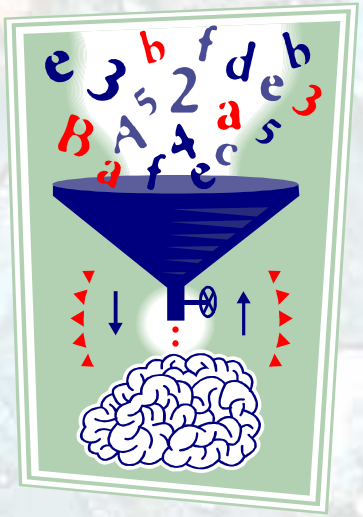
TÉCNICAS DE MINERÍA DE DATOS

TÉCNICAS DE MODELADO PREDICTIVO



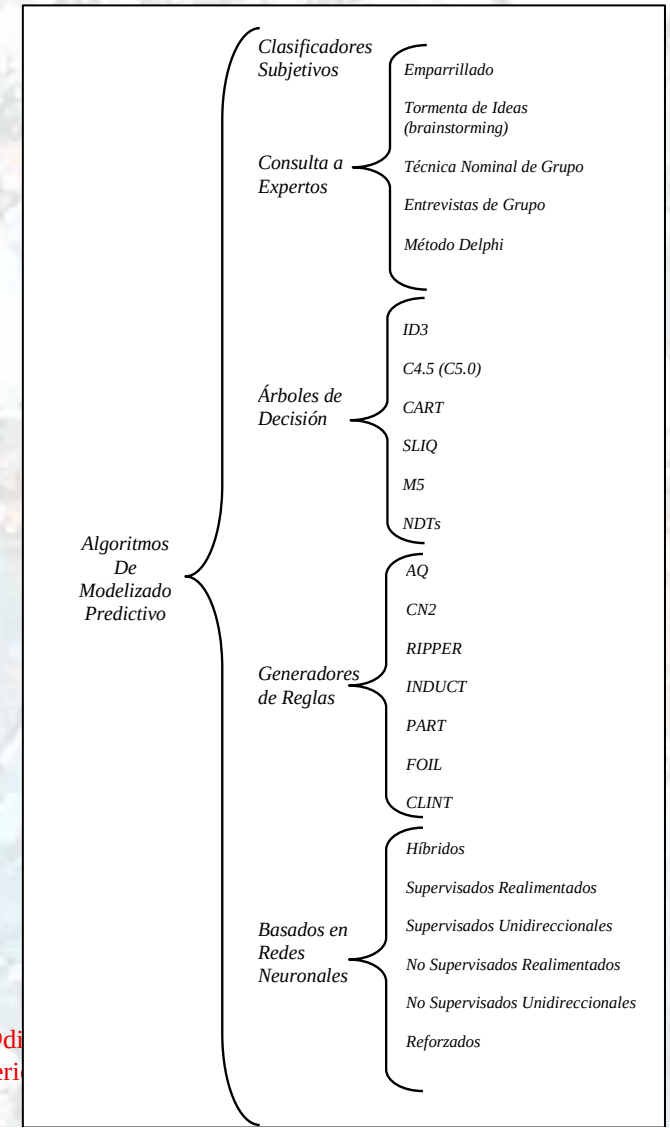
TÉCNICAS DE MODELADO PREDICTIVO

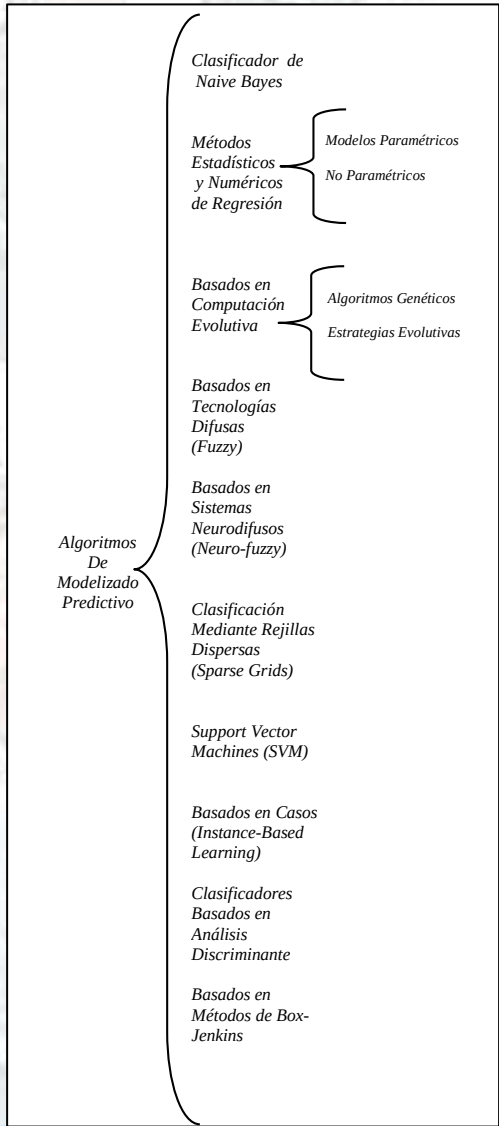
Generalmente, se pueden dividir en:

- **Clasificadores:** Para cada nueva observación el modelo intenta clasificarla en una de las clases previamente establecidas.

- **Regresores:** Para cada nueva observación el modelo intenta predecir el valor continuo más probable

Vamos a ver algunas de las más utilizadas...

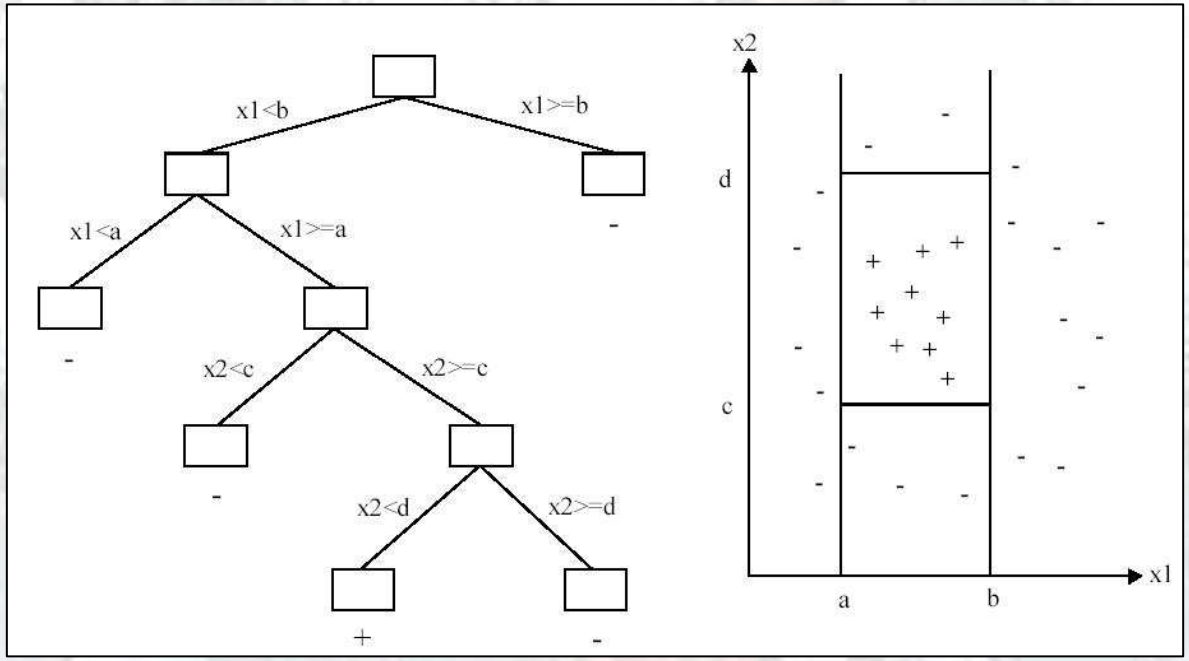


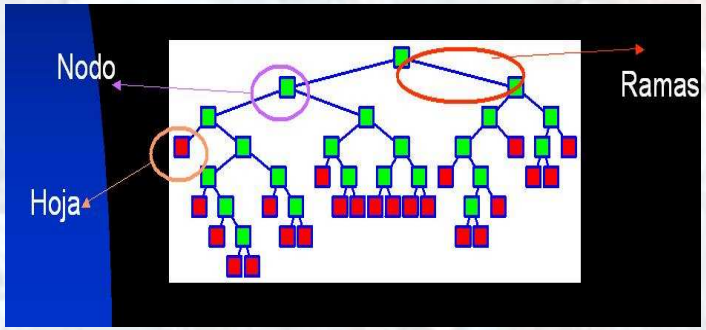


Dr. Fco. Javier Martínez de Pisón Ascacíbar (fjmartin@di
Grupo EDMANS (Engineering Data Mining And Numeri
http://www.mineriadatos.com

TÉCNICAS DE MODELADO PREDICTIVO:
Árboles de Decisión

Los árboles de decisión son unos de los algoritmos clasificadores más conocidos y usados en las tareas de Data Mining [WIT00][DAE02], ya que son una forma de representación sencilla para clasificar ejemplos de un número finito de clases. Se basan en la partición del conjunto de ejemplos según ciertas condiciones que se aplican a los valores de las características. Su potencia descriptiva viene limitada por las condiciones o reglas con las que se divide el conjunto de entrenamiento.

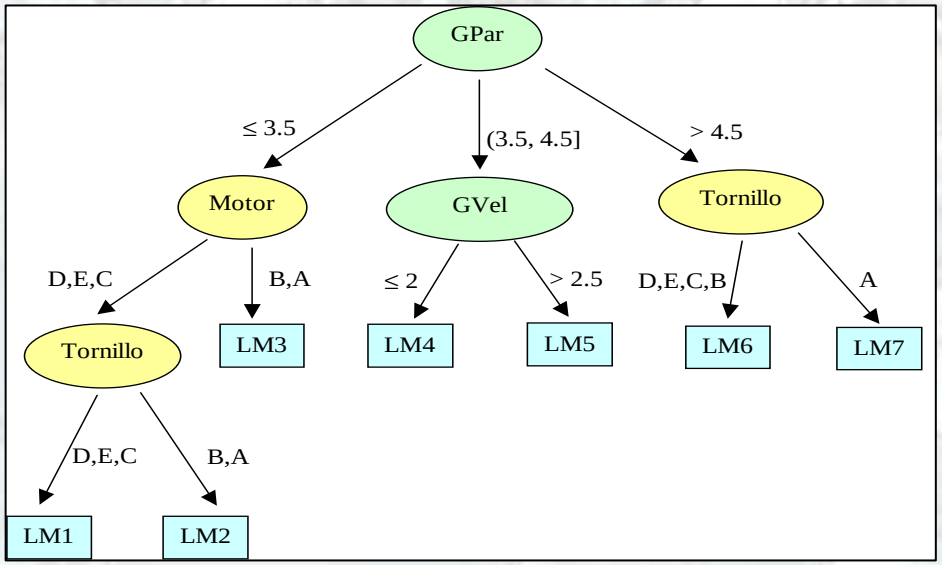


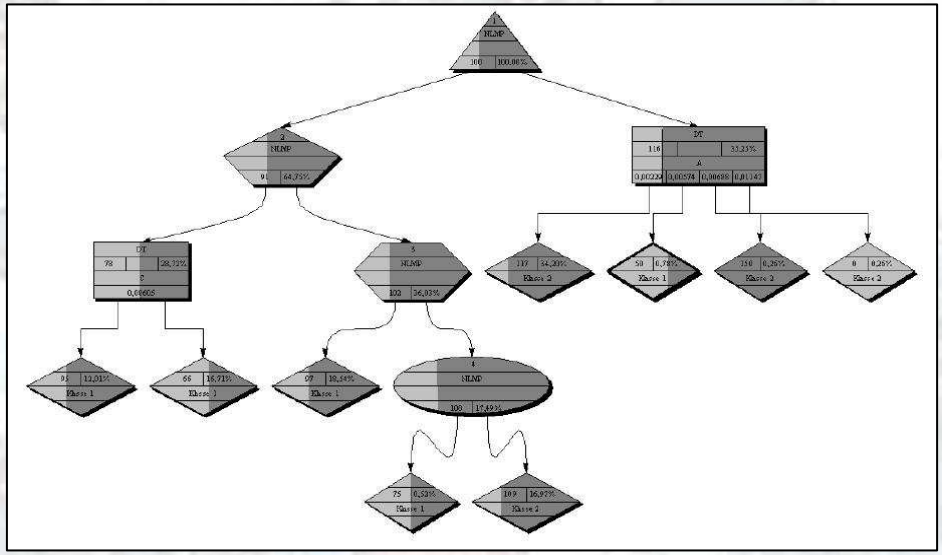


Árboles de Decisión
- *CART*
- *ID3*
- *C4.5 (C5.0)*
- *SLIQ*
- *M5*
- *NDTs*

Dr. Fco. Javier Martínez de Pisón Ascacíbar (fjmartin@dim.unirioja.es)
Grupo EDMANS (Engineering Data Mining And Numerical Simulation)
http://www.mineriadatos.com

TÉCNICAS DE MODELADO PREDICTIVO:
Árboles de Decisión





Árbol Mixto: Árbol generado que utiliza siete modelos lineales para predecir el tiempo de establecimiento de un servomotor a partir de dos variables numéricas y dos nominales [WIT00].

NDTS (NON-LINEAR DECISIÓN TREES)
Realmente este tipo de árboles se basan en las técnicas vistas anteriormente para desarrollar el árbol, pero donde en cada nodo se sustituye por un clasificador cualquiera: lineales, no lineales, etc. que clasifican cada subárbol que parte de cada nodo [PRU02].

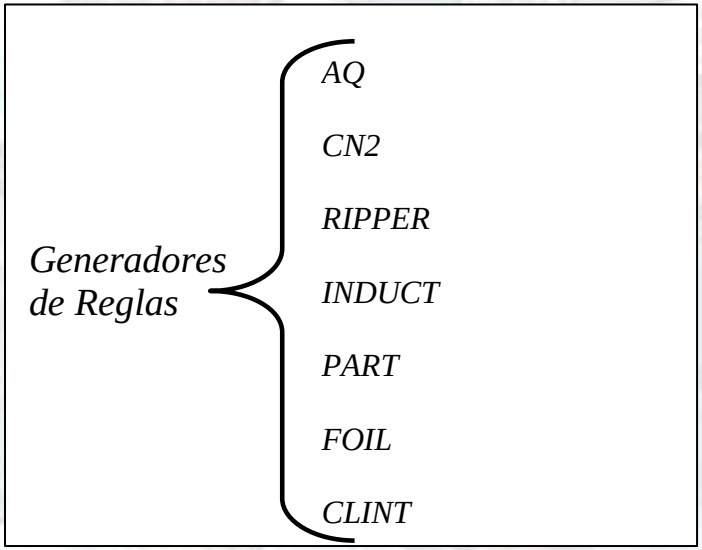
Dr. Fco. Javier Martínez de Pisón Ascacíbar (fjmartin@dim.unirioja.es)
Grupo EDMANS (Engineering Data Mining And Numerical Simulation)
http://www.mineriadatos.com

TÉCNICAS DE MODELADO PREDICTIVO:
Generadores de Reglas

Una desventaja de los árboles de decisión es que tienden a ser demasiado grandes en aplicaciones reales y, por tanto, se hacen difíciles de interpretar desde el punto de vista humano. Para suplir estos inconvenientes aparecen los algoritmos generadores de reglas que intentan generar conjuntos de reglas que sean fácilmente comprensibles.

- **Algunas reglas inducidas pueden derivar de la construcción de un árbol de decisión, siendo primero generado el árbol de decisión y después trasladado a un conjunto de reglas (AQ, CN2, RIPPER, INDUCT, PART).**

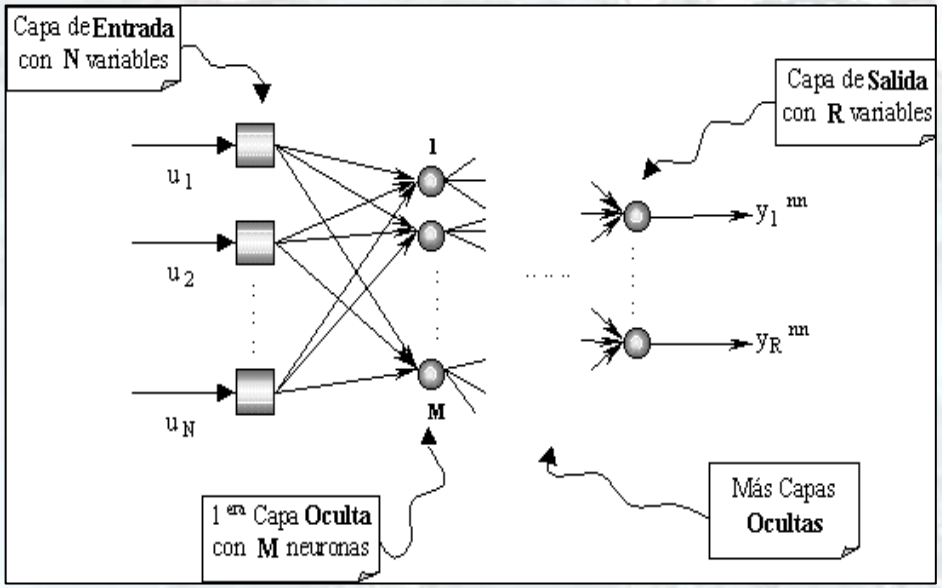
- **Otros algoritmos se basan en el uso de técnicas de aprendizaje con lógica de predicados (ILP, Inductive Logic Programming). (FOIL, FFOIL, CLINT, etc.)**

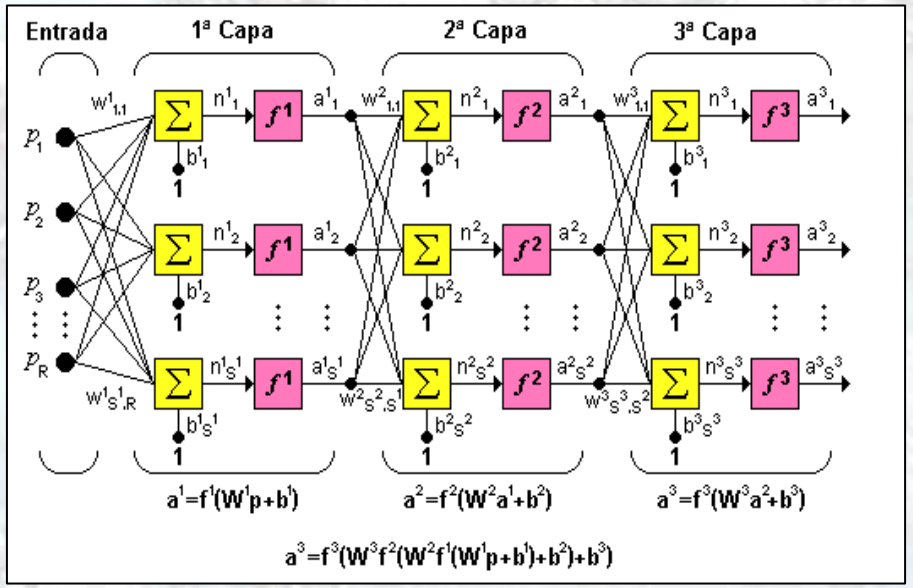


Dr. Fco. Javier Martínez de Pisón Ascacíbar (fjmartin@dim.unirioja.es)
Grupo EDMANS (Engineering Data Mining And Numerical Simulation)
http://www.mineriadatos.com

TÉCNICAS DE MODELADO PREDICTIVO:
Redes Neuronales

Una red neuronal es una estructura compuesta por muchas unidades, muy simples, de procesamiento o neuronas, cada una con memoria local, habitualmente pequeña. Las neuronas se conectan mediante canales de comunicación, denominados conexiones, que manejan datos numéricos. Operan sólo con los datos locales por lo que tienen un gran potencial para el procesamiento paralelo dado que los cálculos de los componentes en cada neurona son independientes.





Dr. Fco. Javier Martínez de Pisón Ascacíbar (fjmartin@dim.unirioja.es)
Grupo EDMANS (Engineering Data Mining And Numerical Simulation)
http://www.mineriadatos.com

TÉCNICAS DE MODELADO PREDICTIVO:
Redes Neuronales

Las Redes Neuronales deben ser entrenadas y después se debe comprobar la capacidad de predicción de las mismas ante nuevas entradas.

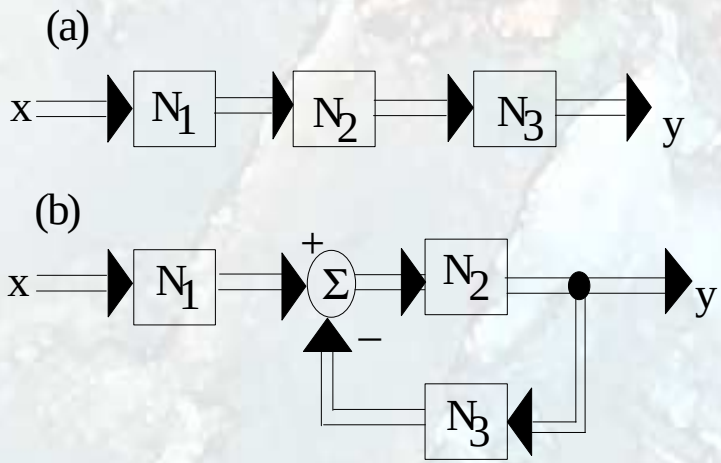
Según el entrenamiento se dividen en:

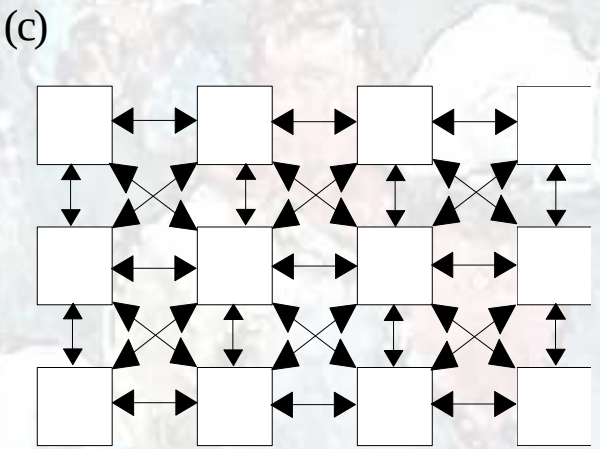
- Redes Supervisadas: Durante la fase de aprendizaje, se indica a la red qué salida debe producir cada patrón, ajustando los pesos en función de ese valor.
- Redes No Supervisadas: La Red localiza en los datos de entrada propiedades que utiliza para separar los patrones en clases. El aprendizaje no supervisado es característico de las redes utilizadas en los casos en que los datos no tienen a priori ningún tipo de clasificación. La red se utiliza para detectar las regularidades intrínsecas de los datos estableciendo así la mejor clasificación posible.

TÉCNICAS DE MODELADO PREDICTIVO:
Redes Neuronales

Las arquitecturas de las redes neuronales se dividen en tres grandes categorías:

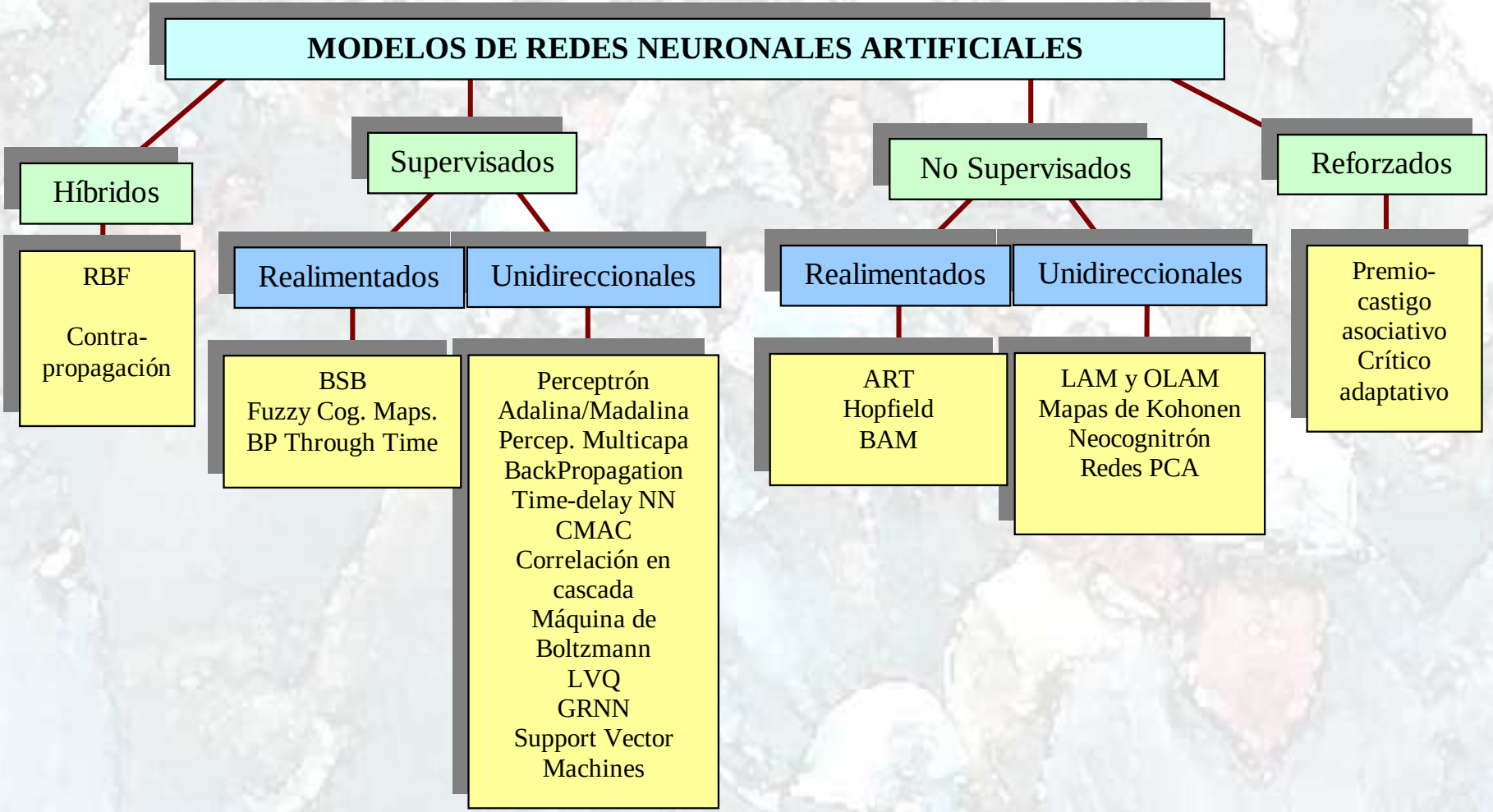
a) Redes Progresivas o Unidireccionales (*Feedforward Networks*).
b) Redes Recurrentes o Realimentadas (*Feedback Networks*).
c) Redes Celulares o en Topología de Malla.





Dr. Fco. Javier Martínez de Pisón Ascacíbar (fjmartin@dim.unirioja.es)
Grupo EDMANS (Engineering Data Mining And Numerical Simulation)
http://www.mineriadatos.com

TÉCNICAS DE MODELADO PREDICTIVO:
Redes Neuronales



TÉCNICAS DE MODELADO PREDICTIVO:
Redes Neuronales

Las Redes Neuronales son Excelentes para obtener Modelos No Lineales de Buena Precisión, por eso SON MUY UTILIZADOS EN LA OPTIMIZACIÓN DE PROCESOS INDUSTRIALES HABITUALMENTE NO LINEALES. El campo de aplicación de las Redes Neuronales es Enorme, no solo para modelado sino para otras técnicas de minería de datos (agrupamiento, proyectores, filtrado, etc.)

Cómo Desventajas Principales:

1. Necesitan mucha información para entrenarlas.
2. Hay que tener experiencia y cuidado a la hora de entrenarlas. Se necesita tiempo y potencia de cálculo. Hay que seleccionar un número adecuado de capas y neuronas para no entrar en el sobreajuste.
3. Son cajas negras. No se pueden extraer de ellas fácilmente las relaciones entre variables (aunque existen algunas técnicas).
4. No son muy robustas frente a espurios (aunque hay algunas redes neuronales robustas).

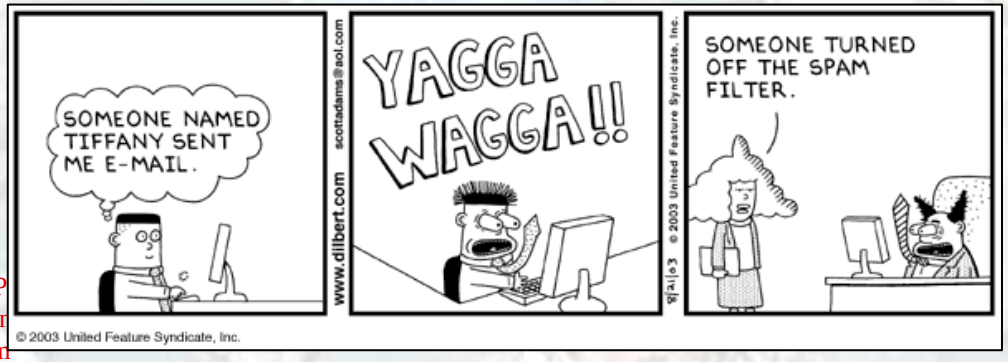
TÉCNICAS DE MODELADO PREDICTIVO:
Técnicas Bayesianas

Se basan en teorías de probabilidad (Teorema de Bayes) para realizar inferencias a partir de los datos induciendo modelos probabilísticos y cuantificando la incertidumbre ante nuevos casos.

Permite realizar tareas:

- Descriptivas: para descubrir relaciones de independencia y/o relevancia entre variables.
- Predictivas: mediante el uso de Redes Bayesianas.

Hoy en día, el campo de aplicaciones es muy alto ya que **son muy robustas en situaciones con alta incertidumbre**. Se está utilizando por ejemplo: clasificadores de conocimiento médico, lucha antiterrorista y policial, filtros anti-spam, etc.



TÉCNICAS DE MODELADO PREDICTIVO:
Técnicas Bayesianas

Dada una hipótesis H y una evidencia E que conduce a esa hipótesis, entonces podemos decir según la regla de Bayes que:

$$\Pr[H \mid E] = \frac{\Pr[E \mid H]\Pr[H]}{\Pr[E]}$$

donde Pr[A] corresponde con la probabilidad de que suceda el suceso A y Pr[A|B] denota la probabilidad del suceso A condicionado a que suceda B.

Si tenemos E1, E2, ..., En evidencias independientes de un suceso H, la probabilidad de que suceda frente a una nueva evidencia E se generaliza como:

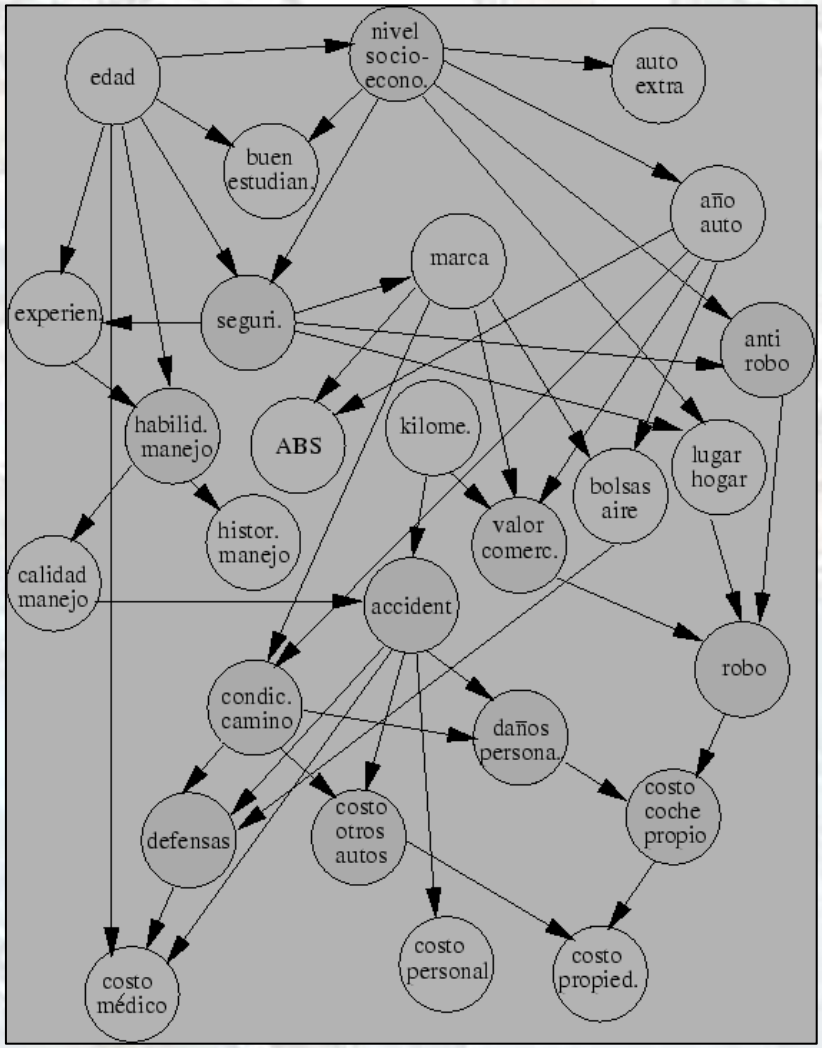
$$\Pr[H \mid E] = \frac{\Pr[E_1 \mid H]\Pr[E_2 \mid H]..\Pr[E_n \mid H]\Pr[H]}{\Pr[E]}$$

Que **nos permite modelizar cada clase, a partir de las probabilidades que existen de que se produzca para cada uno de las evidencias que se conocen**. Este método se denomina de Naive Bayes y se puede utilizar siempre que los eventos sean independientes.

TÉCNICAS DE MODELADO PREDICTIVO:
Técnicas Bayesianas: Redes Bayesianas

Una red bayesiana, una vez construida, constituye un potente dispositivo para el razonamiento probabilístico.

Aunque, muchas veces, el experto es el que construye la red según sus conocimientos, existen diversas técnicas y herramientas que permiten generar automáticamente una red bayesiana a partir de la base de datos que se le suministre.



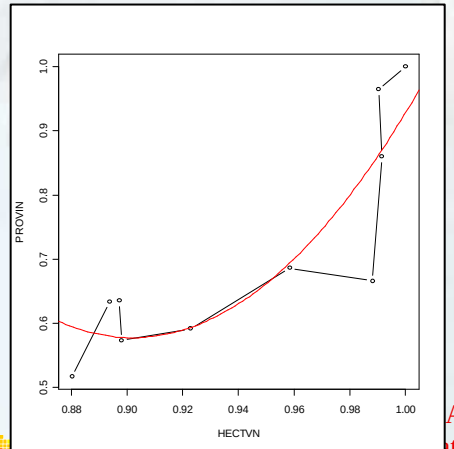
Dr. Fco. Javier Martínez de Pisón Ascacíbar (fjmartin@dim.unirioja.es)
Grupo EDMANS (Engineering Data Mining And Numerical Simulation)
http://www.mineriadatos.com

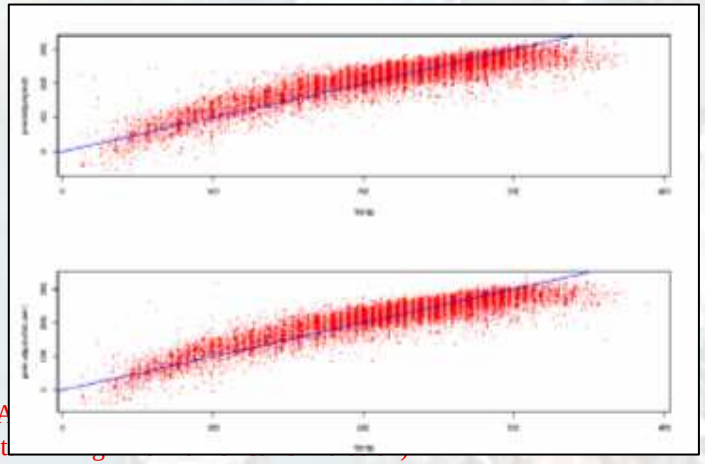
TÉCNICAS DE MODELADO PREDICTIVO:
Métodos Estadísticos

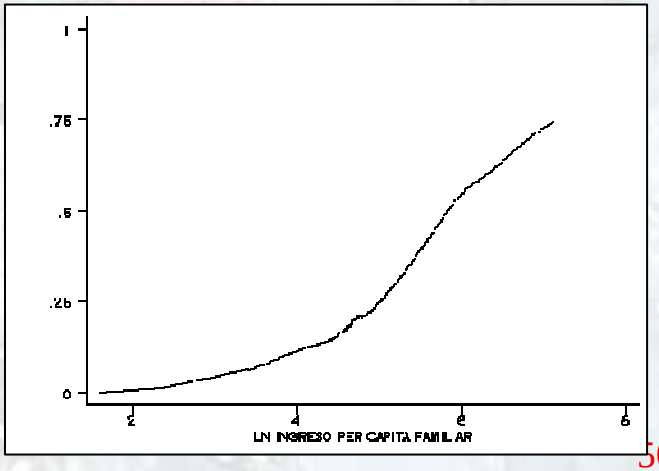
Existen Gran Cantidad de Técnicas de Modelado Estadístico:

Modelos Paramétricos que siguen un cierto modelo matemático y dependen de un número finito de parámetros: Regresión Lineal, Modelos Lineales Generalizados, Regresión sobre Componentes Incorrelados (PCA, PLS), Análisis Discriminante, etc. Son difíciles de aplicar en casos complejos no lineales.

Modelos No Paramétricos que No Siguen un cierto modelo matemático predeterminado: Métodos de Ajuste Local, Modelos Aditivos, Discriminación No Paramétrica, etc. Son más adecuados a casos complejos y multivariantes.







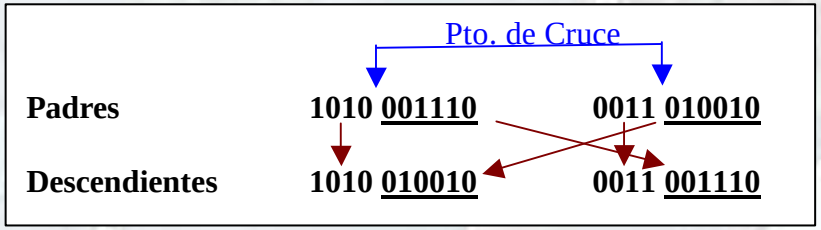
TÉCNICAS DE MODELADO PREDICTIVO:
Programación Evolutiva

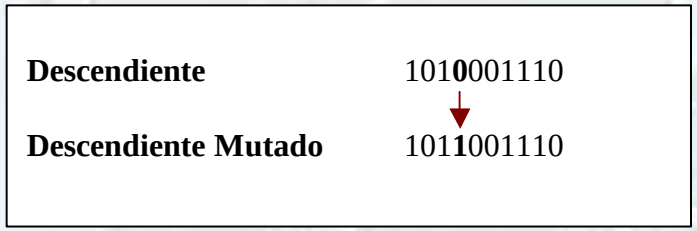
Las estrategias de computación evolutiva suponen un enfoque alternativo para **abordar problemas complejos de búsqueda y aprendizaje a través de modelos computacionales de procesos evolutivos**. Las implantaciones concretas de tales estrategias se conocen como algoritmos evolutivos [DAV91] [HOL92][OPE01].

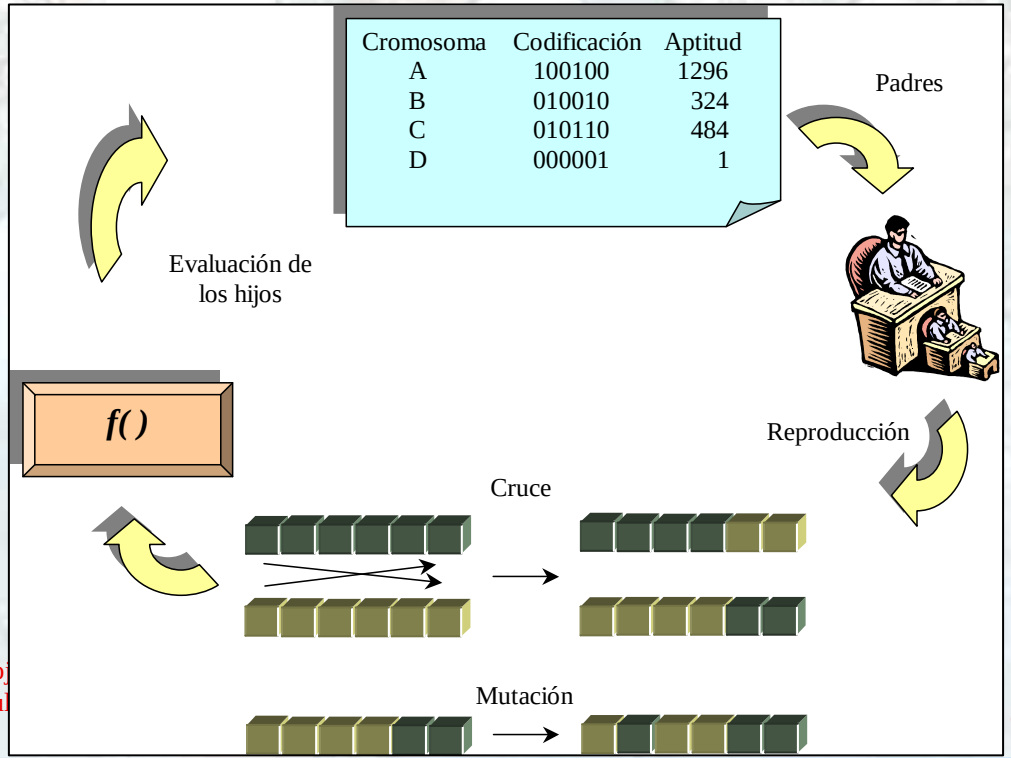
Consiste en el uso de mecanismos de selección de soluciones potenciales y de construcción de nuevos candidatos por recombinación de características de otros ya presentes, de modo parecido a como ocurre en la evolución de los organismos naturales adaptados para la supervivencia en casi cualquier ecosistema.

TÉCNICAS DE MODELADO PREDICTIVO:
Programación Evolutiva

Un caso típico son los Algoritmos Genéticos: Se codifican en cromosomas. Cada generación está compuesta por los mejores de la anterior generación (padres) más otros (descendientes) que provienen del cruzamiento de los padres y más los que surgen por mutación (descendientes mutados). El mejor de la última generación es la solución óptima buscada.







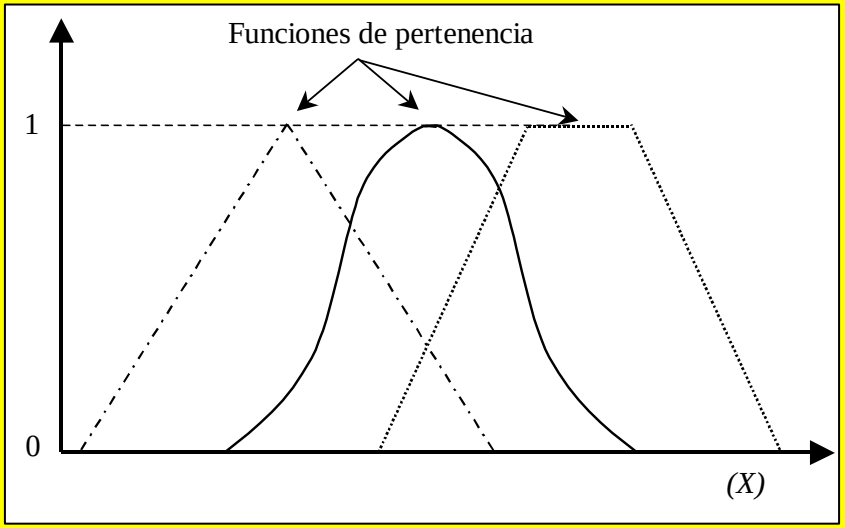
TÉCNICAS DE MODELADO PREDICTIVO:
Métodos Difusos

La lógica difusa designa un conjunto de herramientas de la lógica convencional (booleana) que ha sido extendido para incluir el concepto de verdad parcial (valores de verdad entre completamente cierto y completamente falso).

La convención utilizada para denotar los grados de pertenencia es de asignar el valor 1 al grado de pertenencia mas fuerte, el valor 0 al grado de pertenencia mas débil y valores reales en el interior del intervalo [0,1] a grados de pertenencia intermedios.

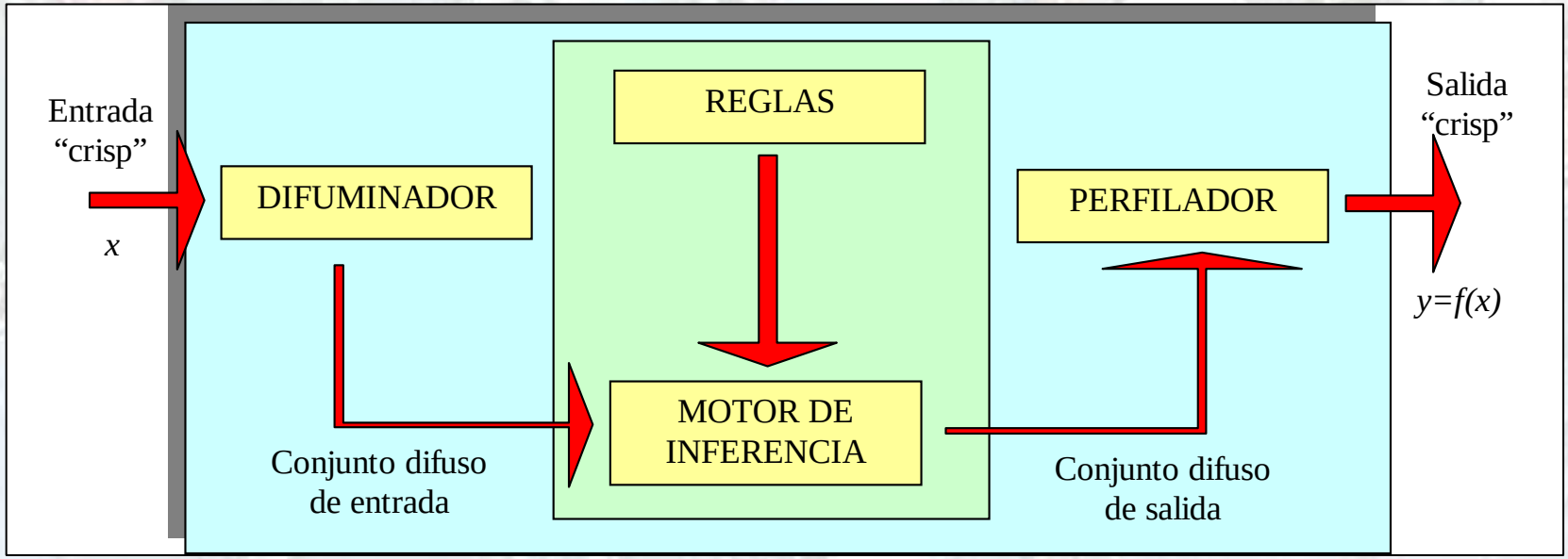
El rol que juega el intervalo unitario en la asignación de grados de pertenencia es solo homegeneizador y puede ser reemplazado por cualquier otro conjunto ordenado.

Sin embargo, esta convención permite observar a la lógica difusa como una extensión natural de la binaria.



TÉCNICAS DE MODELADO PREDICTIVO:
Métodos Difusos

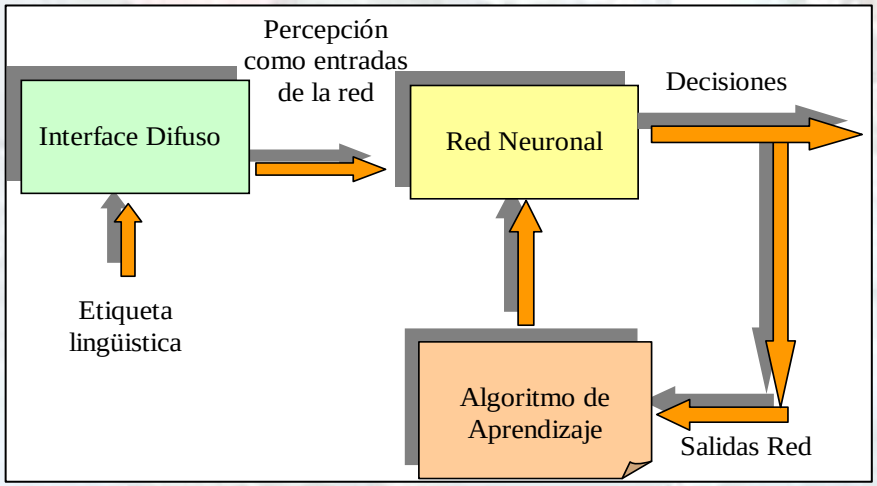
Un sistema de inferencia difusa **es aquel que usa un conjunto de funciones de pertenencia y reglas difusas para dar razón de un grupo de datos**. A los sistemas de inferencia difusa. se les conoce también con los nombres de sistemas basado en reglas difusas, modelos difusos, controladores lógicos difusos o simplemente sistemas difusos.

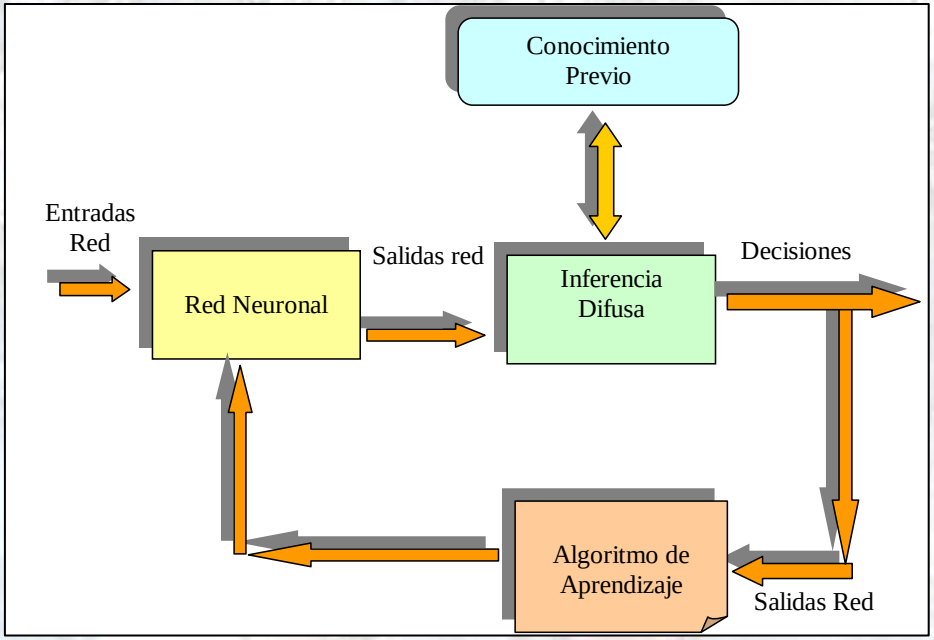


Dr. Fco. Javier Martínez de Pisón Ascacíbar (fjmartin@dim.unirioja.es)
Grupo EDMANS (Engineering Data Mining And Numerical Simulation)
http://www.mineriadatos.com

TÉCNICAS DE MODELADO PREDICTIVO:
Métodos NeuroDifusos

Son modelos que combinan las técnicas Difusas con las Redes Neuronales.





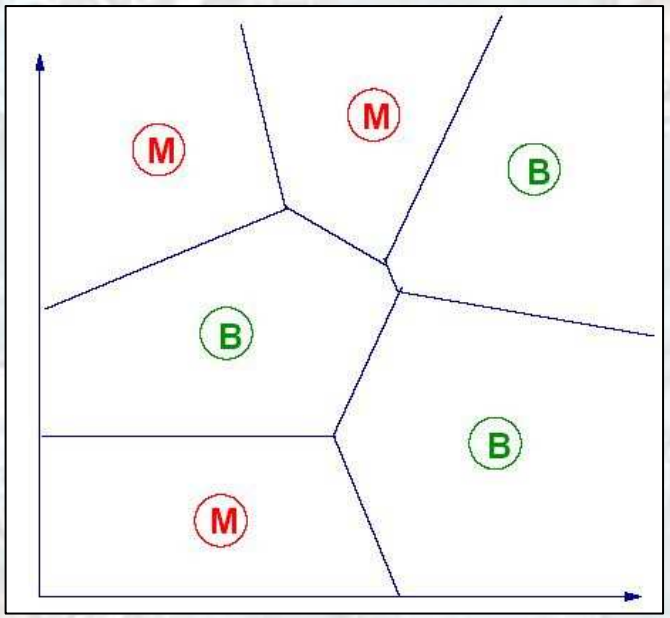
Dr. Fco. Javier Martínez de Pisón Ascacíbar (fjmartin@dim.unirioja.es)
Grupo EDMANS (Engineering Data Mining And Numerical Simulation)
http://www.mineriadatos.com

TÉCNICAS DE MODELADO PREDICTIVO:
Basados en Casos y Vecindad

En este tipo de aprendizaje [AHA92][WIT00], se almacenan los ejemplos de entrenamiento y cuando se quiere clasificar un nuevo objeto, se extraen los objetos más parecidos y se usan para clasificar al nuevo objeto.

Los ejemplos iniciales son almacenados y utilizados como "fuente de conocimiento" (Se crean grupos con K-means, SOM, Jerárquicos).

De esta forma, cuando aparecen nuevos ejemplos, se intentan clasificar mediante alguna medida de distancias o similar (K-Vecinos, LVQ, etc.), y si no se puede asignar a ninguno de los ya existentes, se almacena como un ejemplo nuevo.

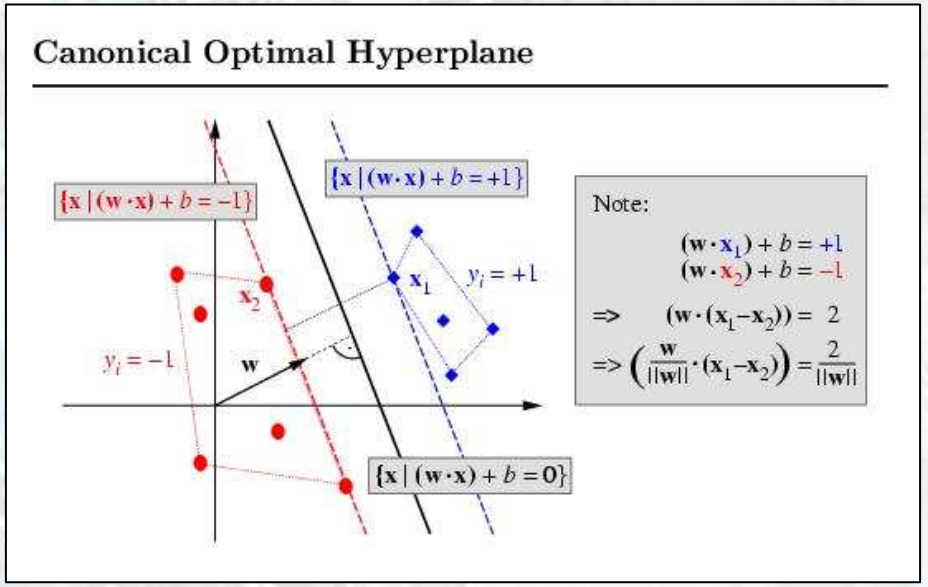


Dr. Fco. Javier Martínez de Pisón Ascacíbar (fjmartin@dim.unirioja.es)
Grupo EDMANS (Engineering Data Mining And Numerical Simulation)
http://www.mineriadatos.com

TÉCNICAS DE MODELADO PREDICTIVO:
Máquinas de Vectores Soporte

Los métodos basados en Máquinas de Vectores Soporte o más comúnmente llamados Support Vector Machines (SVM) Inducen Separadores Lineales o Hiperplanos en Espacios de Características de Muy Alta Dimensionalidad (introducidos por funciones núcleo o funciones kernel) con un sesgo inductivo muy particular (maximización del margen) [HER04][CRI00][WIT00][PRU02].

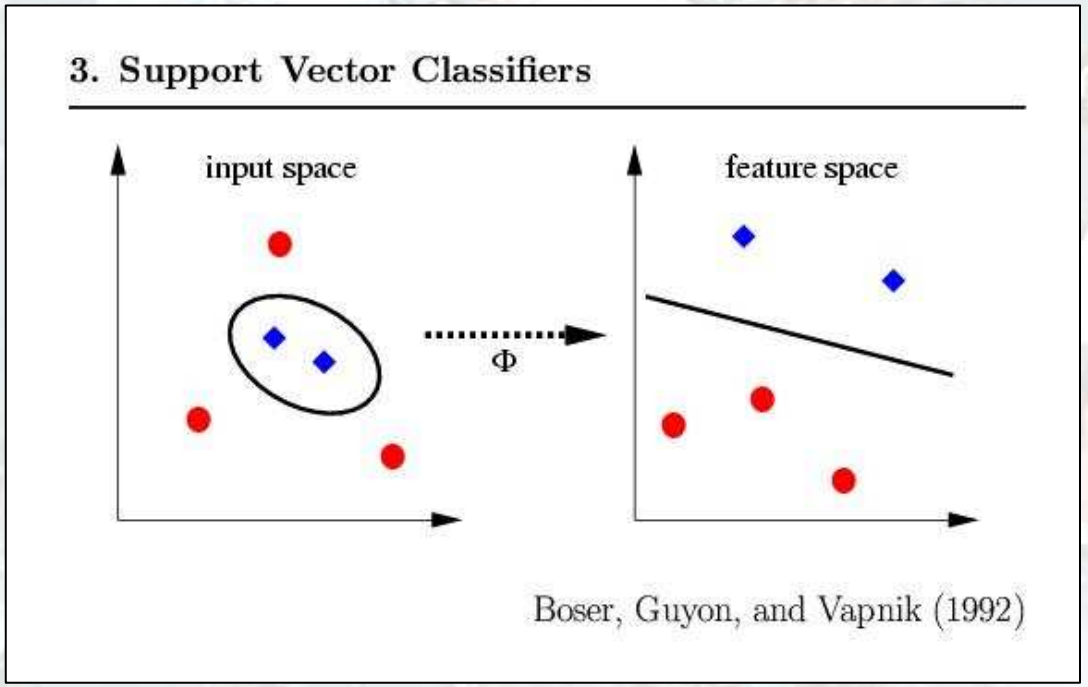
Si el conjunto de ejemplos es linealmente separable, la idea del SVM **es seleccionar el mejor hiperplano separador que está a la misma distancia de los ejemplos más cercanos de cada clase.** Dichos puntos son los denominados Vectores Soporte.



Dr. Fco. Javier Martínez de Pisón Ascacíbar (fjmartin@dim.unirioja.es)
Grupo EDMANS (Engineering Data Mining And Numerical Simulation)
http://www.mineriadatos.com

TÉCNICAS DE MODELADO PREDICTIVO:
Máquinas de Vectores Soporte

El aprendizaje de separadores no lineales con SVM se consigue mediante una transformación no lineal del espacio de atributos (input space) en un espacio de características (feature space) de dimensionalidad mucho mayor donde si es posible separar linealmente los ejemplos [HER04].



Dr. Fco. Javier Martínez de Pisón Ascacíbar (fjmartin@dim.unirioja.es)
Grupo EDMANS (Engineering Data Mining And Numerical Simulation)
http://www.mineriadatos.com