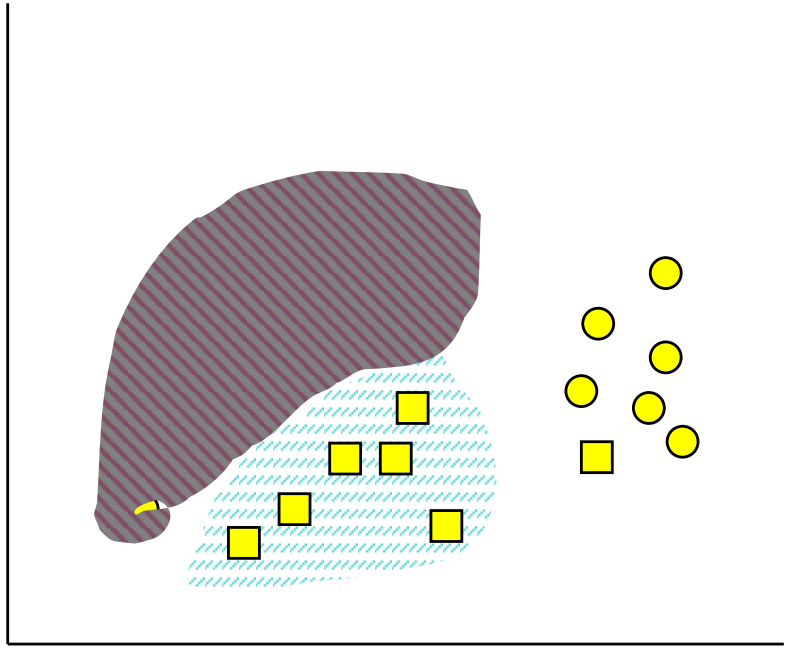
Aprendizaje Supervisado vs No Supervisado



Aprendizaje supervisado:

Aprende, a partir de un conjunto de instancias pre-etiquetadas un metodo para predecir (Ejemplo, clasificación: la clase a que pertenece una nueva instancia)

Aprendizaje Supervisado vs No Supervisado



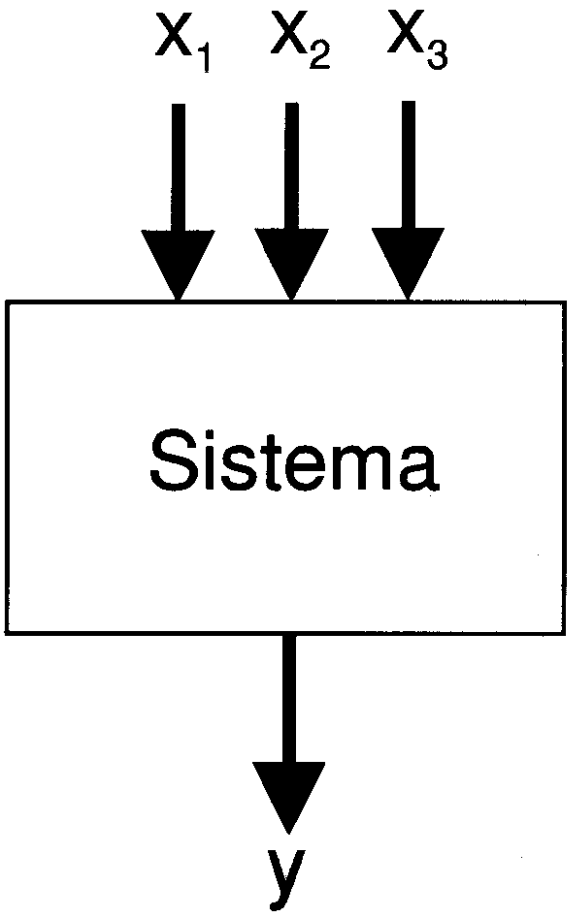
Aprendizaje no supervisado:

No hay conocimiento a priori sobre el problema, no hay instancias etiquetadas, no hay supervisión sobre el procedimiento.

(Ejemplo, clustering: Encuentra un agrupamiento de instancias "natural" dado un conjunto de instancias no etiquetadas)

Regresión

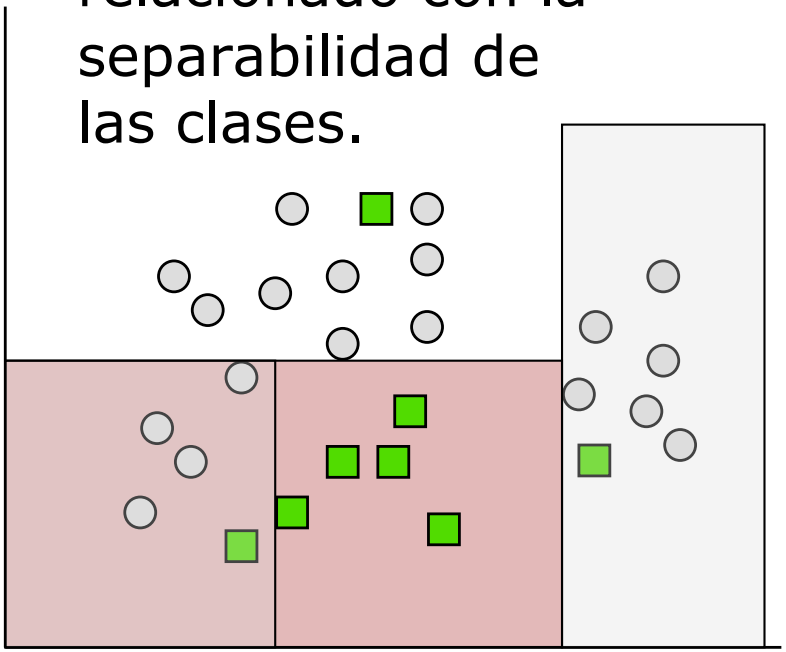
El problema fundamental de la predicción está en modelar la relación entre las variables de estado para obtener el valor de la variable de control.

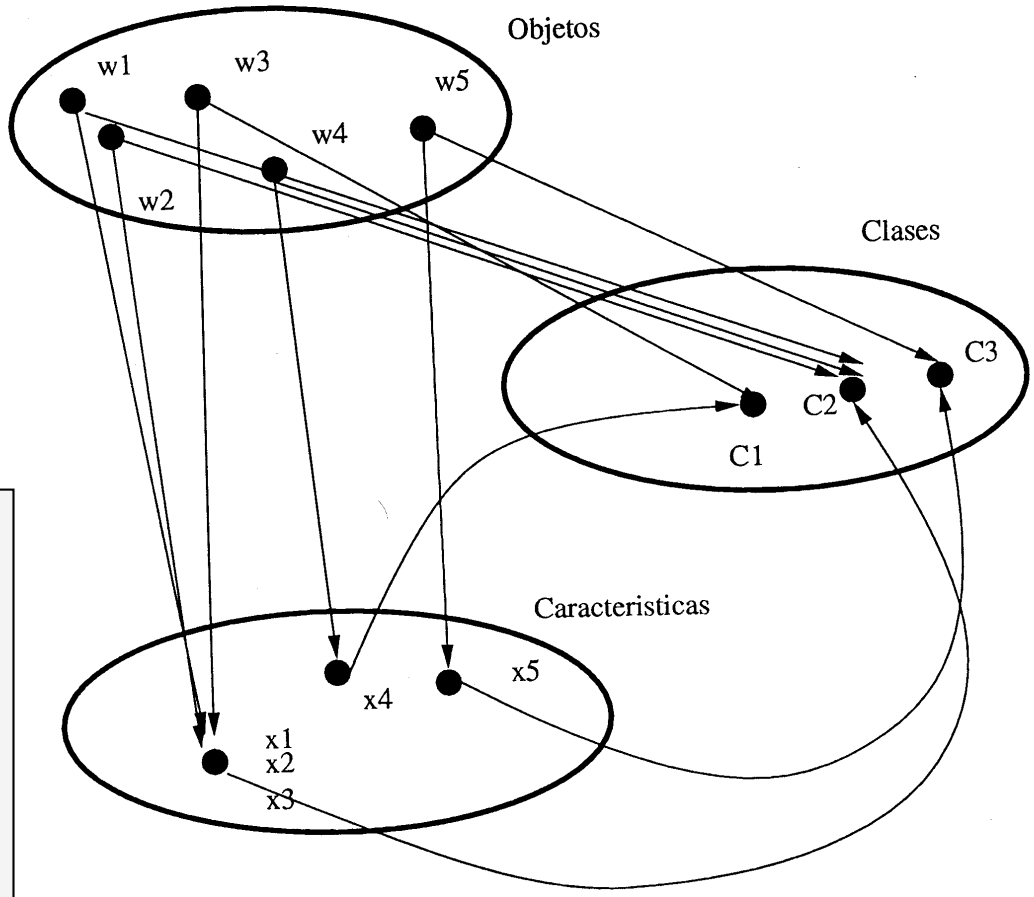


Clasificación

- **Clasificación**

  El problema fundamental de la clasificación está directamente relacionado con la separabilidad de las clases.





Clasificación. Ejemplo

- **Ejemplo:** Diseño de un Clasificador para *Iris*
  - Problema simple muy conocido: *clasificación de lirios.*
  - Tres clases de lirios: *setosa, versicolor* y *virginica*.
  - Cuatro atributos: *longitud* y *anchura* de *pétalo* y *sépalo*, respectivamente.
  - 150 ejemplos, 50 de cada clase.
  - Disponible en http://www.ics.uci.edu/~mlearn/MLRepository.html

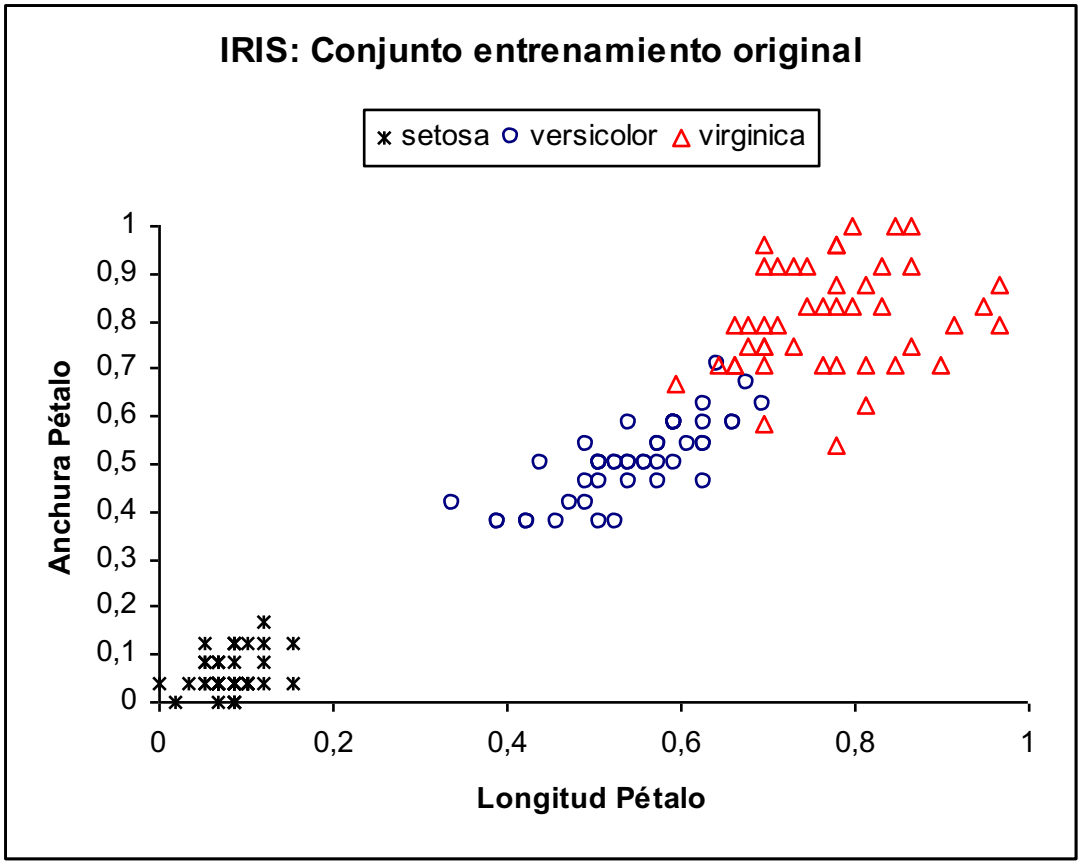
setosa

versicolor

virginica

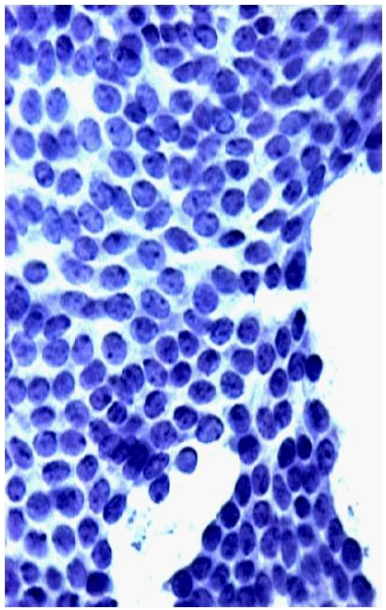
Clasificación. Ejemplo

Ejemplos de conjuntos seleccionados sobre *Iris:*



Clasificación. Ejemplo

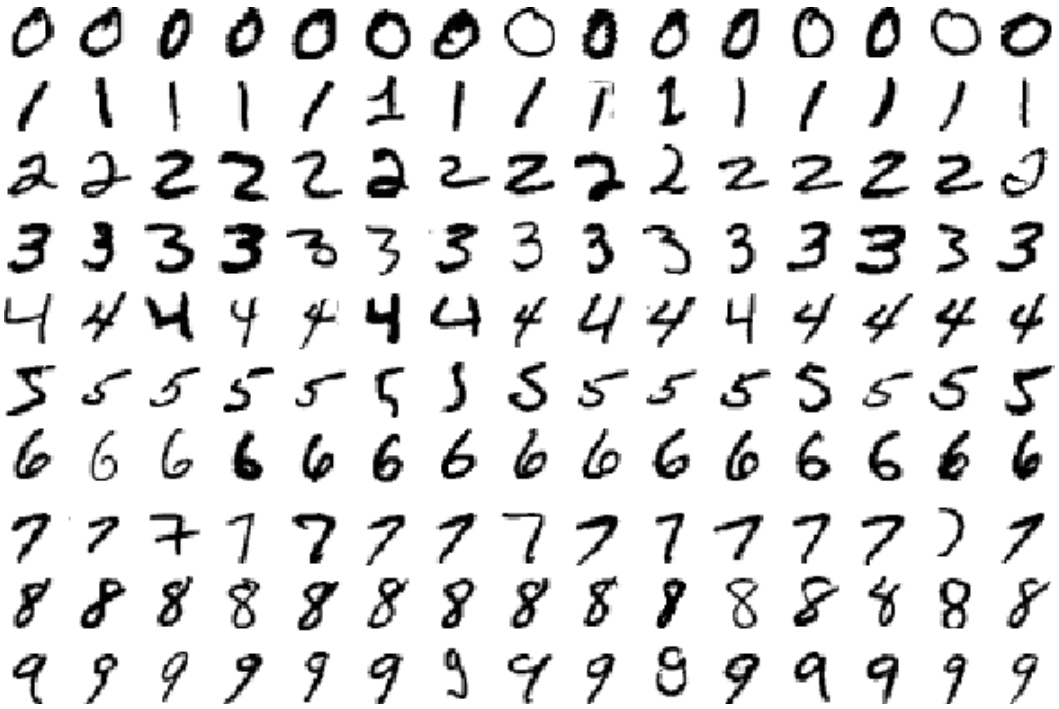
Wisconsin Breast Cancer: Predict malignant/benign



Attribute name	Description
RADIUS	*Mean of distances from center to points on the perimeter*
TEXTURE	*Standard deviation of grayscale values*
PERIMETER	*Perimeter of the mass*
AREA	*Area of the mass*
SMOOTHNESS	*Local variation in radius lengths*
COMPACTNESS	*Computed as: perimeter^2/area – 1.0*
CONCAVITY	*Severity of concave portions of the contour*
CONCAVE POINTS	*Number of concave portions of the contour*
SYMMETRY	*A measure of the nucleii's symmetry*
FRACTAL DIMENSION	*'Coastline approximation'* – 1.0
DIAGNOSIS (Target)	*Diagnosis of cell sample: malignant or benign*

Clasificación. Ejemplo

Handwritting recognition.
Assign a digit from 0 to 9.



Clasificación

Se pueden construir distintos tipos de clasificadores:

Modelos Interpretables:

- Árboles de decisión
- Reglas (p.ej. listas de decisión)

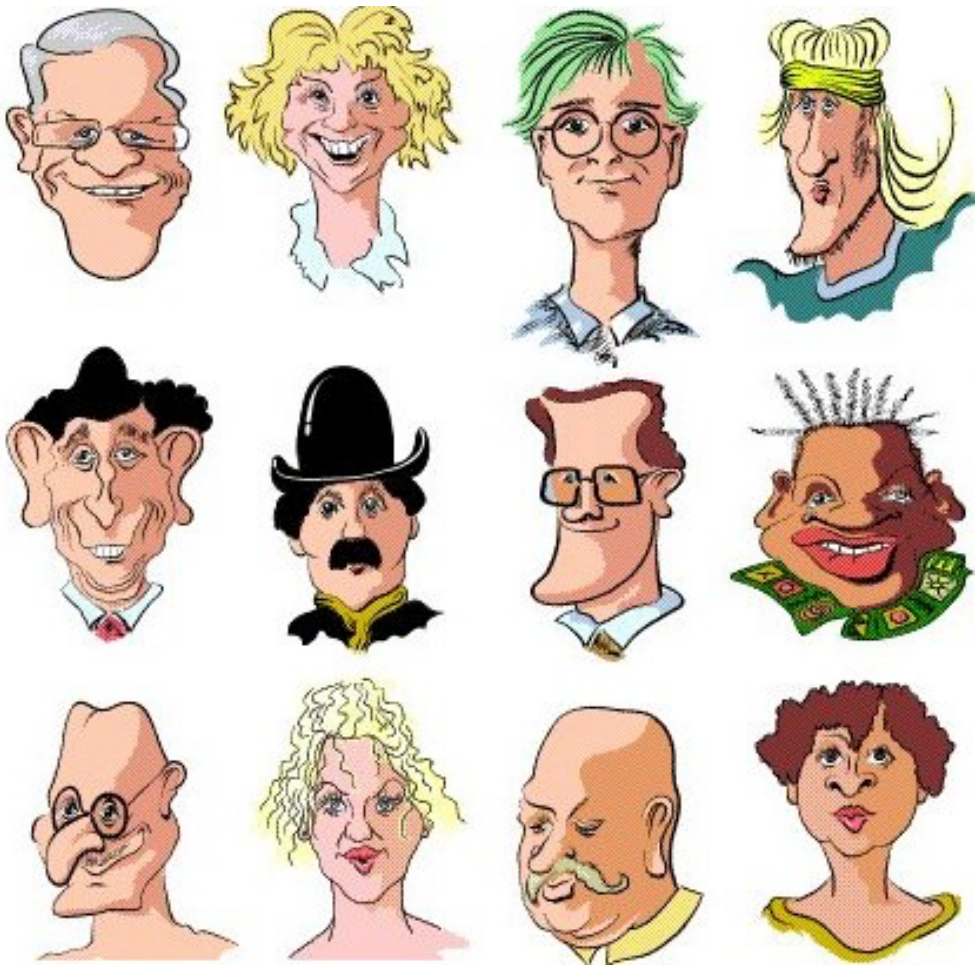
Modelos no interpretables:

- Clasificadores basados en casos (k-NN)
- Redes neuronales
- Redes bayesianas
- SVMs (Support Vector Machines)
- …

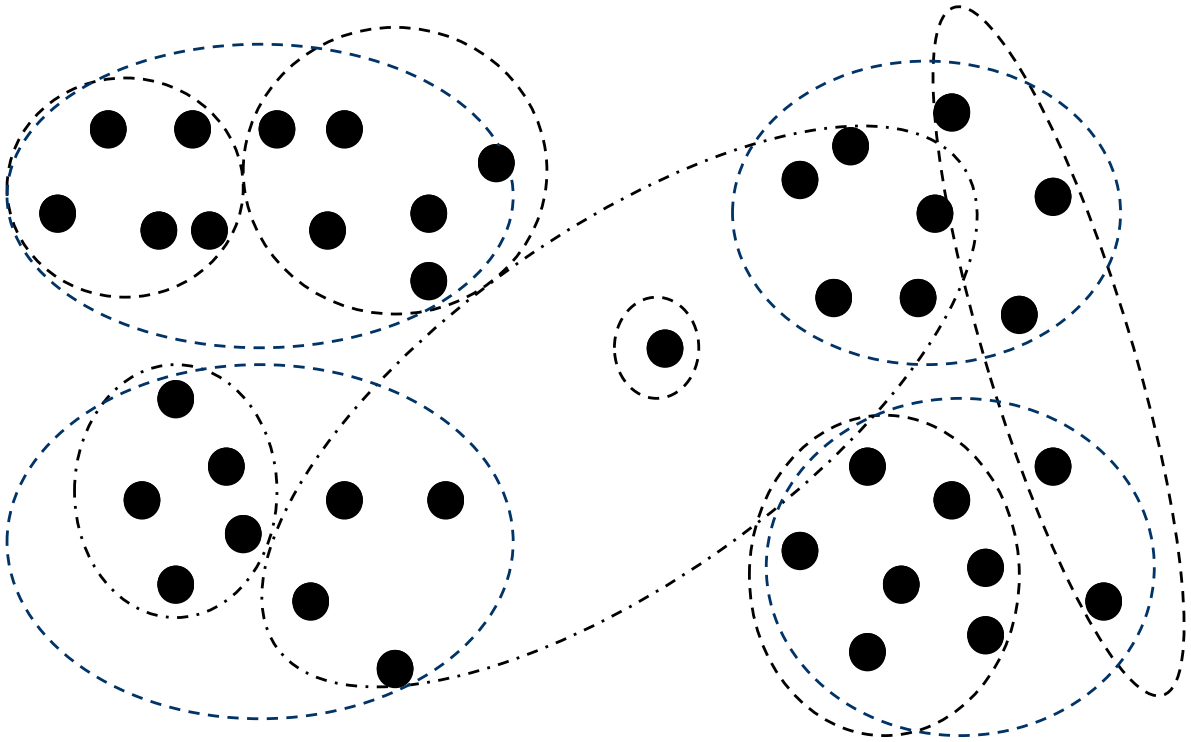
Agrupamiento

Hay problemas en los que deseamos agrupar las instancias creando clusters de similares características

Ej. Segmentación de clientes de una empresa

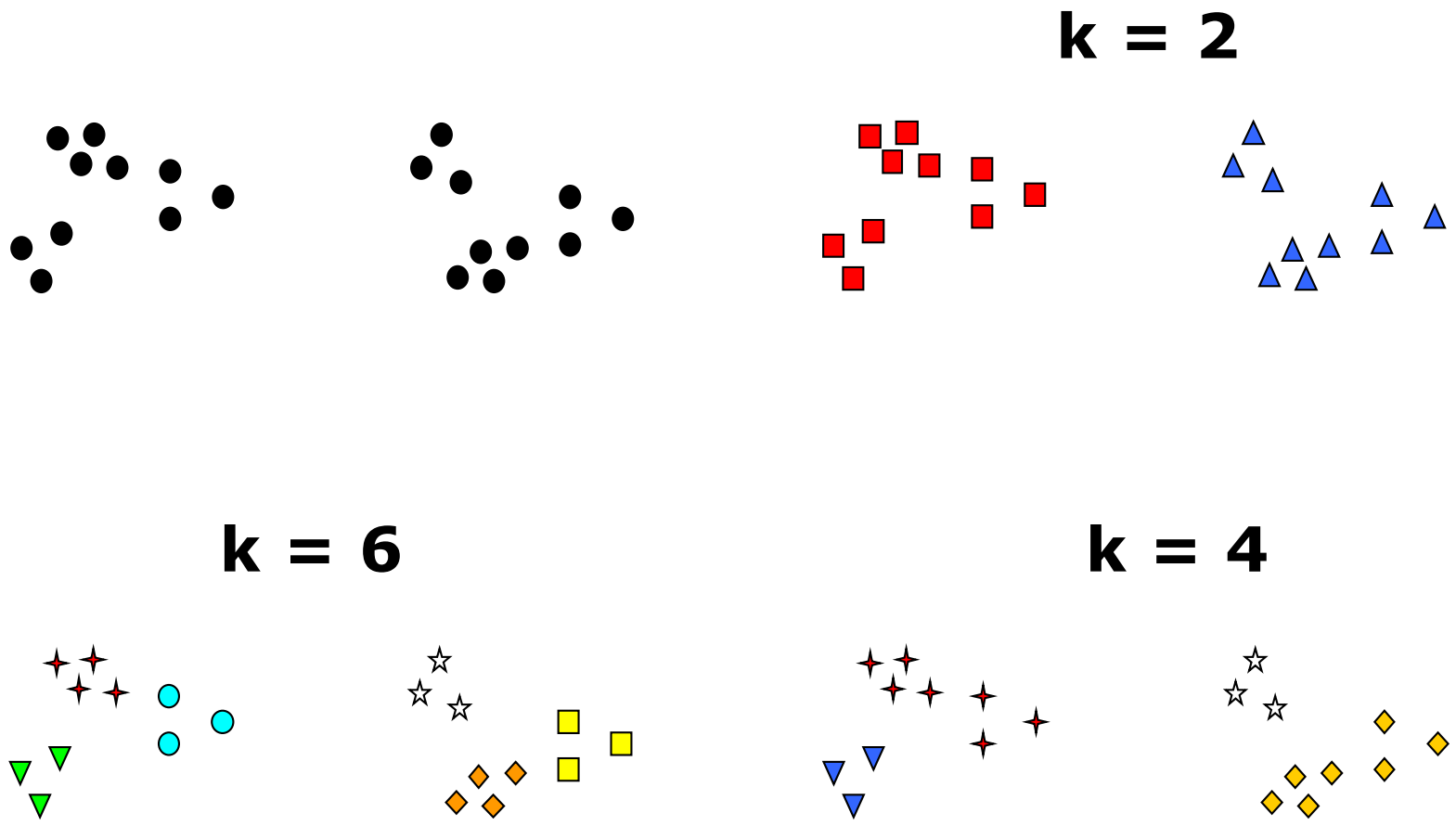


Agrupamiento. Niveles



La decisión del número de clusters es uno de los retos en agrupamiento

Agrupamiento. Niveles

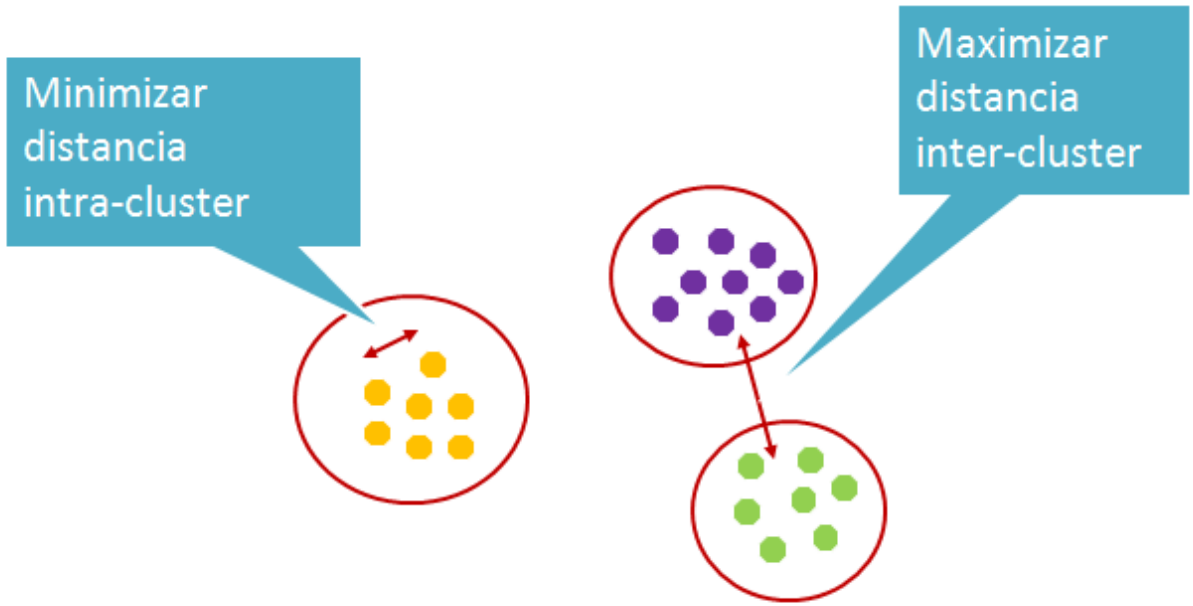


La decisión del número de clusters es uno de los retos en agrupamiento

Agrupamiento. Modelos

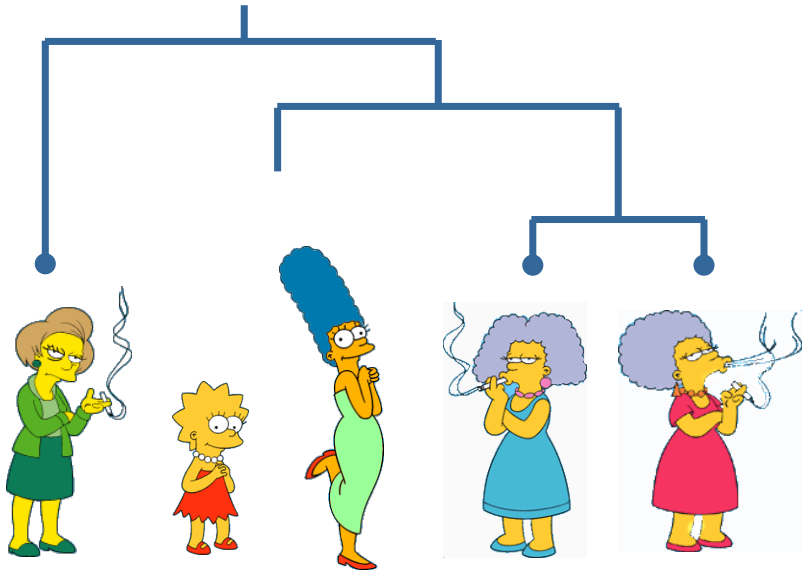
Objetivo

Encontrar agrupamientos de tal forma que los objetos de un grupo sean similares entre sí y diferentes de los objetos de otros grupos [*clusters*].

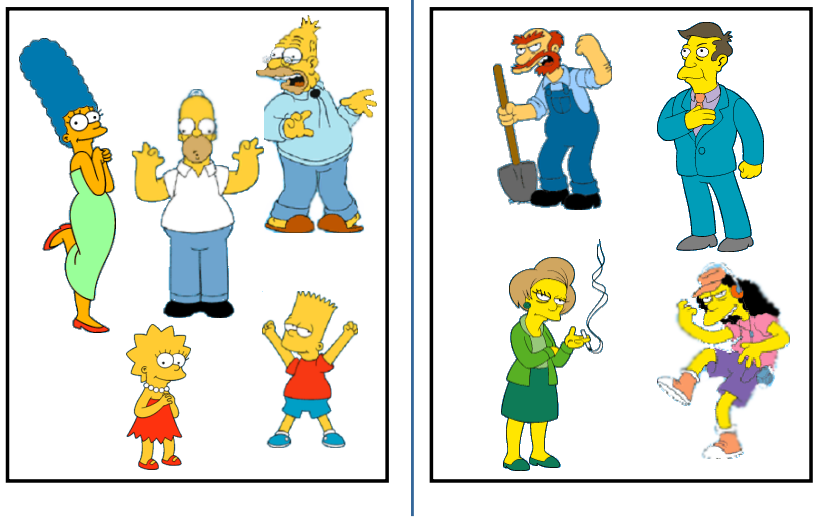


Agrupamiento. Modelos

Modelos Jerárquicos



Modelos Particionales



Ejemplos de Agrupamiento

- **Marketing:** descubrimiento de distintos grupos de clientes en la BD. Usar este conocimiento en la política publicitaria, ofertas, …
- **Uso de la tierra:** Identificación de áreas de uso similar a partir de BD con observaciones de la tierra (cultivos, …)
- **Seguros:** Identificar grupos de asegurados con características parecidas (siniestros, posesiones, ….). Ofertarles productos que otros clientes de ese grupo ya poseen y ellos no
- **Planificación urbana:** Identificar grupos de viviendas de acuerdo a su tipo, valor o situación geográfica
- **WWW:** Clasificación de documentos, analizar ficheros .log para descubrir patrones de acceso similares, …

Descubrimiento de Asociaciones

- Descubrimiento de reglas de asociación:
  - Búsqueda de patrones frecuentes, asociaciones, correlaciones, o estructuras causales entre conjuntos de artículos u objetos (datos) a partir de bases de datos transaccionales, relacionales y otros conjuntos de datos
  - Búsqueda de *secuencias* o patrones temporales
  - Aplicaciones:
    - análisis de cestas de la compra (*Market Basket analysis*)
    - diseño de catálogos,…
    - ¿Qué hay en la cesta? Libros de Jazz
    - ¿Qué *podría* haber en la cesta? El último CD de Jazz
    - ¿Cómo *motivar al cliente* a comprar los artículos que es probable que le gusten?

Descubrimiento de asociaciones
Market Basket Analysis

Compra: zumo de naranja, plátanos, detergente para vajillas, limpia cristales, gaseosa, ...

¿Es típico comprar gaseosa y plátanos? ¿Es importante la marca de la gaseosa?

¿Cómo afecta la demografía de la vecindad a la compra de los clientes?

¿Aumenta la compra del limpia cristales cuando se compran a la vez detergente para vajillas y zumo de naranja?

¿Dónde deberían colocarse los detergentes para maximizar sus ventas?

Técnicas de Minería de Datos

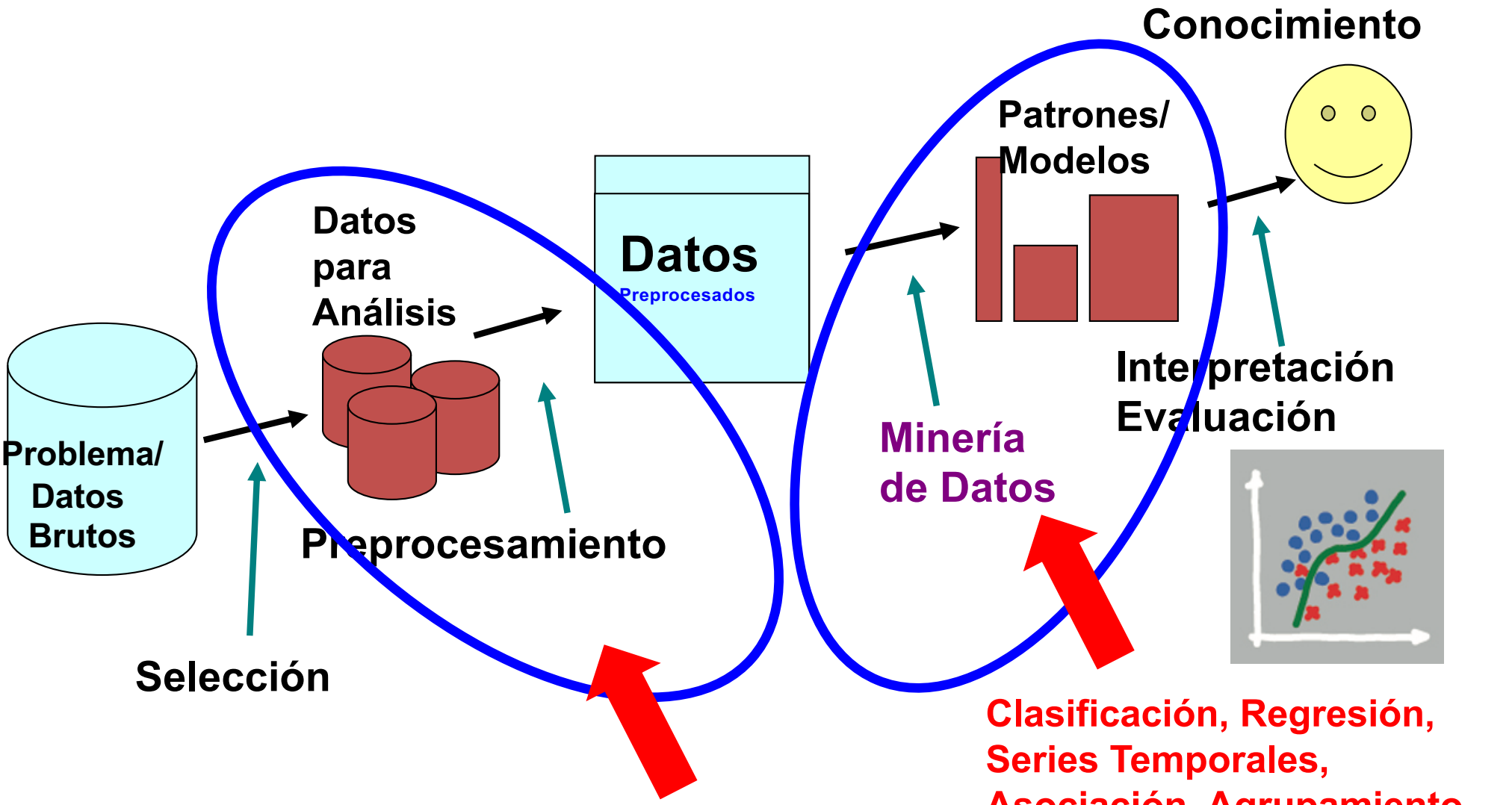
- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation/Anomaly Detection [Predictive]
- Time Series [Predictive]
- Summarization [Descriptive]

Principios básicos de *machine learning*

- Conceptos básicos. Ciencia de Datos, Minería de Datos, Big Data, Machine Learning
- Proceso de Minería de Datos
- Técnicas de Minería de Datos: Clasificación, Regresión, Agrupamiento, Asociación
- Clasificación y Regresión. Predicción por similitud: K-Nearest Neighbour (KNN)
- Validación de Clasificadores
- Clasificación con Árboles de Decisión

Motivación

Predicción es uno de los problemas más estudiados en minería de datos y tiene una gran presencia en problemas reales: problemas de clasificación y regresión



Definición del problema de clasificación

- La clasificación es la técnica de DM más conocida
- El *problema de clasificación* consiste en predecir una determinada clase (categórica) para un objeto
- Otros problemas como estimación o predicción pueden verse como problemas de clasificación
- La *tarea de clasificación:* Dados un conjunto de ejemplos ya clasificados, construir un modelo o clasificador que permita clasificar nuevos casos
- Es un tipo de *aprendizaje supervisado*: Se conoce la clase verdadera de cada uno de los ejemplos que se utilizan para construir el clasificador
- Un *clasificador* puede ser un conjunto de reglas, un árbol de decisión, una red neuronal, etc.
- Aplicaciones típicas:
  - Aprobación de créditos, marketing directo, detección de fraudes, diagnóstico médico…

Definición del problema de clasificación. Ejemplo

Ejemplo: El problema de clasificación de la flor de *Iris*

- Problema simple muy conocido: *clasificación de lirios.*
- Tres clases de lirios: *setosa, versicolor* y *virginica*.
- Cuatro atributos: *longitud* y *anchura* de *pétalo* y *sépalo*, respectivamente.
- 150 ejemplos, 50 de cada clase.
- Disponible en
  http://www.ics.uci.edu/~mlearn/MLRepository.html

	Sepal length X_1	Sepal width X_2	Petal length X_3	Petal width X_4	Class X_5
$\mathbf{x}_1$	5.9	3.0	4.2	1.5	Iris-versicolor
$\mathbf{x}_2$	6.9	3.1	4.9	1.5	Iris-versicolor
$\mathbf{x}_3$	6.6	2.9	4.6	1.3	Iris-versicolor
$\mathbf{x}_4$	4.6	3.2	1.4	0.2	Iris-setosa
$\mathbf{x}_5$	6.0	2.2	4.0	1.0	Iris-versicolor
$\mathbf{x}_6$	4.7	3.2	1.3	0.2	Iris-setosa
$\mathbf{x}_7$	6.5	3.0	5.8	2.2	Iris-virginica
$\mathbf{x}_8$	5.8	2.7	5.1	1.9	Iris-virginica
⋮	⋮	⋮	⋮	⋮	⋮
$\mathbf{x}_{149}$	7.7	3.8	6.7	2.2	Iris-virginica
$\mathbf{x}_{150}$	5.1	3.4	1.5	0.2	Iris-setosa

setosa

versicolor

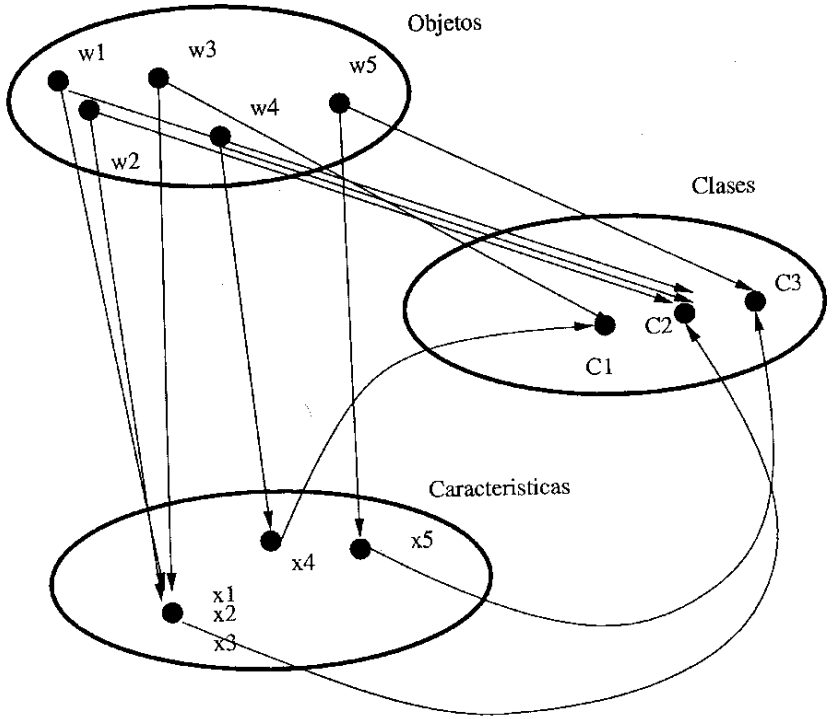
virginica

Definición del problema de clasificación

- Es un tipo de *aprendizaje supervisado*: Se conoce la clase verdadera de cada uno de los ejemplos que se utilizan para construir el clasificador
- El *problema de clasificación* consiste en predecir una determinada clase (categórica) para un objeto

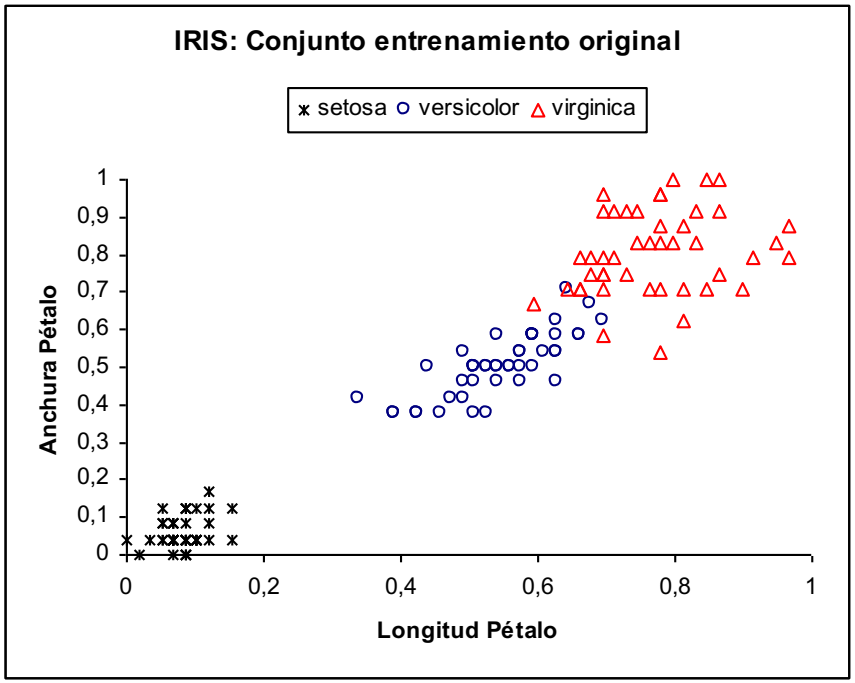
- La *tarea de clasificación:* Dados un conjunto de ejemplos ya clasificados, construir un modelo o clasificador que permita clasificar nuevos casos

El problema fundamental de la clasificación está directamente relacionado con la separabilidad de las clases.



Definición del problema de clasificación. Ejemplo

Ejemplos de conjuntos seleccionados sobre *Iris:*

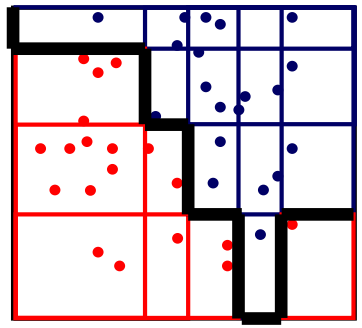


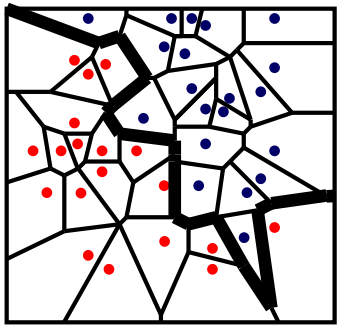
El problema fundamental de la clasificación está directamente relacionado con la separabilidad de las clases.

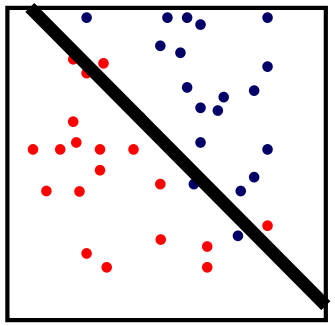
En este problema la dificultad está en la separación entre las clases versicolor y virgínica

Definición del problema de clasificación

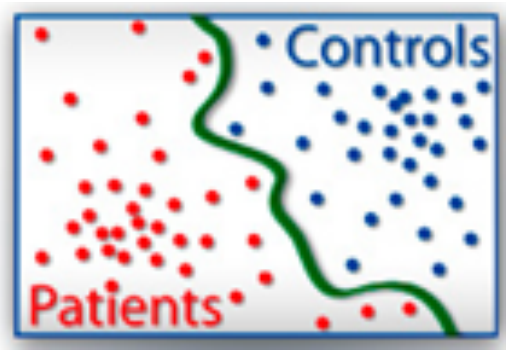
- Un *clasificador* puede ser un conjunto de reglas, un árbol de decisión, una red neuronal, etc.







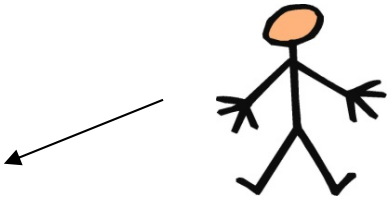
- Aplicaciones típicas:
  - Aprobación de créditos, marketing directo, detección de fraudes, diagnóstico médico…



Definición del problema de clasificación

- Un objeto se describe a través de un conjunto de características (variables o atributos)

$$X \rightarrow \{X_1, X_2, \ldots, X_n\} \qquad e_i \rightarrow \{x_1, x_2, \ldots, x_n, c_k\}$$



- Edad
- Astigmatismo
- Ratio de lagrimeo
- Miopía

CLASIFICACIÓN: Tipo de lentillas

-Ingresos
-Deudas
-Propiedades
…
-CLASIFICACIÓN: Conceder el crédito

- Objetivo de la tarea de clasificación: Clasificar el objetivo dentro de una de las categorías de la clase C={c_1,…,c_k}

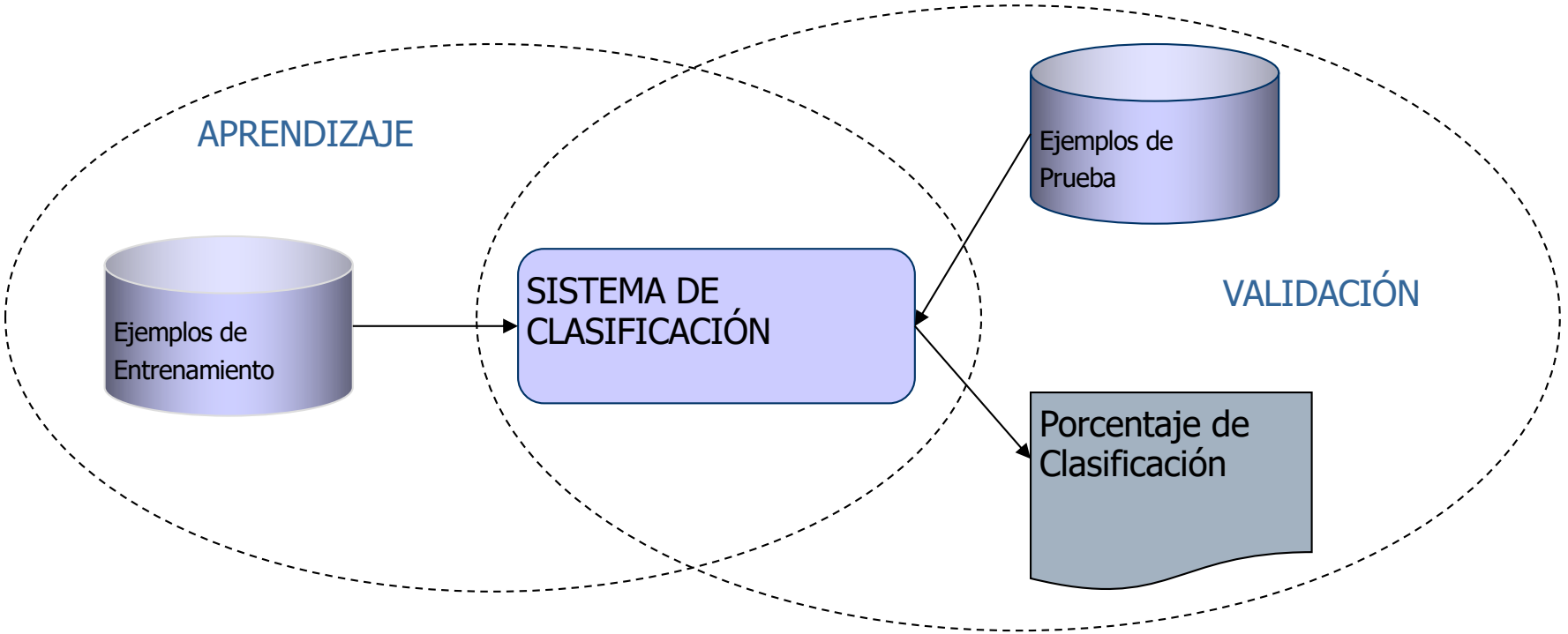
$$f: X_1 \times X_2 \times \ldots \times X_n \rightarrow C$$

- Las características o variables elegidas dependen del problema de clasificación

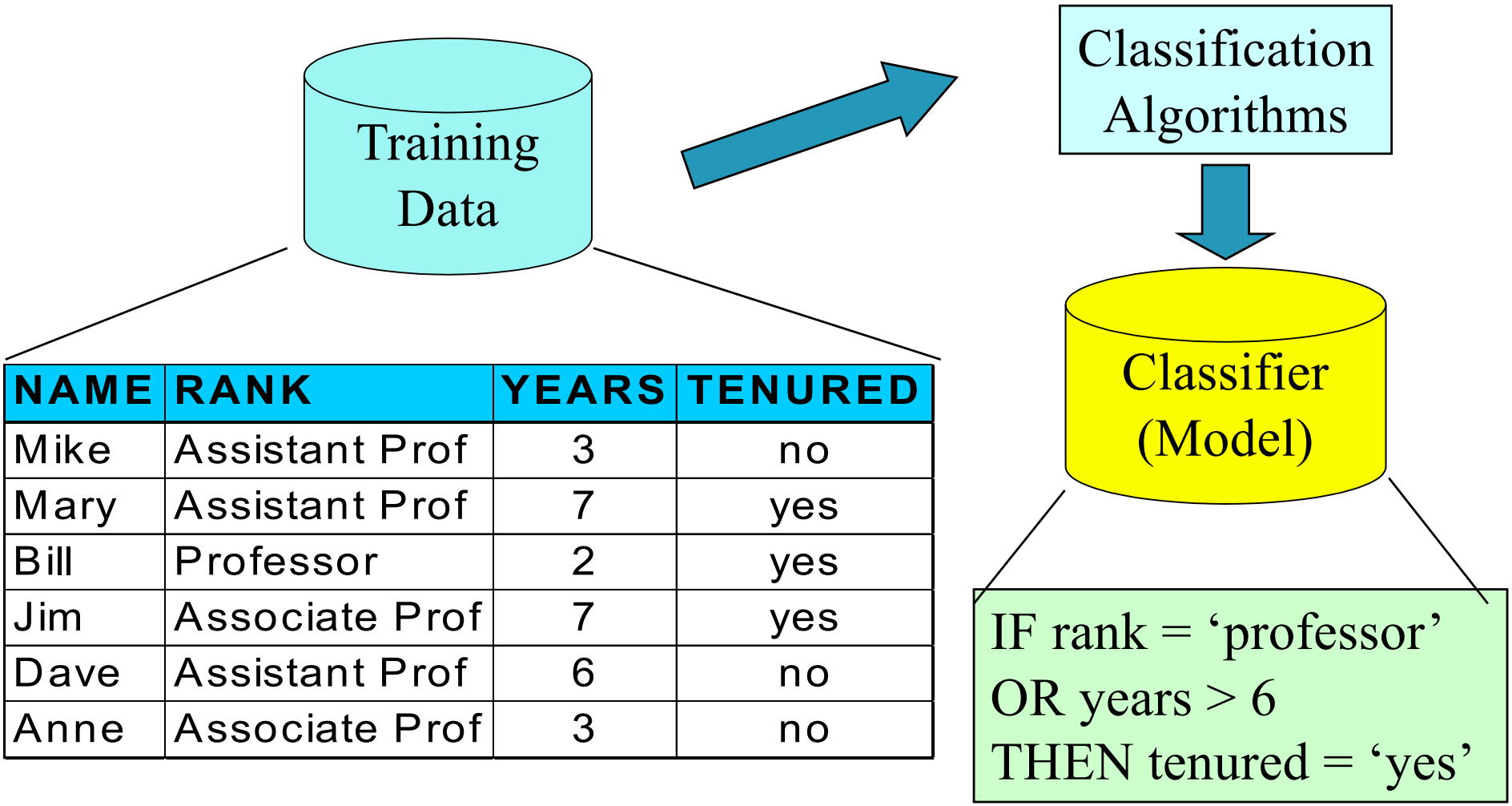
Etapas en el proceso de clasificación

1. Construcción del modelo que describe el conjunto de clases (predeterminadas): Fase de Aprendizaje
   - Cada ejemplo (tupla) se sabe que pertenece a una clase (etiqueta del atributo clase)
   - Se utiliza un conjunto de ejemplos para la construcción del modelo: conjunto de entrenamiento (*training set*)
   - El modelo obtenido se representa como un conjunto de reglas de clasificación, árboles de decisión, fórmula matemática, …
2. Utilización del modelo: Validación
   - Estimación de la precisión del modelo
     - Se utiliza un conjunto de ejemplos distintos de los utilizados para la construcción del modelo: conjunto de prueba (*test set*)
       - Si el conjunto de test no fuese independiente del de entrenamiento ocurría un proceso de sobreajuste (*overfitting*)
     - Para cada ejemplo de test se compara la clase determinada por el modelo con la clase real (conocida)
     - El ratio de precisión es el porcentaje de ejemplos de test que el modelo clasifica correctamente

Etapas en el proceso de clasificación

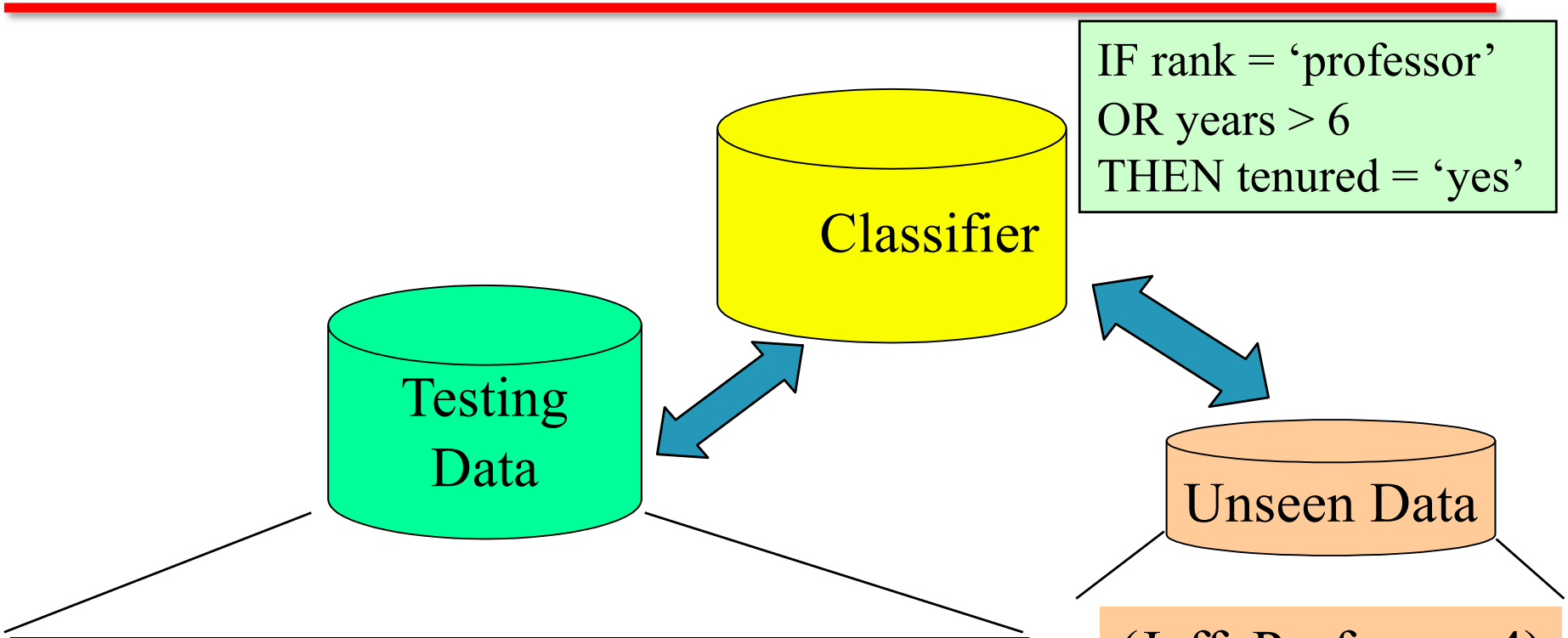


Etapas en el proceso de clasificación



NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

Etapas en el proceso de clasificación



NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

Criterios para evaluar un clasificador

- Precisión o exactitud (s, ε)
  - En ocasiones se debe considerar el costo de la clasificación incorrecta
- Velocidad
  - Tiempo necesario para la construcción del modelo
  - Tiempo necesario para usar el modelo
- Robustez: capacidad para tratar con valores desconocidos
- Escalabilidad: Aumento del tiempo necesario (en construcción y evaluación) con el tamaño de la BD
- Interpretabilidad: comprensibilidad del modelo obtenido
- Complejidad del modelo: Tamaño del árbol de clasificación, número de reglas, antecedentes en las reglas,…

Criterios para evaluar un clasificador

- Matriz de confusión

Dado un problema de clasificación con *m* clases, una matriz de confusión es una matriz *m* x *m* en la que una entrada $c_{i,j}$ indica el número de ejemplos que se han asignado a la clase $c_{j,}$ cuando la clase correcta es c_i

Ejemplo: Para la BD Height, si suponemos que output1 es la clasificación correcta y output2 es la que hace el clasificador, la matriz de confusión es

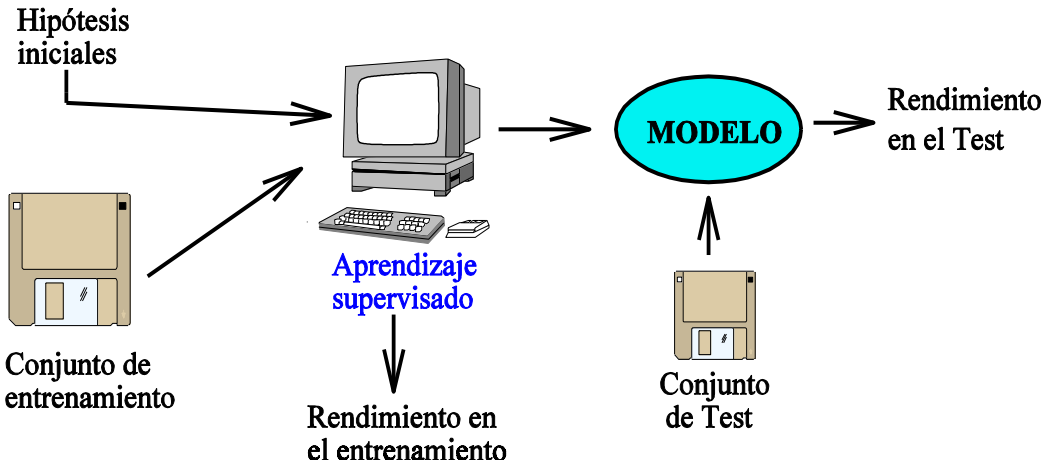
Criterios para evaluar un clasificador

Name	Gender	Height	Output1	Output2
Kristina	F	1.6m	Short	Medium
Jim	M	2m	Tall	Medium
Maggie	F	1.9m	Medium	Tall
Martha	F	1.88m	Medium	Tall
Stephanie	F	1.7m	Short	Medium
Bob	M	1.85m	Medium	Medium
Kathy	F	1.6m	Short	Medium
Dave	M	1.7m	Short	Medium
Worth	M	2.2m	Tall	Tall
Steven	M	2.1m	Tall	Tall
Debbie	F	1.8m	Medium	Medium
Todd	M	1.95m	Medium	Medium
Kim	F	1.9m	Medium	Tall
Amy	F	1.8m	Medium	Medium
Wynette	F	1.75m	Medium	Medium

	Asignación		
	Short	Medium	Tall
Short	0	4	0
Medium	0	5	3
Tall	0	1	2

Matriz de confusión

Criterios para evaluar un clasificador



- Rendimiento (matriz de confusión):

	Clasificación como	
	Si	No
Clase SI	Verdadero positivo (VP)	Falso negativo (FN)
real NO	Falso Positivo (FP)	Verdadero Negativo (VN)

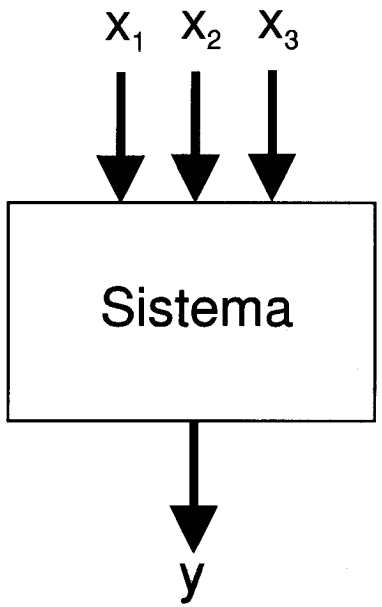
$N=VP+VN+FP+FN$

- Tasa de acierto $s = \frac{VP+VN}{N}$
- Tasa de error $\varepsilon = 1 - s$

Definición del problema de Regresión

- Objetivo: predecir el valor numérico para una variable a partir de los valores para otras
- La definición del problema es parecida a la del problema de clasificación: tenemos variables predictoras y una variable de regresión que en este caso es numérica
- En regresión la mayoría (e incluso todas) las variables predictoras son numéricas

El problema fundamental de la predicción está en modelar la relación entre las variables de estado para obtener el valor de la variable a predecir.



Definición del problema de Regresión

- Objetivo: predecir el valor numérico para una variable a partir de los valores para otras
- La definición del problema es parecida a la del problema de clasificación: tenemos variables predictoras y una variable de regresión que en este caso es numérica
- En regresión la mayoría (e incluso todas) las variables predictoras son numéricas
- **Ejemplos:**
  - **¿qué consumo tendrá un coche en autovía en función de su peso, cilindrada, potencia,...?**
  - **¿qué número de artículos tendremos para el próximo pedido?**
  - **¿cuántos meses necesitaremos para desarrollar un proyecto software?**
  - **¿cuál es la probabilidad de que un cliente determinado sea receptivo a un envío publicitario?**
  - **¿cuántos enfermos tendremos en urgencias la próxima nochebuena?**

Validación en algoritmos de regresión

- Todas las técnicas de validación estudiadas en clasificación son válidas para predicción numérica
- La diferencia está en que ahora debemos medir el error de otra forma
- Debemos medir el error cometido al aproximar un conjunto de valores $\{v_1,\ldots,v_n\}$ por su estimación $\{v'_1,\ldots,v'_n\}$

Error cuadrático medio (ECM)

$$ECM = \frac{\sum_{i=1}^{n}(v_i - v_i')^2}{n}$$

ECM estandarizado (ECME)

$$ECME = \sqrt{\frac{\sum_{i=1}^{n}(v_i - v_i')^2}{n}}$$

Error medio absoluto (EMA)

$$EMA = \frac{\sum_{i=1}^{n}|v_i - v_i'|}{n}$$

Error absoluto relativo (EAR)

$$EAR = \frac{\sum_{i=1}^{n}|v_i - v_i'|}{\sum_{i=1}^{n}|v_i - \bar{v}|}$$

Coeficiente de correlación $r_{vv'} = \frac{\sum_{i=1}^{n}(v_i - \bar{v})(v_i' - \bar{v}')}{(n-1)\sigma_v\sigma_{v'}}$

Análisis de regresión

- El análisis de regresión es el método más utilizado para realizar la tarea de predicción numérica
- **Objetivo:** estimar la variable objetivo *(y)* como una ecuación que contiene como incógnitas al resto de las variables $(x_1,\ldots,x_n)$
- El modelo más sencillo es la **regresión lineal** que reducida a una sola variable predictora tiene la forma:

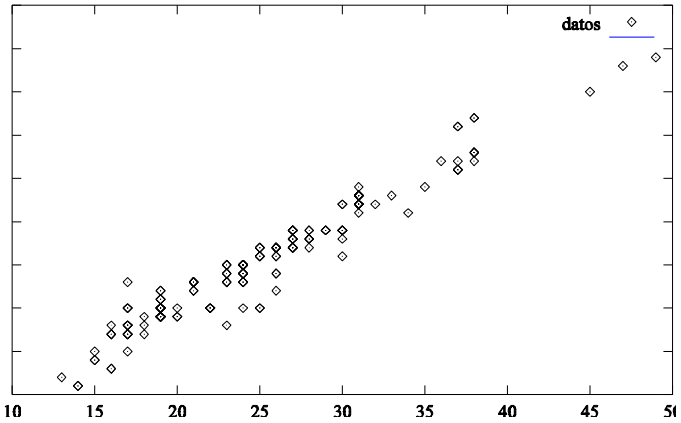
$$y = a + b \cdot x$$

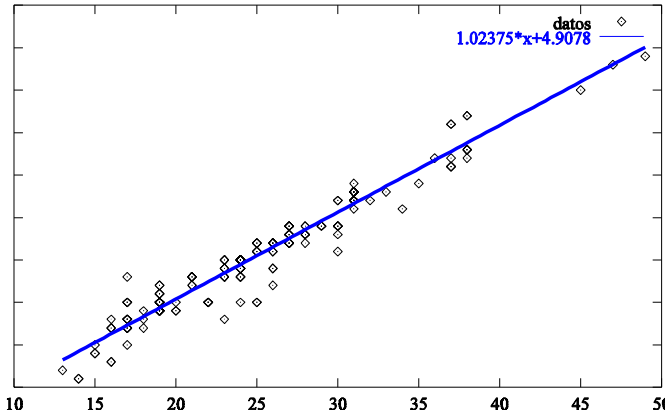
- Estos coeficientes pueden obtenerse fácilmente mediante el método de los mínimos cuadrados

$$b = \frac{\sum_{i=1}^{s}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{s}(x_i - \bar{x})^2}$$

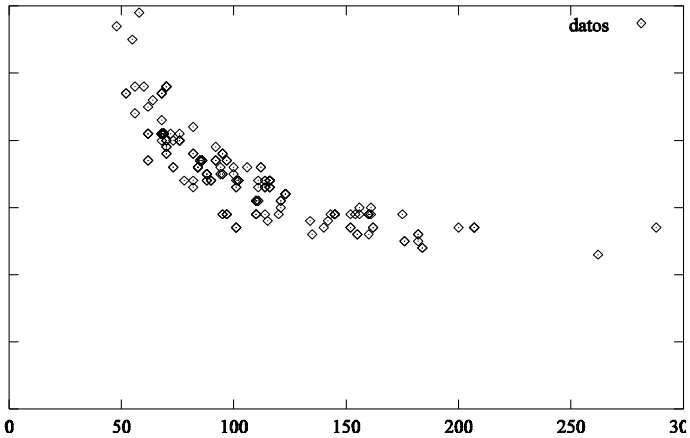
$$a = \bar{y} - b \cdot \bar{x}$$

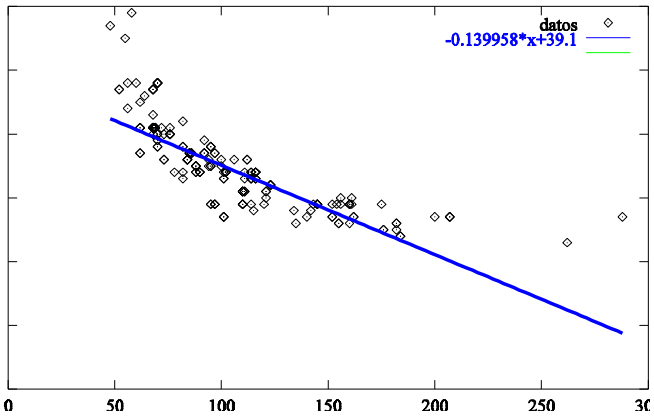
Análisis de regresión





Bastante apropiado





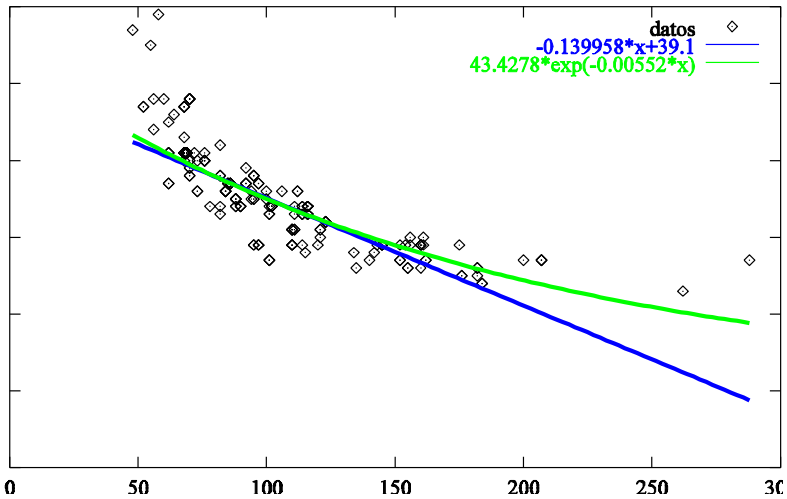
No tan apropiado

Análisis de regresión

- Para estimar curvas es necesario utilizar otra regresión, por ejemplo, **regresión exponencial**: $y=a\cdot e^{bx}$
- ¿Cómo estimamos ahora a y b? Tomando logaritmos

$$\ln(y)=\ln(a\cdot e^{bx}) \Rightarrow \ln(y)=\ln(a)+\ln(e^{bx}) \Rightarrow y^*=a^*+bx$$

- Es decir, tenemos un problema de regresión lineal entre $y^*=ln(y)$ y x. Una vez estimados a^* y b podemos calcular $a=e^{a^*}$



Análisis de regresión

Regresión lineal múltiple

- Cuando hay más de una variable predictora, la ecuación de predicción se transforma en

$$y=a+b_1x_1+b_2x_2+\ldots+b_nx_n$$

  Este problema se conoce como regresión lineal múltiple
- La estimación de los coeficientes es algo más compleja y requiere operar con matrices (y sus inversas)
- Por ejemplo, continuando con el ejemplo CPU.arff, tendríamos

  Class= 0.066 * MYCT + 0.0143*MMIN + 0.0066*MMAX + 0.4945*CACH – 0.1723 * CHMIN + 1.2012 * HMAX – 66.4814

  EAM = 34.31 %

  EAR = 39.26 %
- Existen técnicas más complejas de regresión que aproximan de forma más precisa los datos de entrada

  Regresión no-lineal, regresión logística, ...

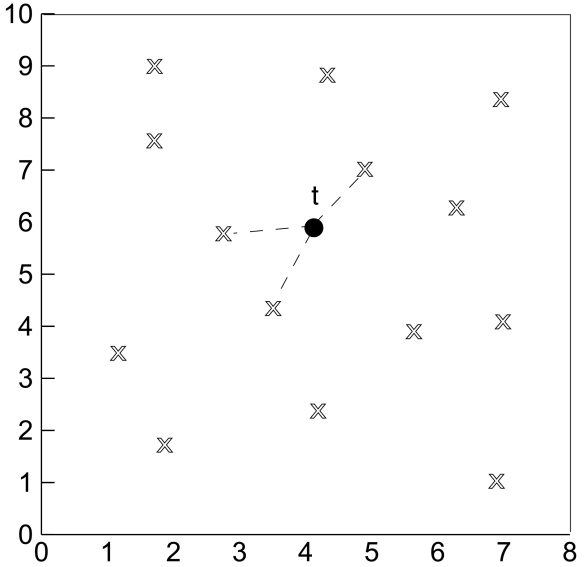
Clasificadores basados en instancias

- Están basados en el aprendizaje por analogía
- Se encuadran en el paradigma perezoso de aprendizaje, frente al voraz al que pertenecen los paradigmas anteriores
  - Perezoso: El trabajo se retrasa todo lo posible
    - No se construye ningún modelo, el modelo es la propia BD o conjunto de entrenamiento
    - Se trabaja cuando llega un nuevo caso a clasificar: Se buscan los casos más parecidos y la clasificación se construye en función de la clase a la que dichos casos pertenecen
- Los algoritmos más conocidos están basados en la regla del vecino más próximo

Clasificadores basados en instancias

Regla del vecino más próximo o *Nearest neighbour* (1-NN)

- Si tenemos *m* instancias $\{e_1, \ldots, e_m\}$ en nuestra base de datos, para clasificar un nuevo ejemplo e' se hará lo siguiente:



1. c_{min} = clase (e_1)
2. d_{min} = d (e_1, e')
3. Para *i=2* hasta *m* hacer
   $d=d(e_i,e')$
   Si ($d<d_{min}$)
   Entonces c_{min} = clase (e_i), $d_{min} = d$
4. Devolver c_{min} como clasificación de e'

- d(·,·) es una función de distancia
- En el caso de variables nominales se utiliza la distancia de *Hamming:*

$$d_h(a,b) = \begin{cases} 0, & si\ a = b \\ 1, & si\ a \neq b \end{cases}$$

Clasificadores basados en instancias

Distancias para las variables numéricas

- Las variables numéricas se suelen normalizar al intervalo [0,1]
- Si e_i^j es el valor de la variable j en e_i, es decir $e_i=(e_i^1,\ldots,e_i^n)$ entonces algunas de las distancias más utilizadas son
  - Euclídea $$d_e(e_1,e_2)=\sqrt{\sum_{i=1}^{n}(e_1^i-e_2^i)}$$
  - Manhattan: $$d_m(e_1,e_2)=\sum_{i=1}^{n}\left|e_1^i-e_2^i\right|$$
  - Minkowski $$d_m^k(e_1,e_2)=\left(\sum_{i=1}^{n}\left|e_1^i-e_2^i\right|^k\right)^{1/k}$$

Como se puede observar, $d_m^1=d_m$ y $d_m^2=d_e$

Clasificadores basados en instancias

- Por tanto, la distancia entre dos instancias e_1 y e_2, utilizando p.e. d_e para las variables numéricas sería

$$d_e(e_1, e_2) = \sqrt{\sum_i (e_1^i - e_2^i)^2 + \sum_j d_h(e_1^j, e_2^j)}$$

siendo *i* el índice que se utiliza para recorrer las variables numéricas y *j* el índice que se utiliza para recorrer las variables nominales

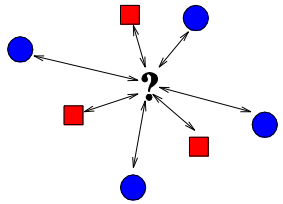
- Tratamiento de valores desconocidos: si ej_1=? Y ej_2=? Entonces la distancia asociada a la j-ésima componente es la máxima, es decir, 1

- Una crítica a esta regla es que todas las variables se consideran con igual importancia. La solución es asignar pesos a los atributos de forma que se pondere su importancia dentro del contexto

$$d_e(e_1, e_2) = \sqrt{\sum_i w_i \cdot (e_1^i - e_2^i)^2 + \sum_j w_j \cdot d_h(e_1^j, e_2^j)}$$

Clasificadores basados en instancias

- La extensión a la regla del vecino más próximo, es considerar los k vecinos más próximos: k-NN (NN = 1-NN)
- Funcionamiento: Dado *e* el ejemplo a clasificar
  1. Seleccionar los k ejemplos con $K = \{e_1,\ldots, e_k\}$ tal que, no existe ningún ejemplo *e'* fuera de *K* con $d(e,e')<d(e,e_i),\ i=1,\ldots,k$
  2. Devolver la clase que más se repite en el conjunto $\{clase(e_1), \ldots, clase\ (e_k)\}$ (la clase mayoritaria)
- Por ejemplo, si k=7 el siguiente caso (?) se clasificaría como ●



- Se podría tratar de forma diferente a los k-vecinos, p.e., dependiendo de la distancia al objeto a clasificar. De esta forma tendríamos: Clasificación
  - Voto por la mayoría ●
  - Voto con pesos en función de la distancia □

Clasificadores basados en instancias

- El algoritmo k-NN es robusto frente al ruido cuando se utilizan valores de k moderados (k>1)
- Es bastante eficaz, puesto que utiliza varias funciones lineales locales para aproximar la función objetivo
- Es válido para clasificación y para predicción numérica (devolviendo la media o la media ponderada por la distancia)
- Es muy ineficiente en memoria ya que hay que almacenar toda la BD
- La distancia entre vecinos podría estar dominada por variables irrelevantes
  - Selección previa de características
- Su complejidad temporal (para evaluar un ejemplo) es $O(dn^2)$ siendo $O(d)$ la complejidad de la distancia utilizada
  - Una forma de reducir esta complejidad es mediante el uso de prototipos
- El algoritmo *k*-NN está disponible en WEKA (KNIME) bajo el nombre de ibk. Permite voto por mayoría o voto ponderado por la distancia *(1/d y 1-d)*. No permite ponderar la variables

Clasificadores basados en instancias: lazy learning y vecinos cercanos

Clasificador Del Vecino Más Cercano (k-NN)

Dado el siguiente conjunto con 4 instancias, 3 atributos y 2 clases:

x_1: 0.4 0.8 0.2 positiva

x_2: 0.2 0.7 0.9 positiva

x_3: 0.9 0.8 0.9 negativa

x_4: 0.8 0.1 0.0 negativa

Calculamos la distancia del ejemplo con todos los del conjunto:

$$d(x_{1,}y_1) = \sqrt{(0.4-0.7)^2 + (0.8-0.2)^2 + (0.2-0.1)^2} = 0.678$$

$$d(x_{2,}y_1) = \sqrt{(0.2-0.7)^2 + (0.7-0.2)^2 + (0.9-0.1)^2} = 1.068$$

$$d(x_{3,}y_1) = \sqrt{(0.9-0.7)^2 + (0.8-0.2)^2 + (0.9-0.1)^2} = 1.020$$

$$d(x_{4,}y_1) = \sqrt{(0.8-0.7)^2 + (0.1-0.2)^2 + (0.0-0.1)^2} = 0.173$$

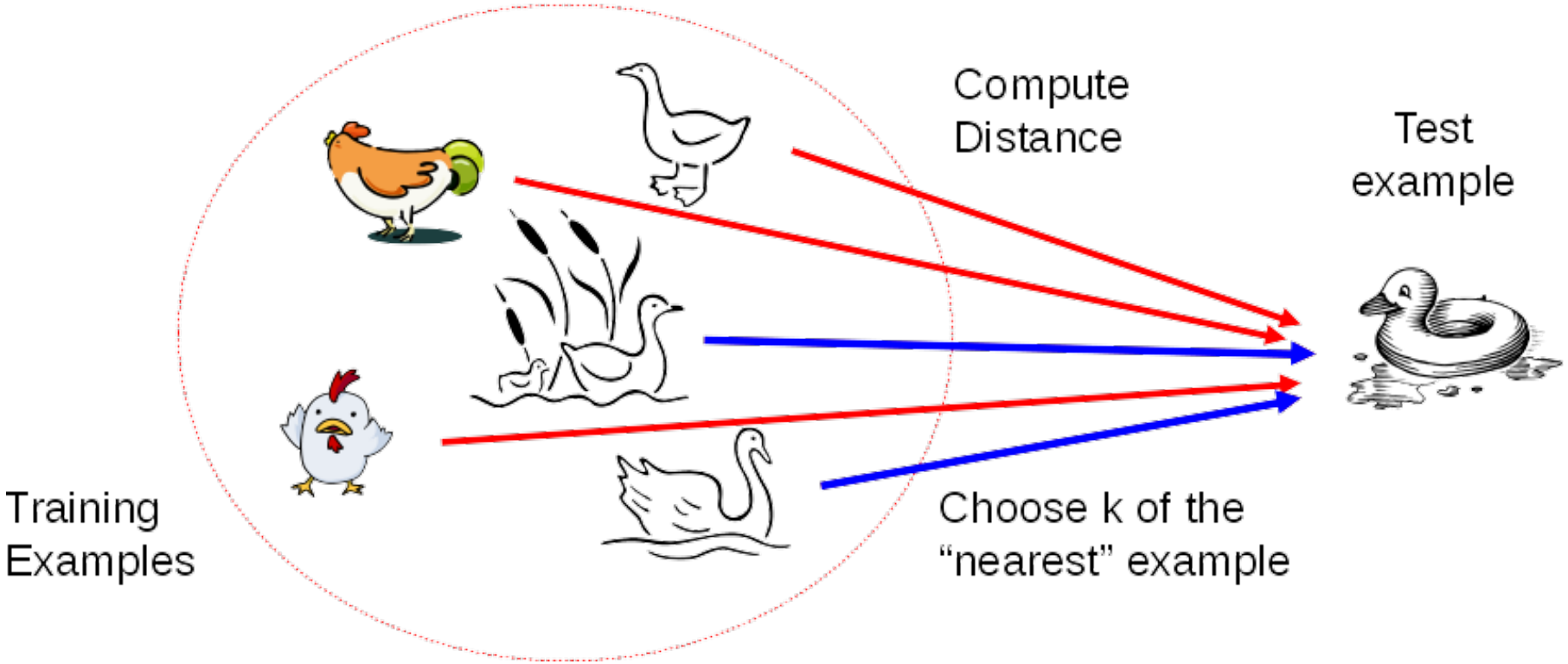
Por tanto, el ejemplo se clasificará con respecto a la clase *negativa.*

IMPORTANTE: Los atributos deben estar normalizados [0,1] para no priorizarlos sobre otros.

Clasificadores basados en instancias: lazy learning y vecinos cercanos

Clasificador del Vecino Más Cercano (k-NN)

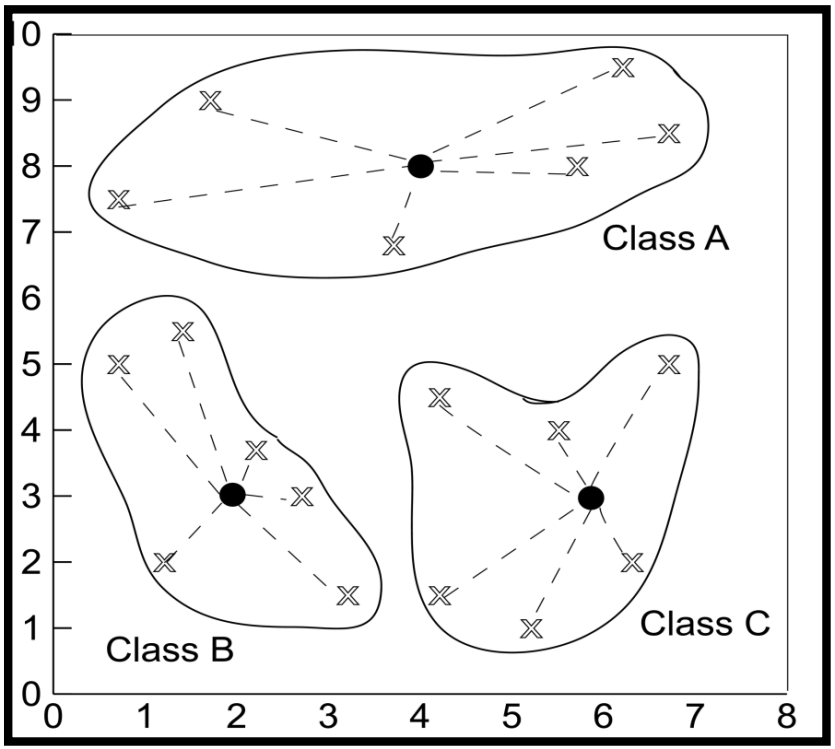
- **If it walks like a duck, quacks like a duck, then it's probably a duck**



Clasificador del Vecino Más Cercano kNN

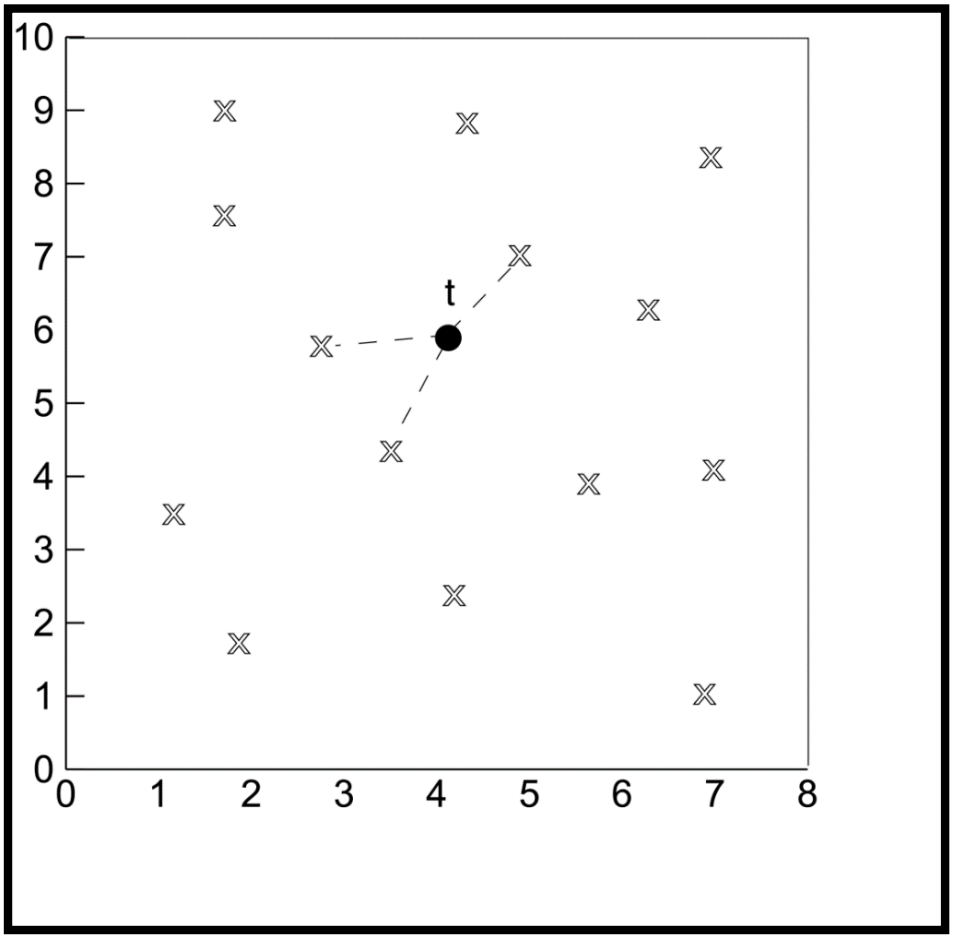
Clasificador Del Vecino Más Cercano (k-NN): Basado en distancias

Basado en Distancias



Clasificador del Vecino Más Cercano kNN

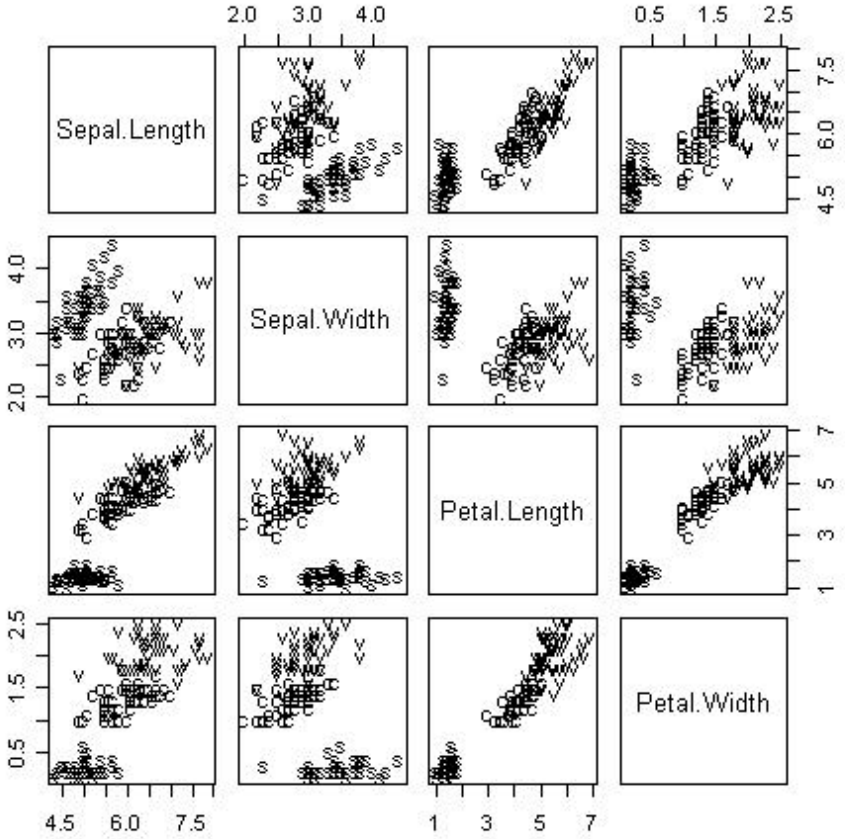
Clasificador Del Vecino Más Cercano (k-NN): Ejemplo para k=3

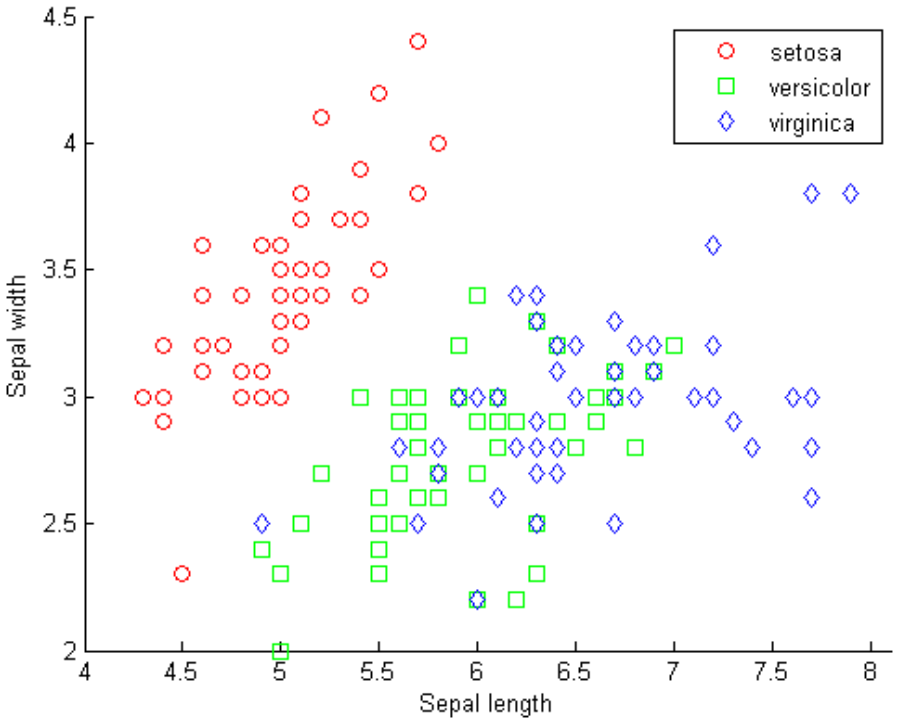


Clasificador del Vecino Más Cercano kNN

Clasificador Del Vecino Más Cercano (k-NN). Problema en las fronteras

setosa, versicolor (C) y virginica

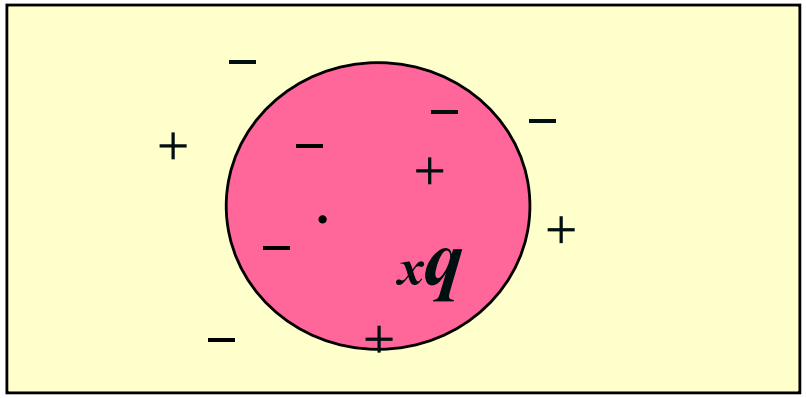


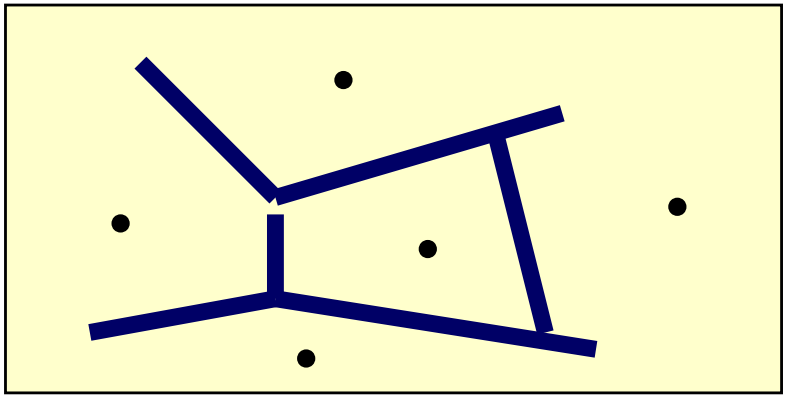


Clasificador del Vecino Más Cercano kNN

Clasificador Del Vecino Más Cercano (k-NN)

- ***k*-NN devuelve la clase más repetida de entre todos los k ejemplos de entrenamiento cercanos a *xq*.**
- **Diagrama de Voronoi: superficie de decisión inducida por 1-NN para un conjunto dado de ejemplos de entrenamiento.**





Técnicas de clasificación válidas para regresión: kNN, RNN

- **Métodos basados en ejemplos/instancias:**

  Al utilizar kNN, si los *k* vecinos más próximos $\{e_1,\ldots,e_k\}$ tienen valores $\{v_1,..,v_k\}$ para la variable objetivo, entonces el valor a devolver para el objeto analizado *e'* sería

$$v = \begin{cases} \dfrac{\sum_{i=1}^{k} v_i}{k} & \text{si todos cuentan igual} \\ \dfrac{\sum_{i=1}^{k} w_i \cdot v_i}{\sum_{i=1}^{k} w_i} & \text{si se hace un voto ponderado} \end{cases}$$

por ejemplo, con $w_i = 1/d(e_i,e')$

Principios básicos de *machine learning*

- Conceptos básicos. Ciencia de Datos, Minería de Datos, Big Data, Machine Learning.
- Proceso de Minería de Datos
- Técnicas de Minería de Datos: Clasificación, Regresión, Agrupamiento, Asociación
- El Poder de los Datos. Casos de estudio
- Clasificación y Regresión. Predicción por similitud: K-Nearest Neighbour (KNN).
- Validación de Clasificadores
- Clasificación con Árboles de Decisión

Evaluación

Métricas

Cómo evaluar la “calidad” de un modelo de clasificación

Métodos

Cómo estimar, de forma fiable, la calidad de un modelo.

Comparación

Cómo comparar el rendimiento relativo
de dos modelos de clasificación alternativos

Evaluación: Métricas

Matriz de confusión (confusion matrix)

		Predicción	
		C_P	**C_N**
Clase real	**C_P**	**TP**: True positive	**FN**: False negative
	C_N	**FP**: False positive	**TN**: True negative

Precisión del clasificador

accuracy = (TP+TN)/(TP+TN+FP+FN)

Evaluación: Métricas

Limitaciones de la precisión ("accuracy") :

Supongamos un problema con 2 clases:

- 9990 ejemplos de la clase 1
- 10 ejemplos de la clase 2

Si el modelo de clasificación siempre dice que los ejemplos son de la clase 1, su precisión es

9990/10000 = 99.9%

Totalmente engañosa, ya que nunca detectaremos ningún ejemplo de la clase 2.

Evaluación: Métricas

Alternativa: Matriz de costes

C(i|j)		Predicción	
		C_P	C_N
Clase real	C_P	C(P|P)	C(N|P)
	C_P	C(P|N)	C(N|N)

El coste de clasificación será proporcional a la precisión del clasificador sólo si

$\forall i,j: i \neq j \quad C(i|j) = C(j|i)$

$C(i|i) = C(j|j)$

Evaluación: Métricas

Medidas "cost-sensitive"

		Predicción	
		C_P	C_N
Clase real	C_P	**TP**: True positive	**FN**: False negative
	C_N	**FP**: False positive	**TN**: True negative

precision = TP/(TP+FP)

True positive recognition rate

recall = sensitivity = TP/P = TP/(TP+FN)

True negative recognition rate

specificity = TN/N = TN/(TN+FP)

Evaluación: Métricas

Medidas "cost-sensitive"

		Predicción	
		C_P	C_N
Clase real	C_P	**TP**: True positive	**FN**: False negative
	C_N	**FP**: False positive	**TN**: True negative

F-measure

F = 2*precision*recall / (precision+recall)

F= 2TP / (2TP+FP+FN)

Evaluación: Métricas

		Predicción	
		C_P	C_N
Real	C_P	TP	FN
	C_N	FP	TN

Accuracy

		Predicción	
		C_P	C_N
Real	C_P	TP	FN
	C_N	FP	TN

Recall

		Predicción	
		C_P	C_N
Real	C_P	TP	FN
	C_N	FP	TN

Precision

		Predicción	
		C_P	C_N
Real	C_P	TP	FN
	C_N	FP	TN

F-measure

Evaluación: Métodos

Para evaluar la precisión de un modelo de clasificación nunca debemos utilizar el conjunto de entrenamiento (lo que nos daría el "**error de resustitución**" del clasificador), sino un conjunto de prueba independiente:

Por ejemplo, podríamos reservar 2/3 de los ejemplos disponibles para construir el clasificador y el 1/3 restante lo utilizaríamos de conjunto de prueba para estimar la precisión del clasificador.

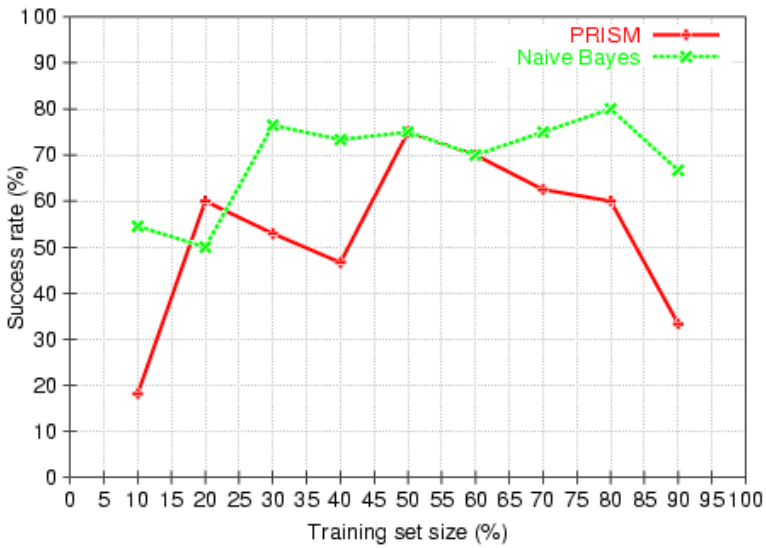
Validación de clasificadores

- El objetivo de los métodos de validación es realizar una estimación honesta de la bondad del clasificador construido
- Utilizar como bondad la tase de acierto sobre el conjunto de entrenamiento no es realista
  - El porcentaje obtenido suele ser demasiado optimista debido a que el modelo estará sobreajustado a los datos utilizados durante el proceso de aprendizaje
- Existen distintas técnicas de validación de clasificadores, entre ellas
  - *Hold-out*
  - Validación cruzada
  - *Leave-one-out*
  - *Boostraping*

Validación de clasificadores

Hold-out

- Consiste en dividir la BD en dos conjuntos independientes: entrenamiento (CE) y test (CT)
- El tamaño del CE normalmente es mayor que el del CT (2/3, 1/3, 4/5, 1/5,…)
- Los elementos del CE suelen obtenerse mediante muestreo sin reemplazamiento de la BD inicial. El CT está formado por los elementos no incluidos en el CE
- Suele utilizarse en BBDD de tamaño grande



Test set (%) + Training set (%) = 100%

Validación de clasificadores

Validación cruzada

- Consiste en:
  1. Dividir la BD en *k* subconjuntos (*folds*), $\{S_1,\ldots,S_k\}$ de igual tamaño
  2. Aprender *k* clasificadores utilizando en cada uno de ellos un CE distinto. Validar con el CT correspondiente

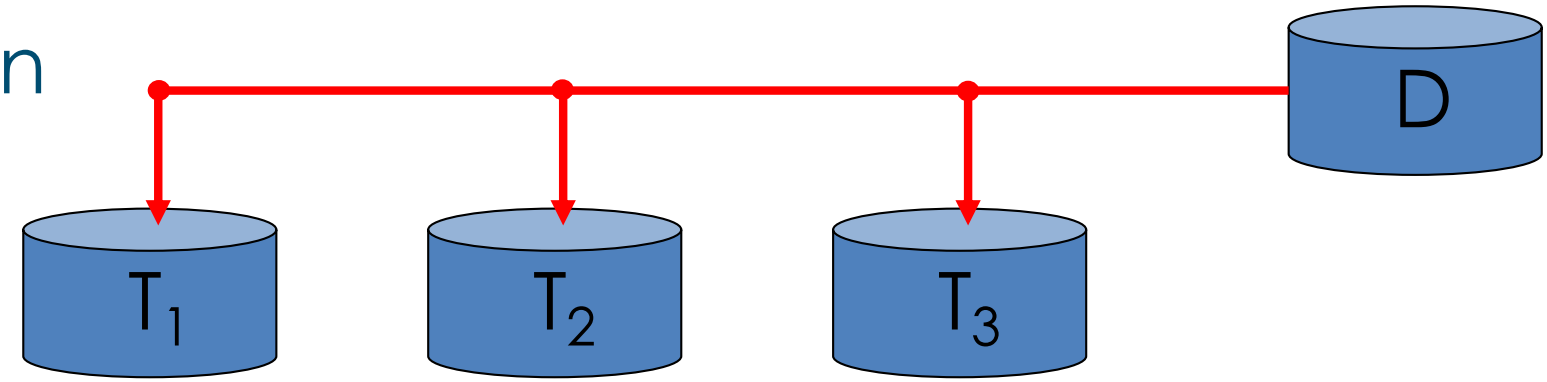
$$CE = S_1 \cup \ldots \cup S_{i-1} \cup S_{i+1} \cup \ldots S_k$$

$$CT = S_i$$

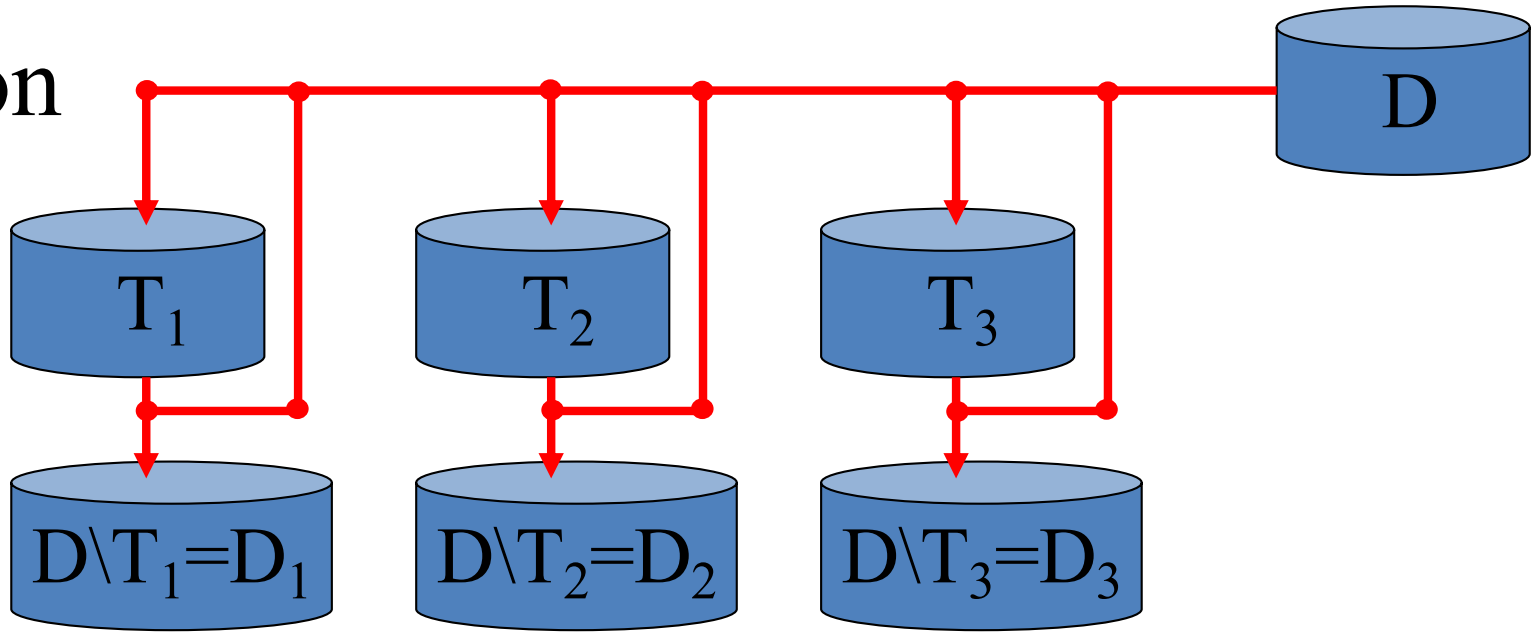
  3. Devolver como tasa de acierto (error) el promedio obtenido en las *k* iteraciones

- Validación cruzada estratificada: Los subconjuntos se estratifican en función de la variable clase
- Valores típicos de k=5,10
- Suele utilizarse en BBDD de tamaño moderado

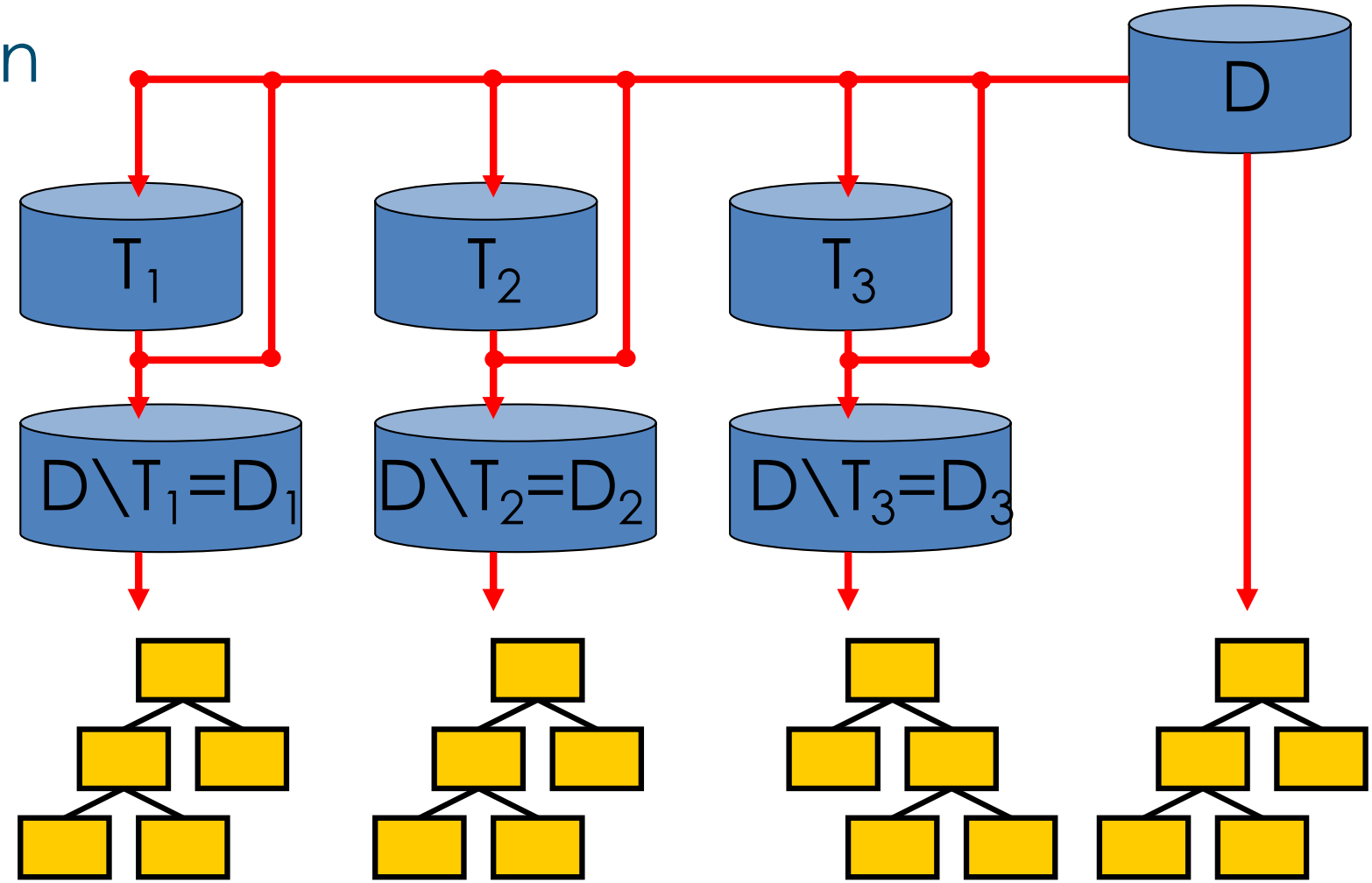
- Partition



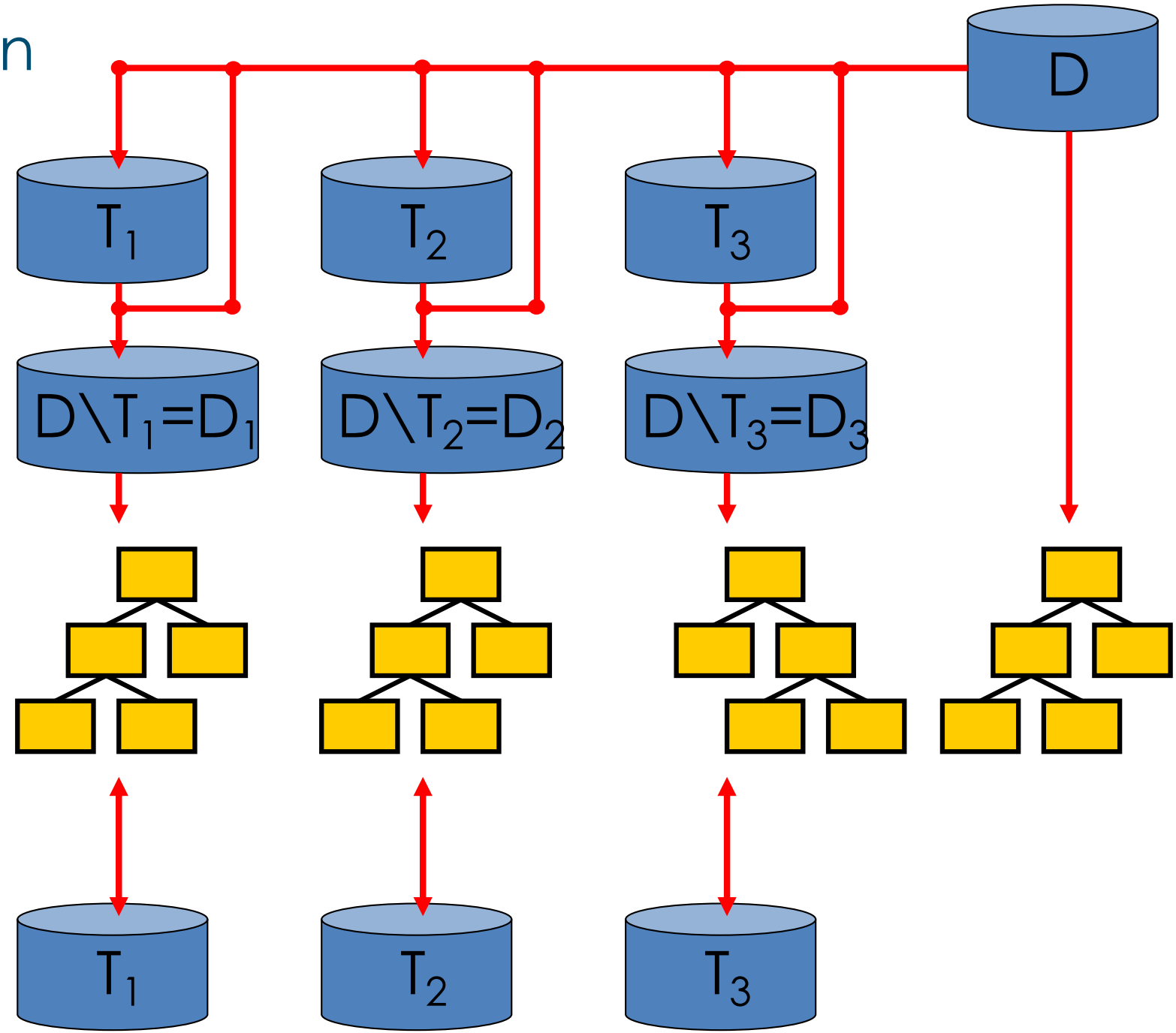
- Partition
- Train



- Partition
- Train



- Partition
- Train
- Test



Validación de clasificadores

Leaving-one-out

- Es un caso especial de validación cruzada en el que *k* es igual al número de registros
  - Tiene la ventaja de que el proceso es determinista y de que en todo momento se utiliza el máximo posible de datos para la inducción del clasificador
  - Se utiliza en BBDD muy pequeñas, debido a su alto costo computacional

Validación de clasificadores

Bootstrap

- Está basado en el proceso de muestreo con reemplazo
- A partir de una BD con *n* registros se obtiene un CE con *n* casos
- Como CT se utilizan los registros de la BD no seleccionados para el CE

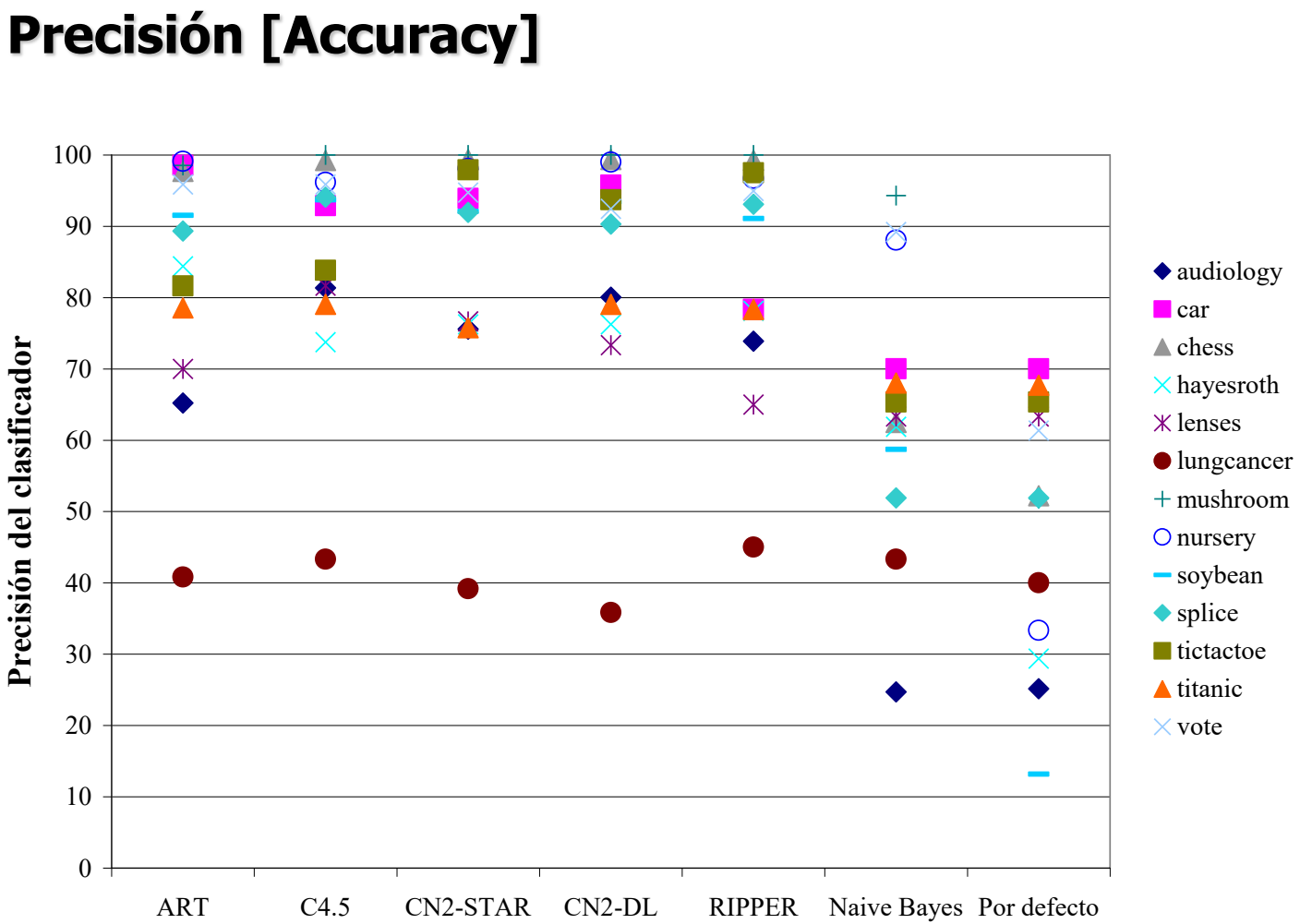
  ¿Cuántos casos habrá en CT? ¿qué porcentaje respecto a *n*?
  - La probabilidad de que se elija un registro es *1/n*. La probabilidad de que no se elija es *1-1/n*
  - Se hacen *n* extracciones, por tanto la probabilidad de que un ejemplo no sea elegido es

$$\left(1-\frac{1}{n}\right)^{n} \approx e^{-1} = 0.368$$

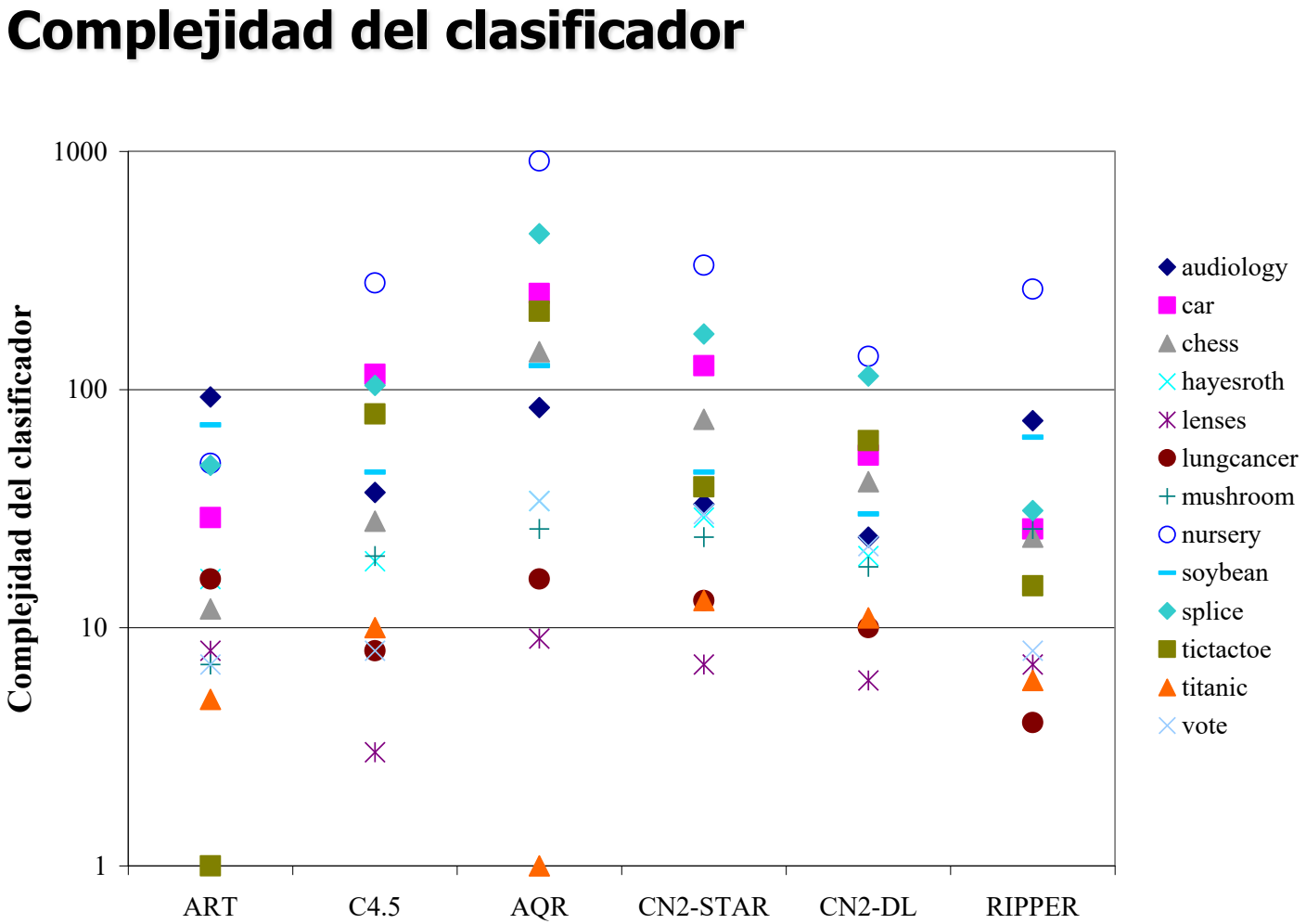
  - El CE tendrá aproximadamente el 63.2% de los registros de la BD y el CT el 36.8 %
- Esta técnica se conoce como 0.632 bootstrap
- El error sobre el CT suele ser bastante pesimista por lo que se corrige

$$error = 0.632 \cdot error_{CT} + 0.368 \cdot error_{CE}$$

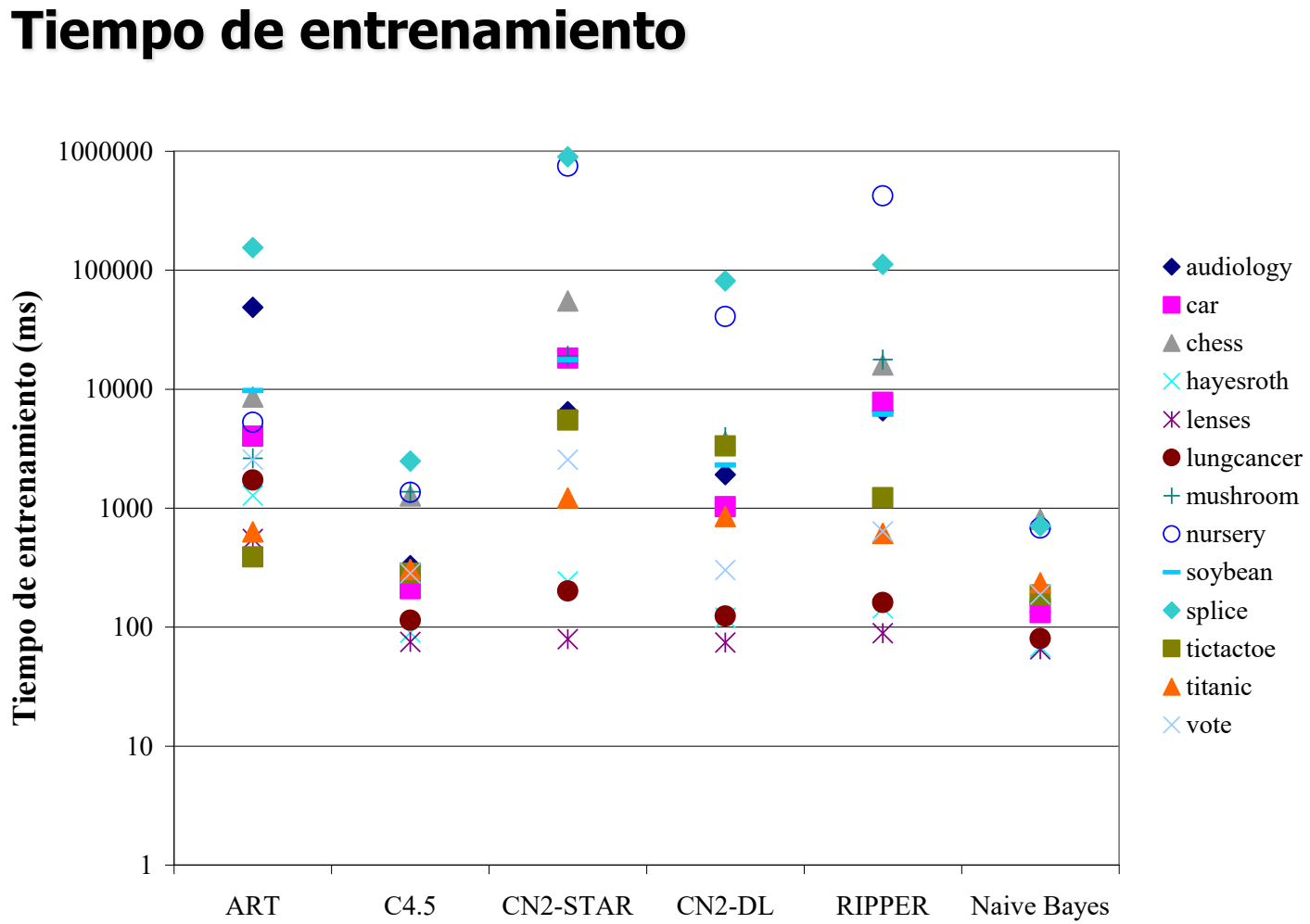
Evaluación: Comparación



Evaluación: Comparación

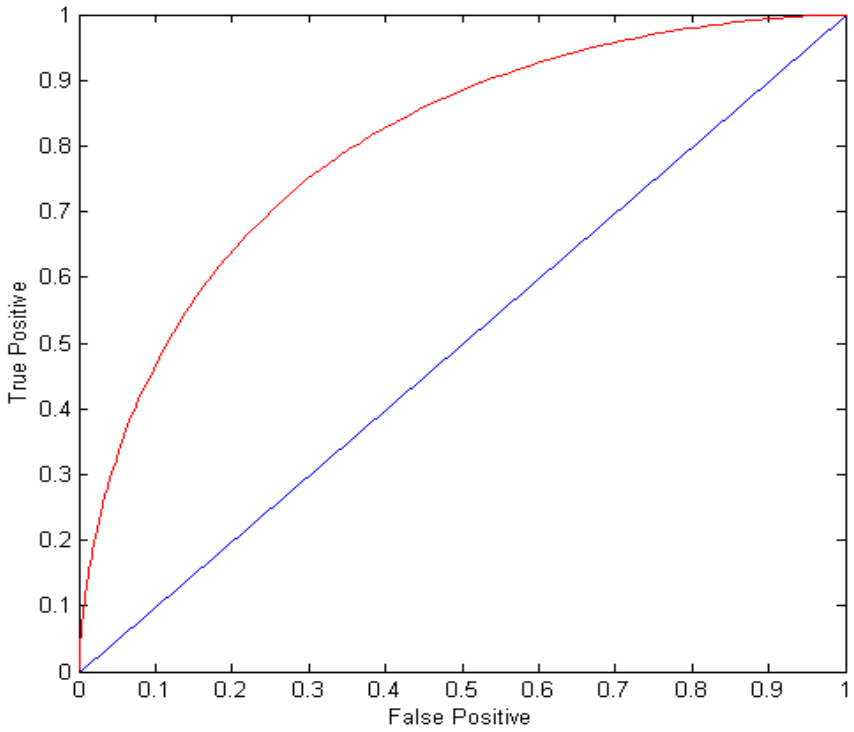


Evaluación: Comparación



Evaluación: Comparación

Curvas ROC (Receiver Operating Characteristics)



Eje vertical: "true positive rate" TPR = TP/(TP+FN)

Eje horizontal: "false positive rate" FPR = FP/(FP+TN)

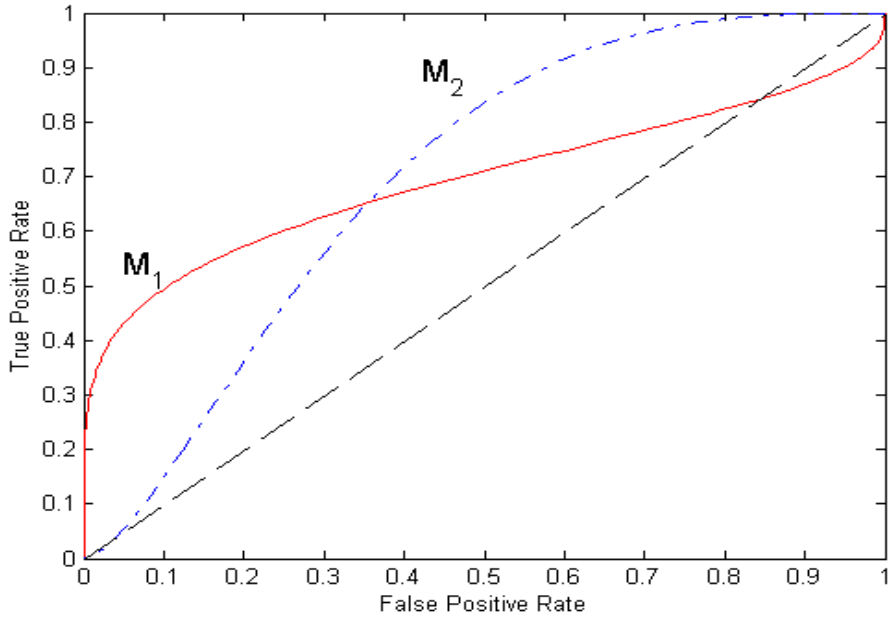
Evaluación: Comparación

Curvas ROC

- **Desarrolladas en los años 50 para analizar señales con ruido: caracterizar el compromiso entre aciertos y falsas alarmas.**
- **Permiten comparar visualmente distintos modelos de clasificación.**
- **El área que queda bajo la curva es una medida de la precisión (accuracy) del clasificador:**
  - **Cuanto más cerca estemos de la diagonal (área cercana a 0.5), menos preciso será el modelo.**
  - **Un modelo "perfecto" tendrá área 1.**

Evaluación: Comparación

Curvas ROC



Ningún modelo es consistentemente mejor que el otro: M1 es mejor para FPR bajos, M2 para FPR altos.

Evaluación: Comparación

Curvas ROC

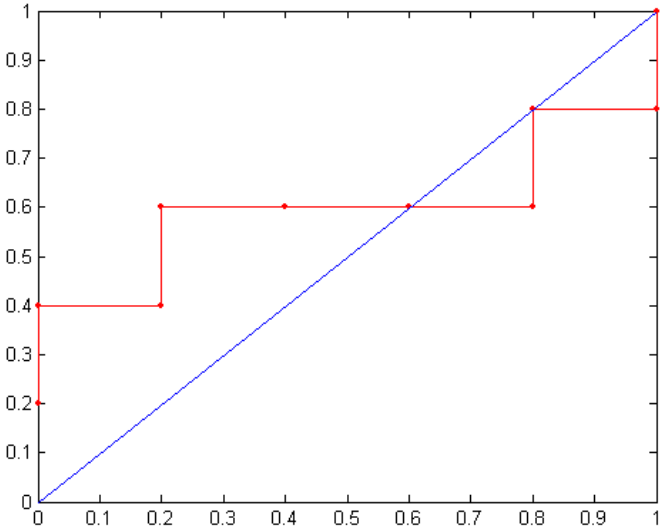
¿Cómo construir la curva ROC?

- **Se usa un clasificador que prediga la probabilidad de que un ejemplo E pertenezca a la clase positiva P(+|E)**
- **Se ordenan los ejemplos en orden decreciente del valor estimado P(+|E)**
- **Se aplica un umbral para cada valor distinto de P(+|E), donde se cuenta el número de TP, FP, TN y FN.**

$$TPR = TP/(TP+FN)$$
$$FPR = FP/(FP+TN)$$

Evaluación: Comparación

Curvas ROC



Ejemplo	P(+|E)	Clase
1	0.95	+
2	0.93	+
3	0.87	-
4	0.85	-
5	0.85	-
6	0.85	+
7	0.76	-
8	0.53	+
9	0.43	-
10	0.25	+

Class	+	-	+	-	-	-	+	-	+	+	
	0.25	0.43	0.53	0.76	0.85	0.85	0.85	0.87	0.93	0.95	1.00
TP	5	4	4	3	3	3	3	2	2	1	0
FP	5	5	4	4	3	2	1	1	0	0	0
TN	0	0	1	1	2	3	4	4	5	5	5
FN	0	1	1	2	2	2	2	3	3	4	5
TPR	1	0.8	0.8	0.6	0.6	0.6	0.6	0.4	0.4	0.2	0
FPR	1	1	0.8	0.8	0.6	0.6	0.2	0.2	0	0	0

Principios básicos de *machine learning*

- Conceptos básicos. Ciencia de Datos, Minería de Datos, Big Data, Machine Learning.
- Proceso de Minería de Datos
- Técnicas de Minería de Datos: Clasificación, Regresión, Agrupamiento, Asociación
- El Poder de los Datos. Casos de estudio
- Clasificación y Regresión. Predicción por similitud: K-Nearest Neighbour (KNN).
- Regresión. Medidas de evaluación, similitud y regresión lineal
- Validación de Clasificadores
- Clasificación con Árboles de Decisión

Clasificación con árboles

Árboles de decisión

1. Definición de árboles de decisión
2. Construcción de árboles de decisión
3. Criterios de selección de variables
4. Particionamiento del espacio con un árbol de decisión
5. Ventajas e inconvenientes del uso de árboles de decisión en clasificación
6. Algunos algoritmos de minería de datos basados en árboles de decisión

Definición de árboles de decisión

- Un árbol de decisión es un clasificador que en función de un conjunto de atributos permite determinar a que clase pertenece el caso objeto de estudio

- La estructura de un árbol de decisión es:
  - Cada hoja es una categoría (clase) de la variable objeto de la clasificación
  - Cada nodo es un nodo de decisión que especifica una prueba simple a realizar
  - Los descendientes de cada nodo son los posibles resultados de la prueba del nodo

Definición de árboles de decisión

Insurance Risk Assessment

Age	Car Type	Risk
23	family	High
17	sports	High
43	sports	High
68	family	Low
32	truck	Low
20	family	High