Métodos de Minería de Datos

ALGORITMOS SUPERVISADOS

JAIME MIRANDA

Departamento de Ingeniería Industrial

Universidad de Chile

CONCEPTOS BÁSICOS

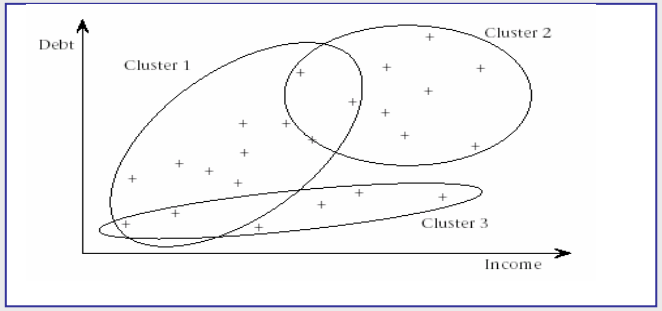
→ Medida de distancia

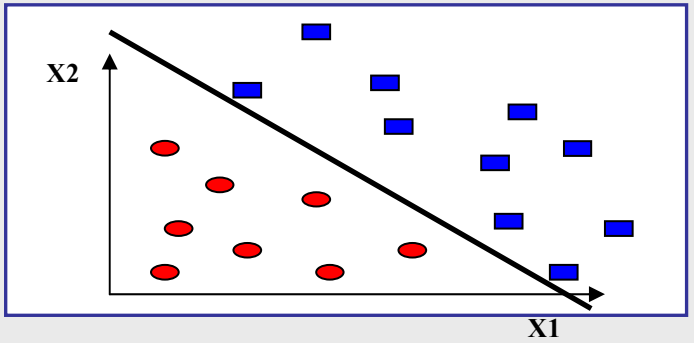
→ Prototipo o centro de clase más cercana

→ Entre más cerca mayor pertenencia de a la clase

→ Hipersuperficies

→ Clasificación de acuerdo a si los objetos están a uno u otro lado de una hipersuperficie o conjunto de hiperplanos





Regla de Bayes

$$p(C_K / X_L) = \frac{p(X_L / C_K) * p(C_K)}{p(X_L)}$$

Donde:

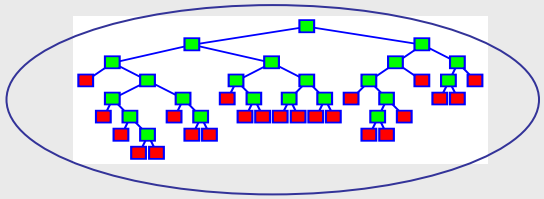
C_k denota la clase k del atributo elegido como base.

X_L son los atributos a los cuales se condicionara en probabilidad.

TÉCNICAS EMPLEADAS

MÉTODOS SUPERVISADOS

→ Redes neuronales
→ Árboles de decisión
→ SVM



MÉTODOS NO SUPERVISADOS

→ Fuzzy C-means (Cluster)
→ Mapas Kohonen

NUEVAS TÉCNICAS

→ MCS (Multiclassifier Systems)
→ Clamping (Heurística selección de atributos)

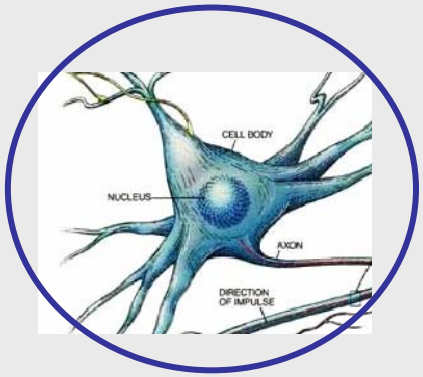


DIAGRAMA DE APRENDIZAJE SUPERVISADO

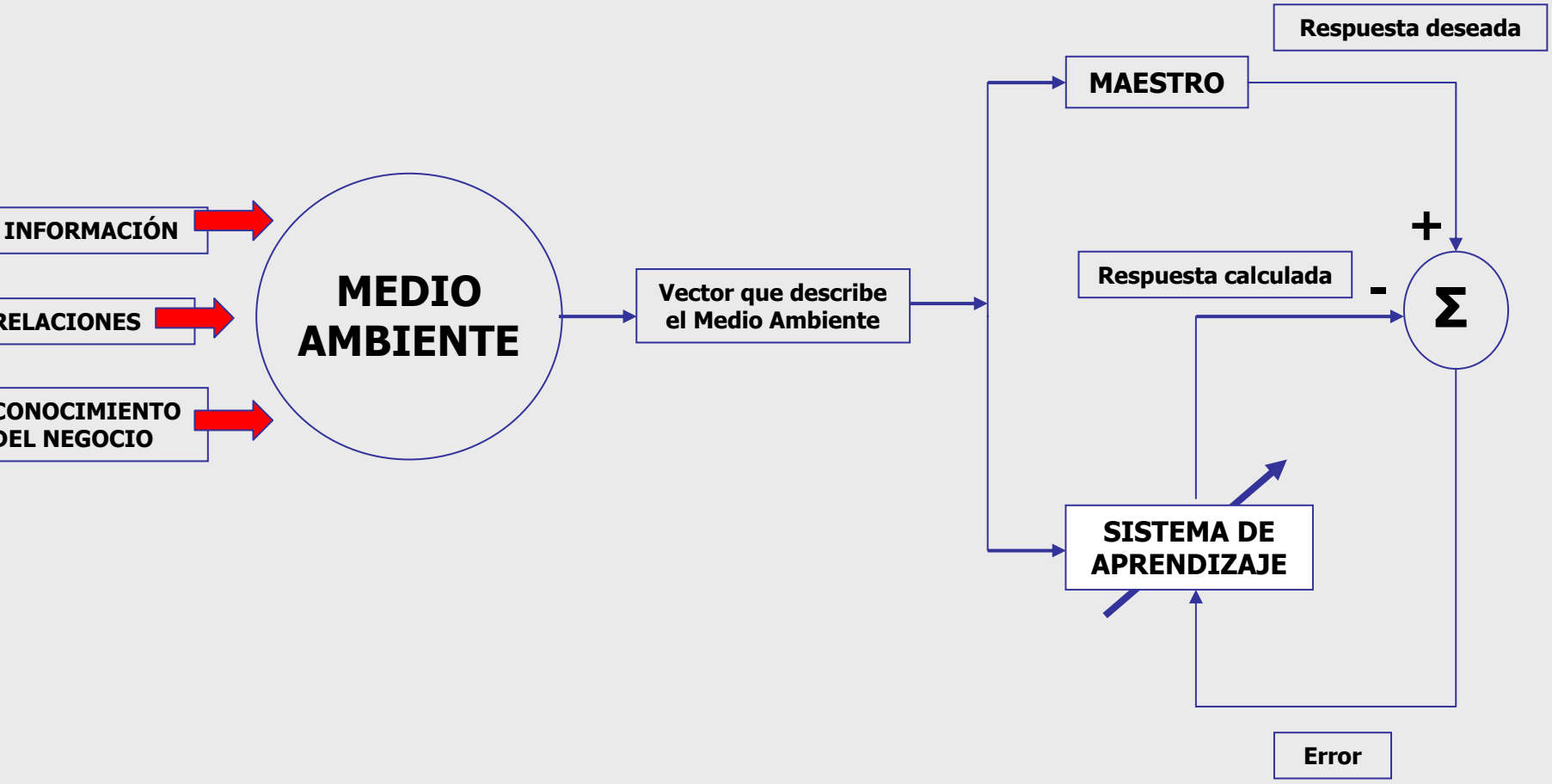
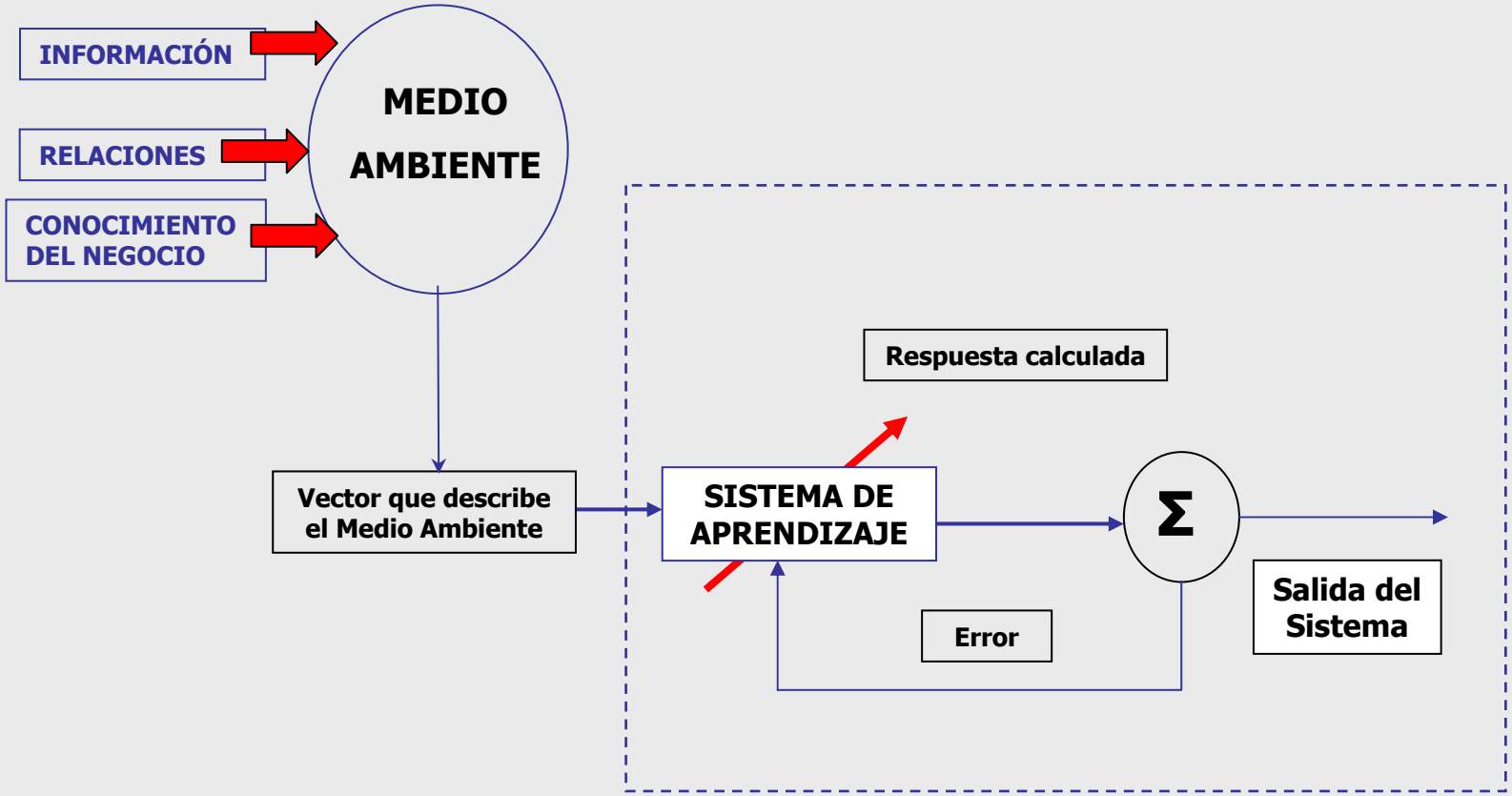


DIAGRAMA DE APRENDIZAJE NO SUPERVISADO



Sistema no supervisado

MODELOS SUPERVISADOS

REDES NEURONALES MLP

JAIME MIRANDA

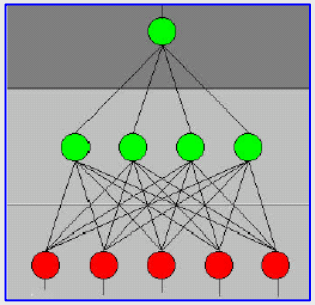
Departamento de Ingeniería Industrial

Universidad de Chile

ALGUNOS CASOS EXITOSOS DMGROUP

Scoring de riesgo crediticio para mutuos hipotecarios

→ Técnica de minería usada : Red neuronal multicapa



Modelo predictivo de fugas de cuentacorrentistas

→ Técnica de minería usada: SVM

Modelo predictivo de fugas de cuentacorrentistas

→ Técnica de minería usada: Red neuronal multicapa especializada

Modelo predictivo de ofertas focalizadas

→ Técnica de minería usada: Red neuronal multicapa – Árbol de decisión

Modelo detección de fraudes en transacciones electrónicas de fondos

→ Técnica de minería usada: MCS

- SVM
- Árboles de decisión
- Mapas de Kohonen

CASO DE ESTUDIO

PROBLEMA: ALTO NUMERO DE FRAUDES EN TARJETAS DE CRÉDITO

→ Construir un modelo de predicción de fraudes en tarjetas de crédito

→ La empresa posee un Datawarehouse con la información transaccional y de caracterización (perfil) de los clientes

→ La base consta de 2.000 clientes descritos por 15 variables

→ Se desea encontrar un patrón característico que describa el comportamiento de fraude

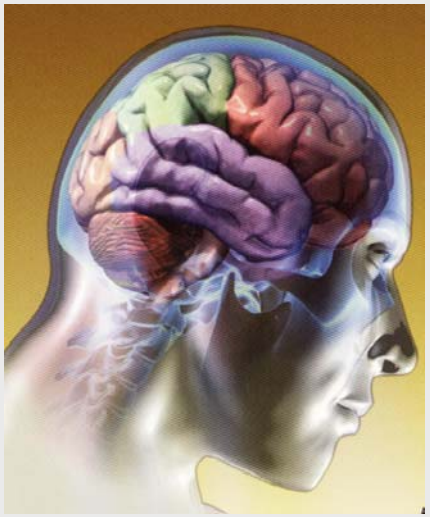
ALGUNAS DEFINICIONES …

APRENDIZAJE

→ “El aprendizaje es una habilidad de la que disponen gran parte de los sistemas naturales para adaptarse al entorno en el que vive”.

→ “Adquisición de conocimiento de un proceso por medio del análisis, ejercicio o experiencia“.

→ “Un proceso por el cual los parámetros libres del sistema se adaptan a través de un proceso continuo de estimulación a partir del entorno en el que el sistema está inmerso”.



ENFOQUE BAYESIANO

$$p(C_K / X_L) = \frac{p(X_L / C_K) * p(C_K)}{p(X_L)}$$

DONDE:

$$\sum_i P(C_i) = 1$$

$$\sum_k P(X_k) = 1$$

Ck **denota la clase k del atributo elegido como base.**

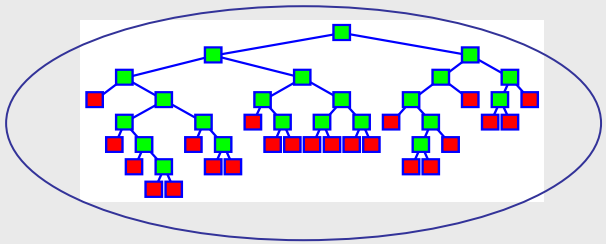
XL **son los atributos a los cuales se condicionara en probabilidad**

REGLA DE DECISIÓN DE BAYES

X_k pertenece a la clase C_j si y solo si:

$$g(x_k) > g(x_k)$$

$$P(C_j / x_k) > P(C_i / x_k)$$



CLASIFICACIÓN LINEAL

$$f: X \in \mathbb{R}^n \rightarrow \mathbb{R}$$

$$X = (x_1, x_2, \ldots\ldots, x_n)$$

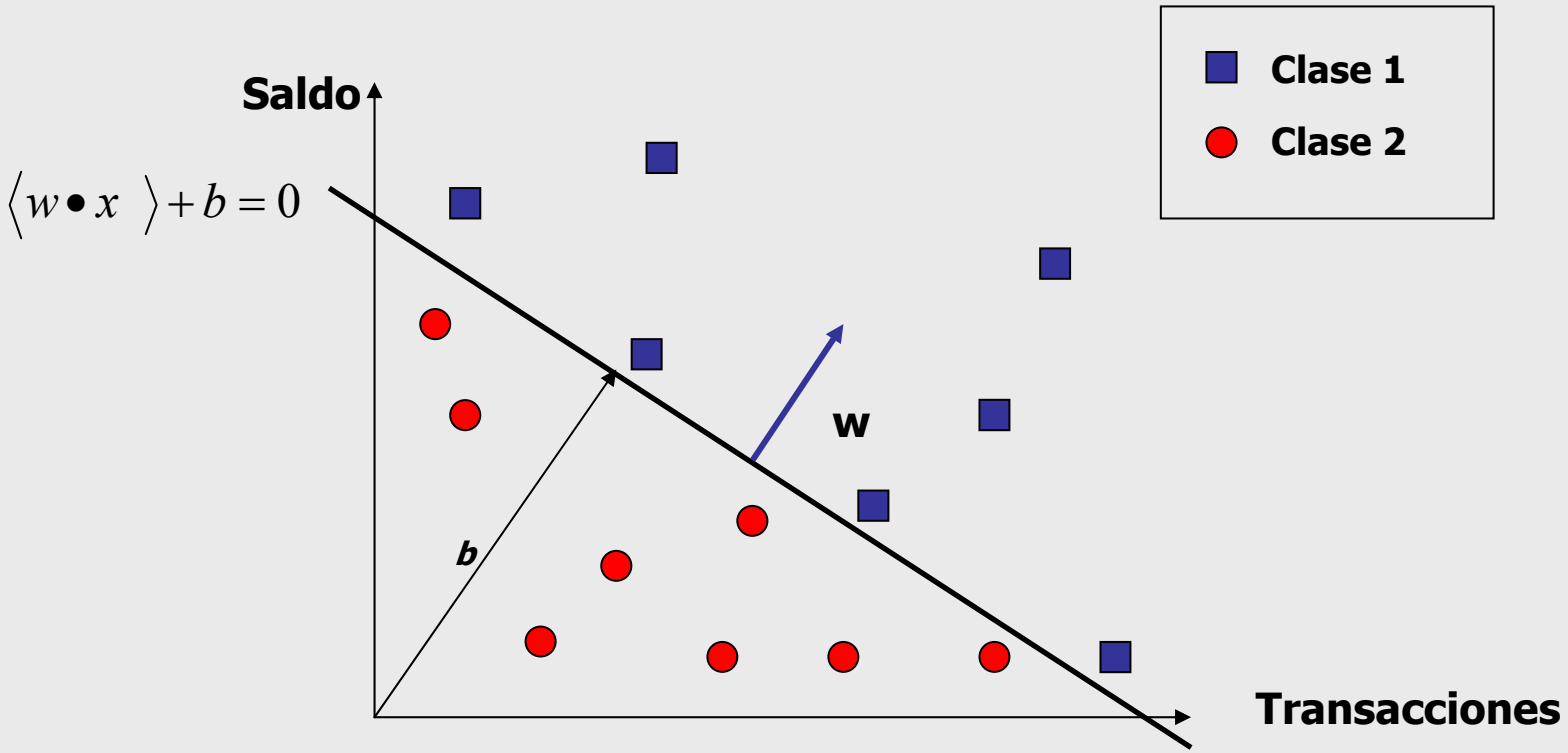
EJEMPLO: Clasificación binaria

$f(X) \geq 0$ ➡ **Clase 1**

$f(X) \leq 0$ ➡ **Clase 2**

CLASIFICACIÓN LINEAL (2)

$$f(X) = \langle w \bullet x \rangle + b$$



ALGUNAS NOCIONES

→ Propuesto en 1956 por Frank Rosenblatt.

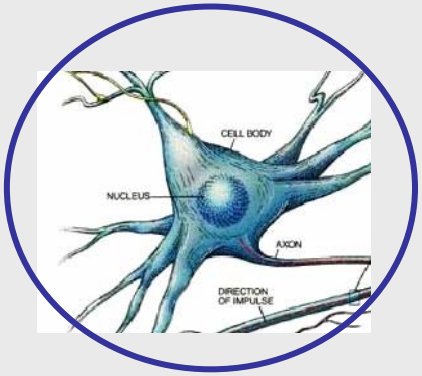
→ Fue objeto de gran interés a comienzos de los 60´s.

→ Primer algoritmo iterativo para clasificación lineal.

→ Este algoritmo garantiza encontrar un hiperplano separador de clases, para datos linealmente separables.

→ Unidad básica de la arquitectura de las redes neuronales.

→ Se basa en una representación neuronal biológica.



MODELO PERCEPTRÓN (2)

MODELO

→ Una neurona con pesos sinápticos y nivel umbral ajustable.

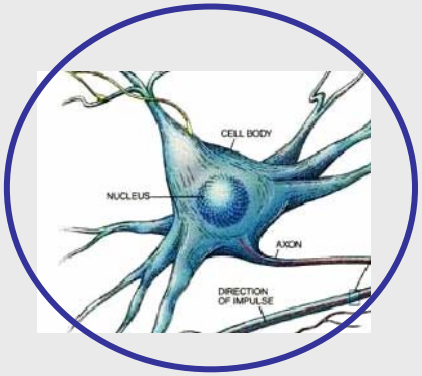
→ Neurona biológica.

PROPÓSITO

→ Clasificar estímulos externos de un objeto respecto a una clase.

APRENDIZAJE

→ Determinar el vector de pesos óptimo (w) que clasifique bien a cada objeto.



ELEMENTOS BÁSICOS

UNIDADES DE ENTRADA

→ Elementos que estimulan la red.

→ Atributos o variables de entrada.

CONEXIONES

→ Por las que se propaga la señal que conforma el patrón.

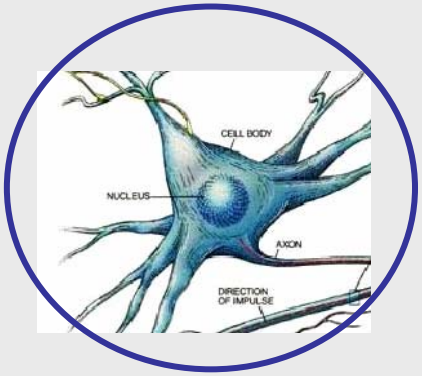
→ Pesos wi indican que tan fuerte es el atributo.

COMBINADOR LINEAL

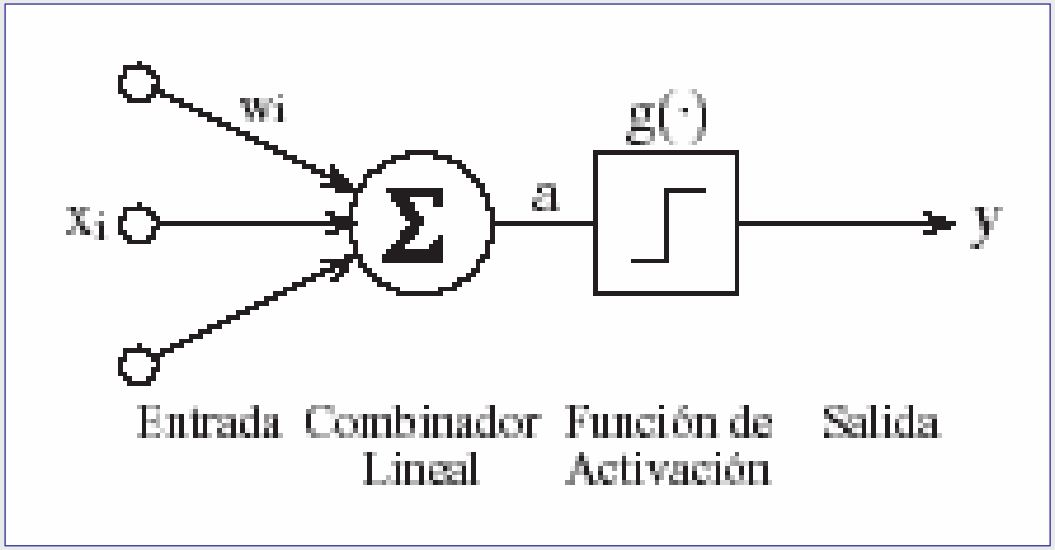
→ Capta las señales y las combina en una sola.

FUNCIÓN DE ACTIVACIÓN

→ Activa o no la señal.



MODELO GENERAL



DEFINICIONES

→ Vector de pesos

$$W = \langle w_0, w_1, \ldots\ldots w_n, \rangle$$

→ Conjunto de ejemplos

$$E = \langle (x^1, t^1), (x^2, t^2), \ldots\ldots\ldots (x^p, t^p), \rangle$$

→ Salidas del modelo

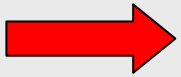
$$t^u = +1 \qquad t^u = -1$$

ALGORITMO PERCEPTRON

1. **Inicializar $W=(w_1,...w_n) = 0$.**
2. **Seleccionar un ejemplo $X^u=(x^u,t^u)$ en orden cíclico o al azar.**
3. **Si W clasifica correctamente a Xu, es decir:**

$$\langle W \bullet X^u \rangle > 0 \quad \text{y} \quad t^u = +1$$

$$\langle W \bullet X^u \rangle < 0 \quad \text{y} \quad t^u = -1$$



No tomar ninguna acción

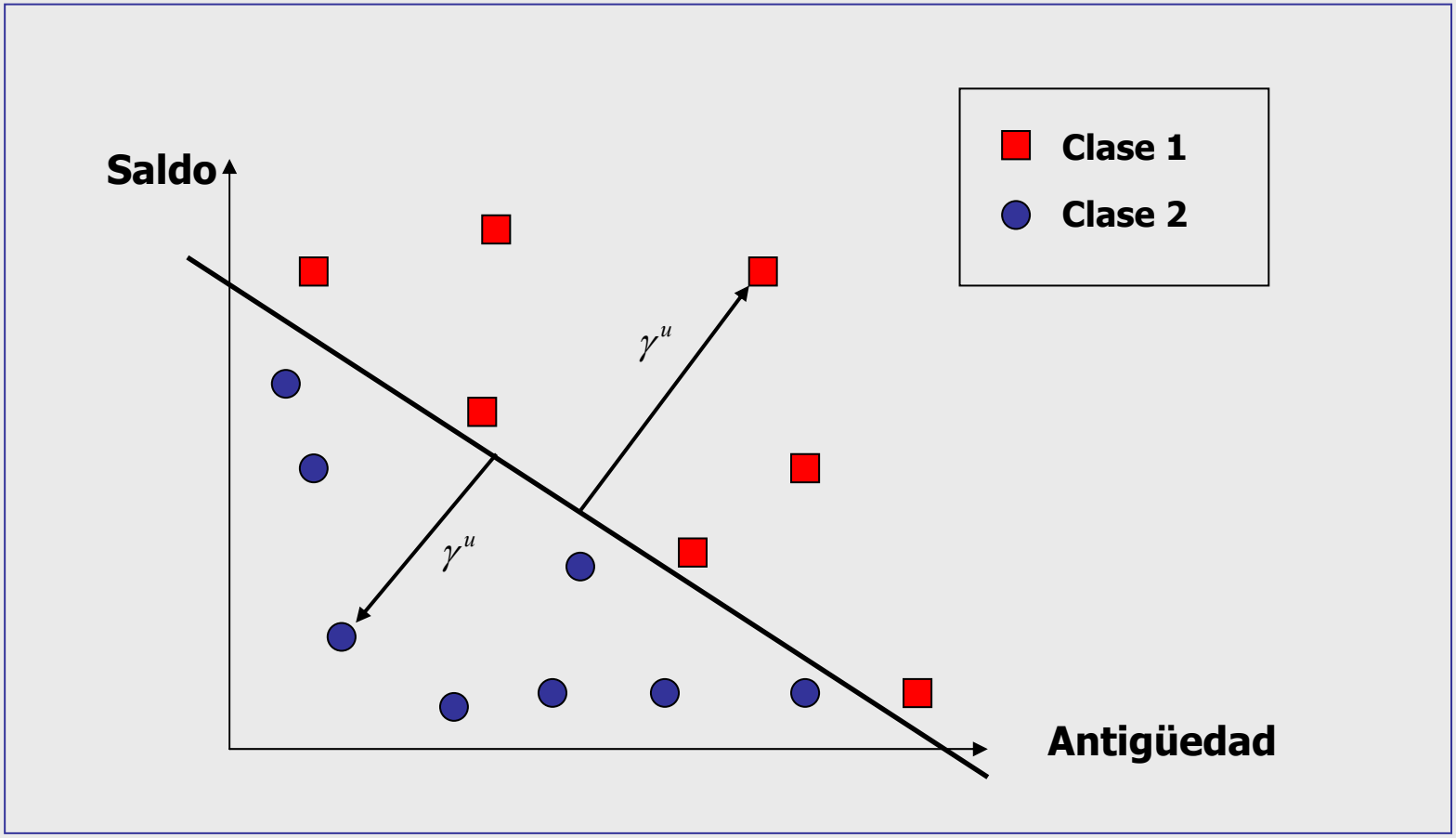
4. **Si clasifica mal:**

$$W^{"} = W + t^u x^u$$

5. **Volver a 2.**

MARGEN EN FORMA GRÁFICA

$$\gamma^u = t^u * \left\langle (w \bullet x^u + b) \right\rangle$$



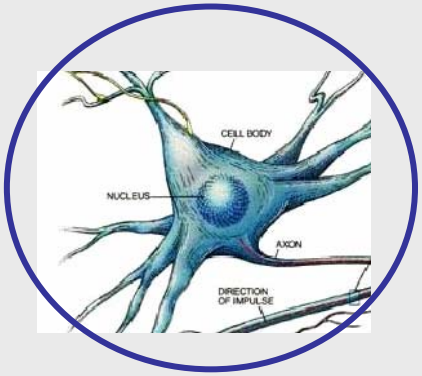
PRINCIPALES TEOREMAS PERCEPTRON

CONVERGENCIA

→ El algoritmo perceptron converge siempre en tiempo finito, para un conjunto separable y finito de ejemplos

PESOS CÍCLICOS ACOTADOS

→ El conjunto de pesos que algoritmo recorre para cualquier problema, sea éste separable o no, es acotado



PRINCIPALES TEOREMAS PERCEPTRON (2)

EN TIEMPO FINITO EL ALGORITMO PRODUCIRÁ:

→ Un vector de pesos que satisfaga todos los ejemplos

→ Conjunto linealmente separable

→ Volverá a visitar un vector de pesos

→ Conjunto no linealmente separable

Test de separabilidad lineal

PROBLEMAS:

No se conoce cota de tiempo

Costoso: almacenar pesos pasados

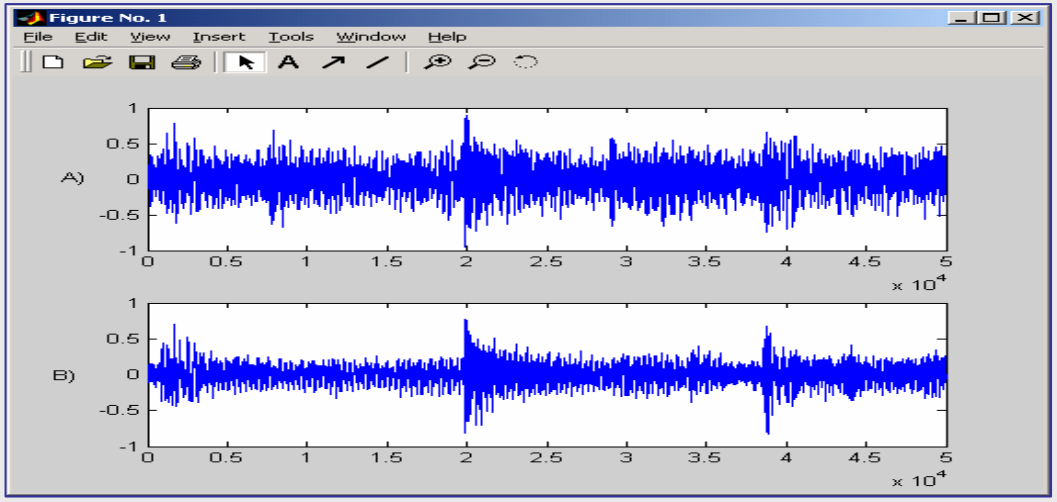
APLICACIÓN : COMBINADORES LINEALES:

CANCELADORES DE RUIDO (ADALINE)

→ Fueron introducidos por Widrow

→ Limpieza de una señal acústica que posee ruido

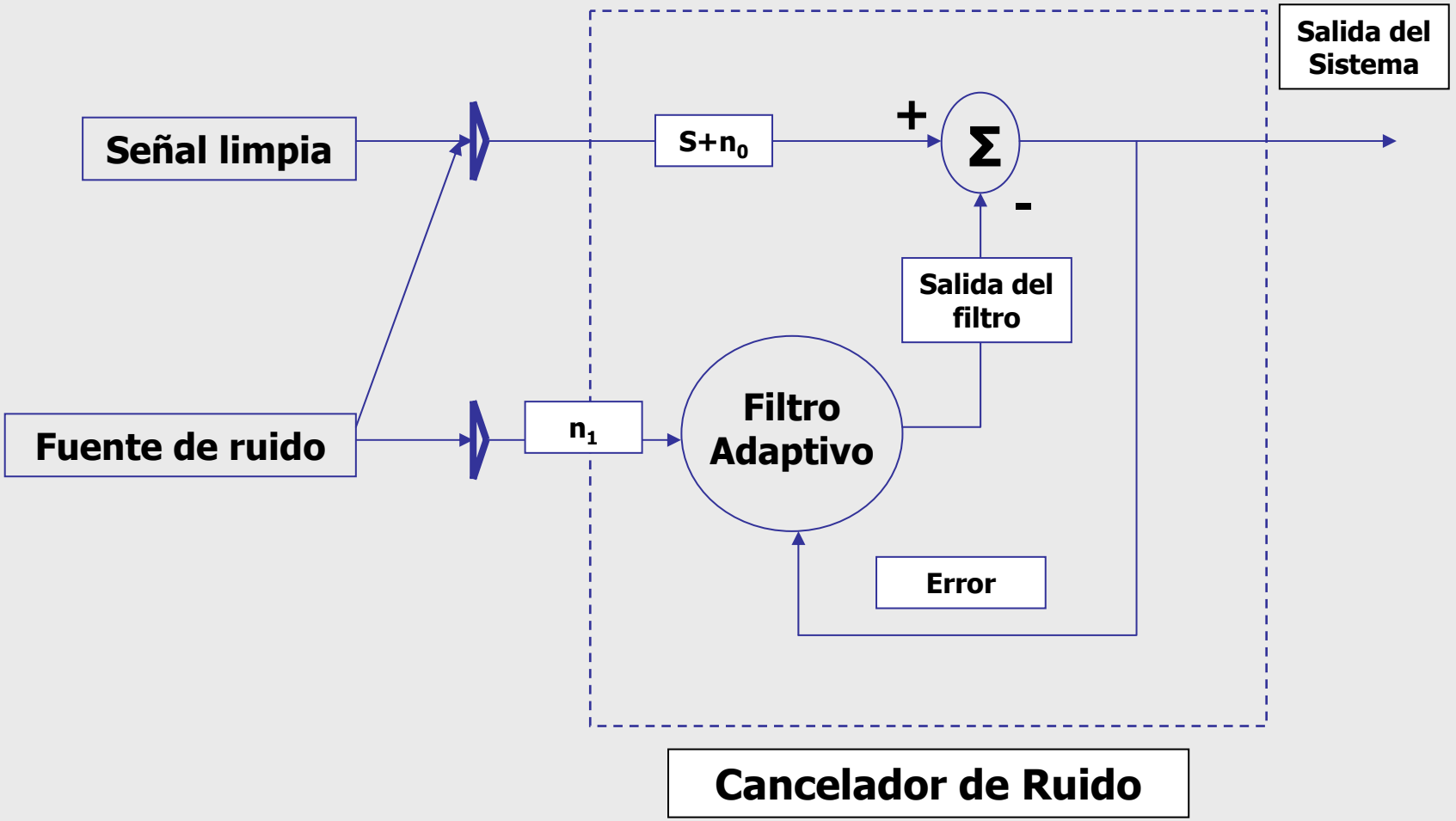
→ Uso de una estimación de una serie temporal



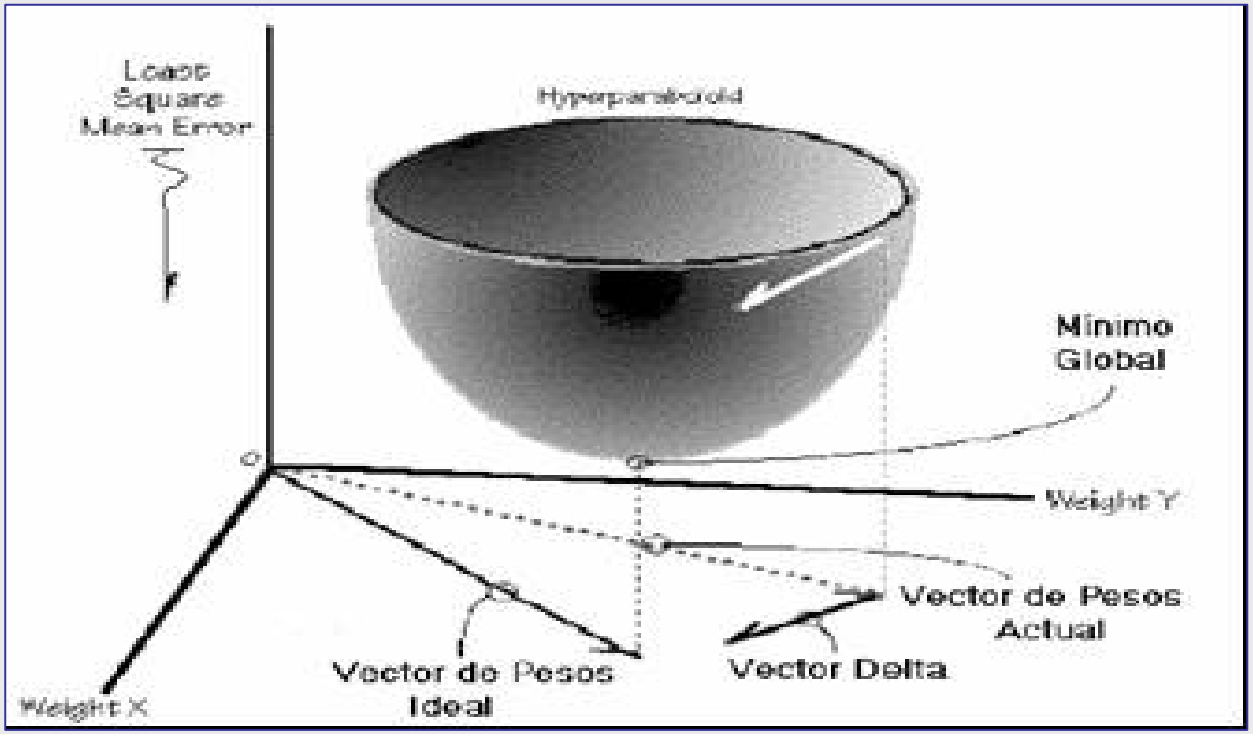
Señal sucia

Señal limpia

ESQUEMA GENERAL CANCELADOR DE RUIDO

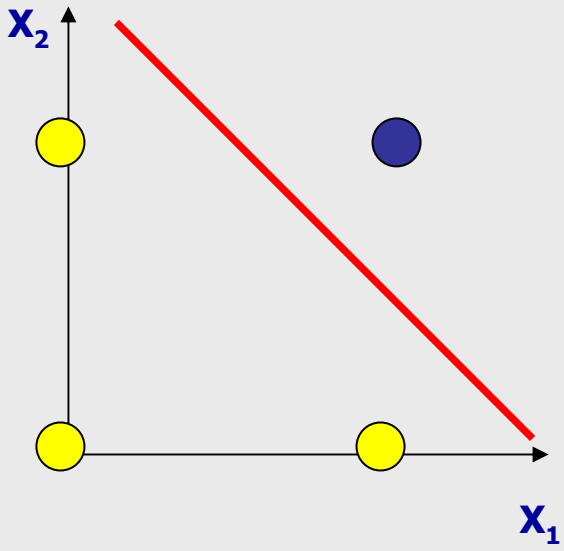


BÚSQUEDA DE PESOS ÓPTIMOS



SEPARABILIDAD EN FUNCIONES LINEALES

FUNCION BOOLEANAS		
X1	X2	AND
1	1	1
0	1	0
1	0	0
0	0	0

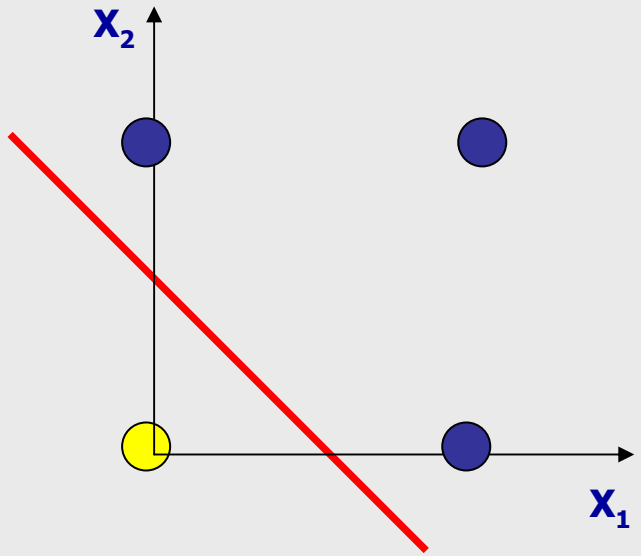


Los valores amarillos valen uno (TRUE)

Los valores azules valen cero (FALSO)

SEPARABILIDAD EN FUNCIONES LINEALES (2)

FUNCION BOOLEANAS		
X1	X2	OR
1	1	1
0	1	1
1	0	1
0	0	0

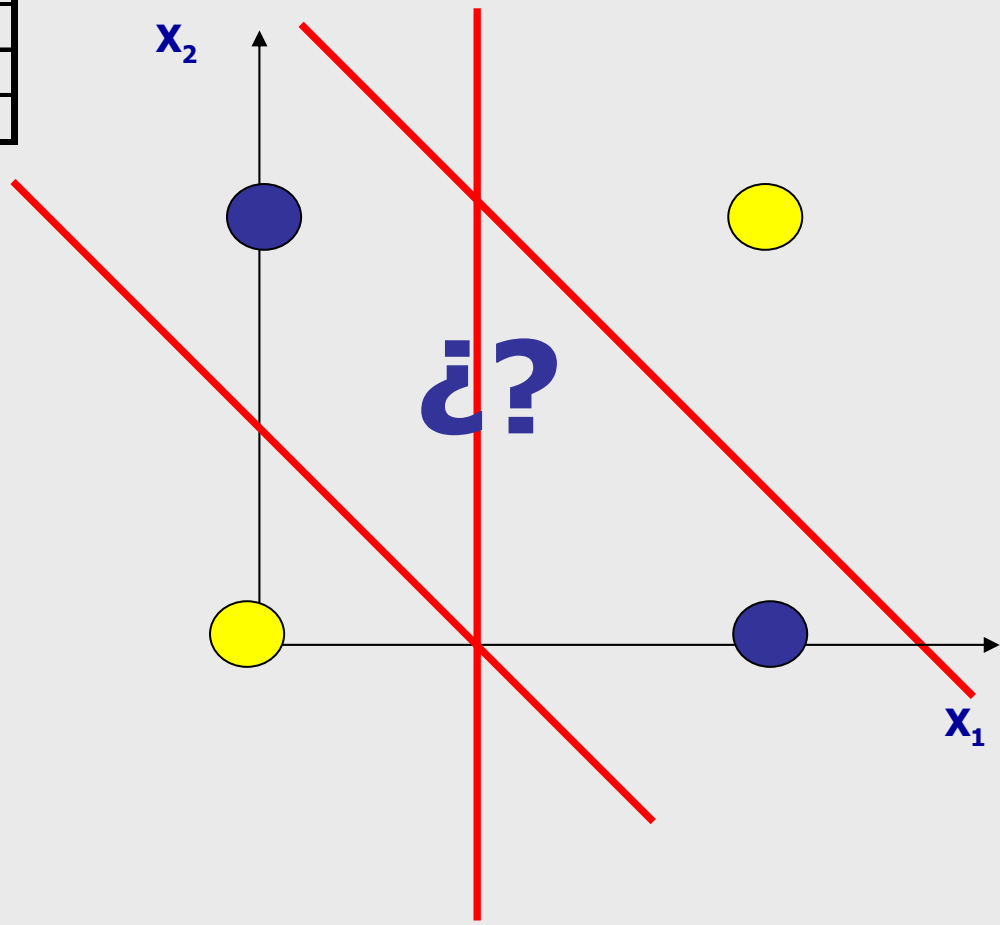


Los valores amarillos valen uno (TRUE)

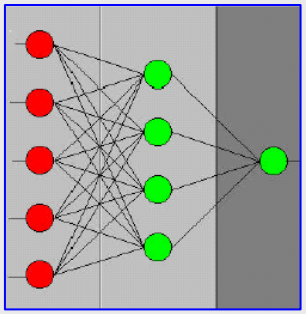
Los valores azules valen cero (FALSO)

UN PEQUEÑO PROBLEMA

FUNCION BOOLEANAS		
X1	X2	XOR
1	1	0
0	1	1
1	0	1
0	0	0



Aplicaciones en la industria



→ Retención o fuga de clientes
→ Detección de fraudes
→ Scoring

Fortalezas

→ Fuerte en lo referente a la modelación no lineal
→ Trabaja tanto con variables categóricas como continuas
→ Alta aplicabilidad (variadas áreas de estudio)

Debilidades

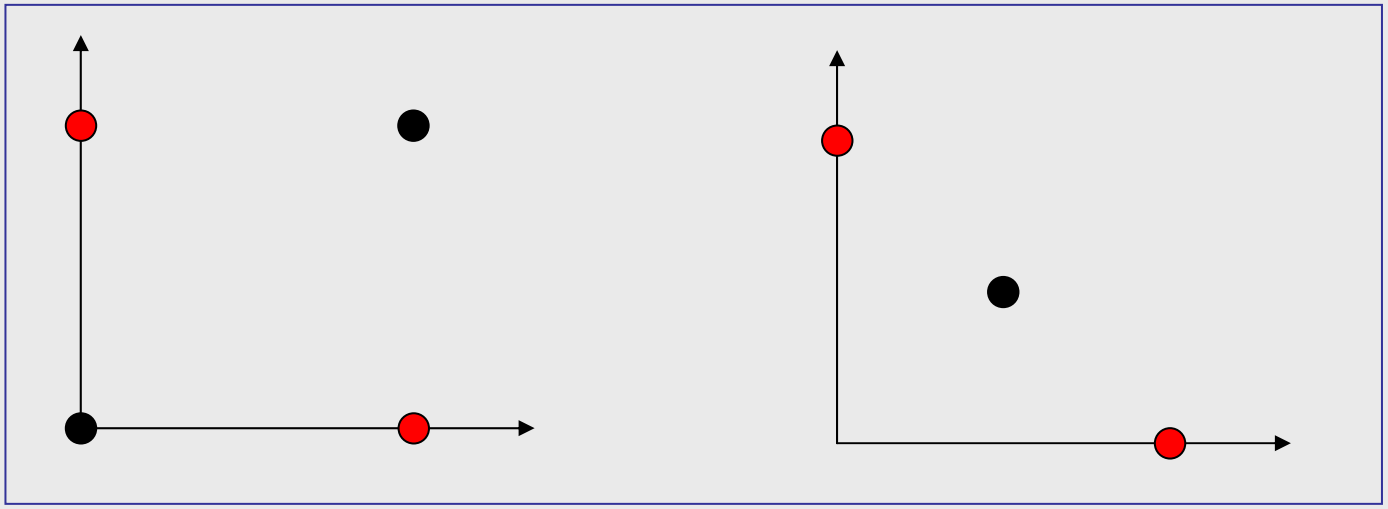
→ Difícil interpretación de las relaciones entre las variables (Heurísticas)
→ Sobreajuste

La gran mayoría de problemas no son linealmente separables

→ No es posible usar el modelo perceptron. Modelo limitado

→ No existe ningún hiperplano separador.

Se buscan funciones no lineales que definen las hipersuperficies



PRINCIPALES TEOREMAS

REPRESENTACIÓN DE FUNCIONES

→ Cualquier función booleana de un número finito de entradas puede representarse en forma exacta por un patrón multicapas

HETCH-NIELSEN

→ Cualquier función continua dentro de un cubo n-dimensional puede implementarse en forma exacta por una red con una capa oculta

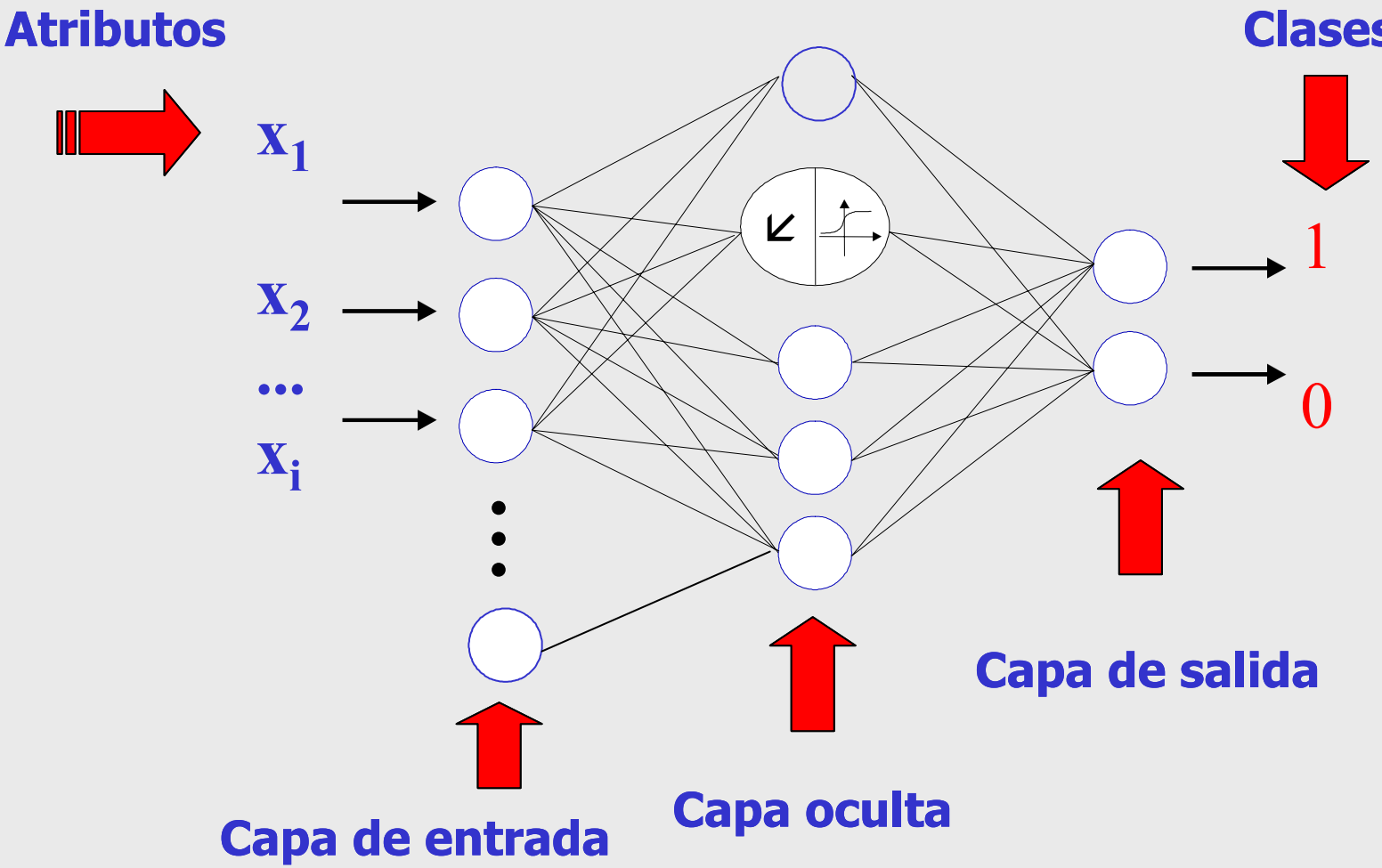
HORNIK

→ La función puede ser representada con una red multicapa, siempre y cuando tenga un número adecuado de unidades escondidas

KOLGOMOROV

→ Una capa oculta es suficiente para la aproximación de cualquier función

ARQUITECTURA RED NEURONAL MLP



ROL DE LAS CAPAS OCULTAS

Construyen las distintas dimensiones de los polígonos (fronteras).

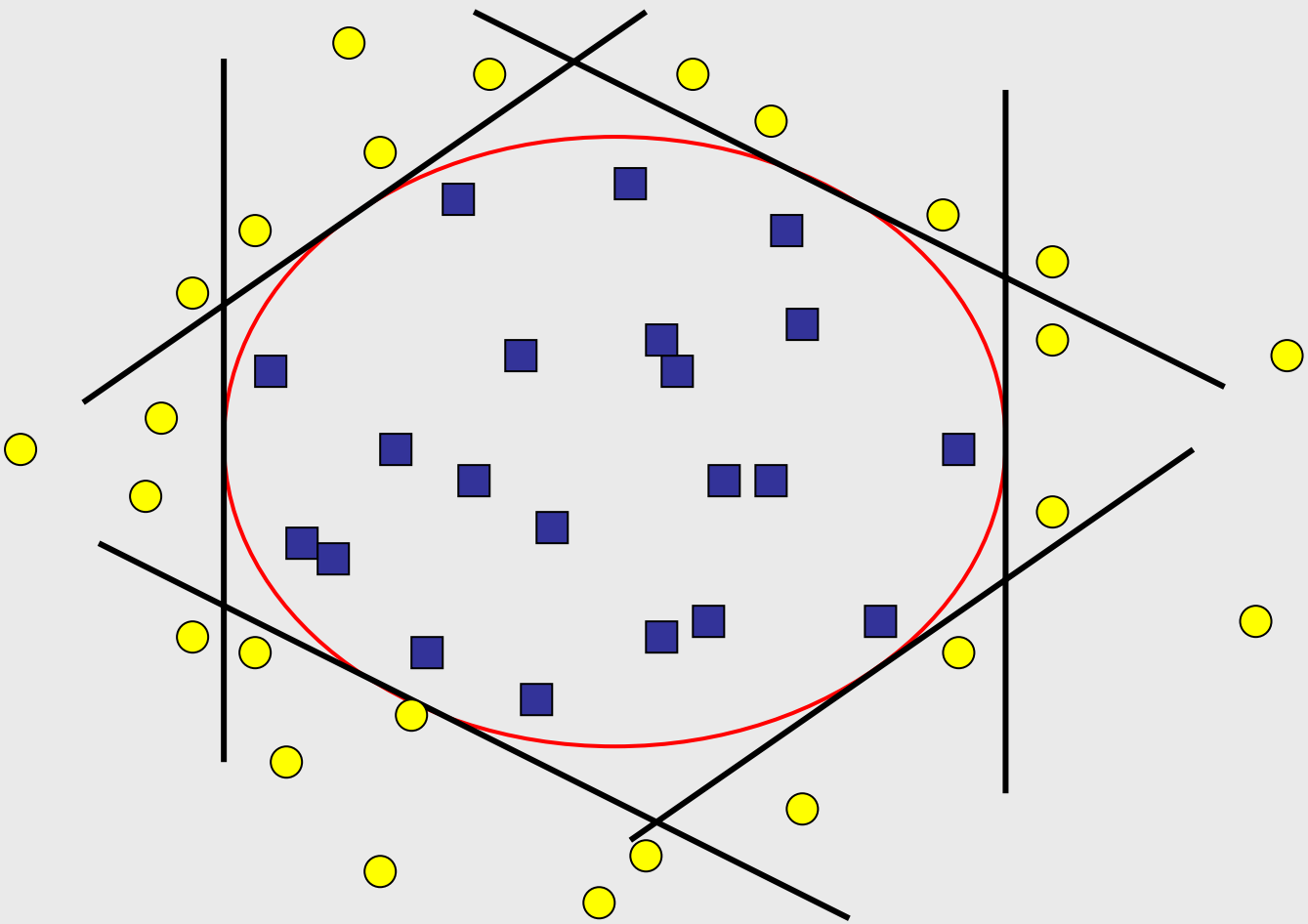
Si el polígono posee una forma muy complicada necesita más capas escondidas.

Muchas capa produce sobreajuste.

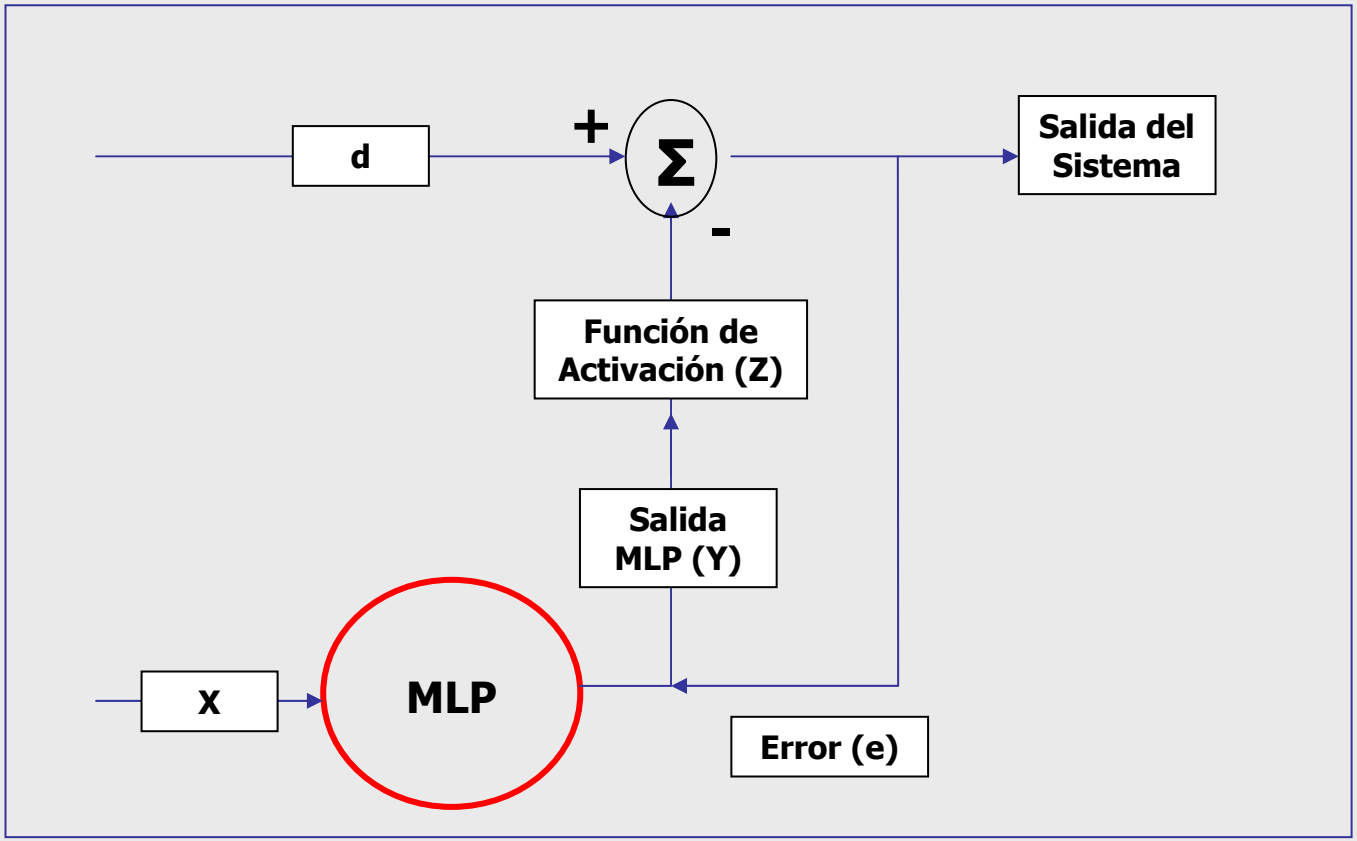
Pocas capas no puede ser modelado el problema.

"Regla Aproximada"~ Usar el promedio de las neuronas de entrada y de salida

ROL DE LAS CAPAS OCULTAS: FORMA GRAFICA



ESQUEMA GENERAL: FUNCIONES DE ACTIVACIÓN



FUNCIONES SIGMOIDES

→ Logística (0,1)

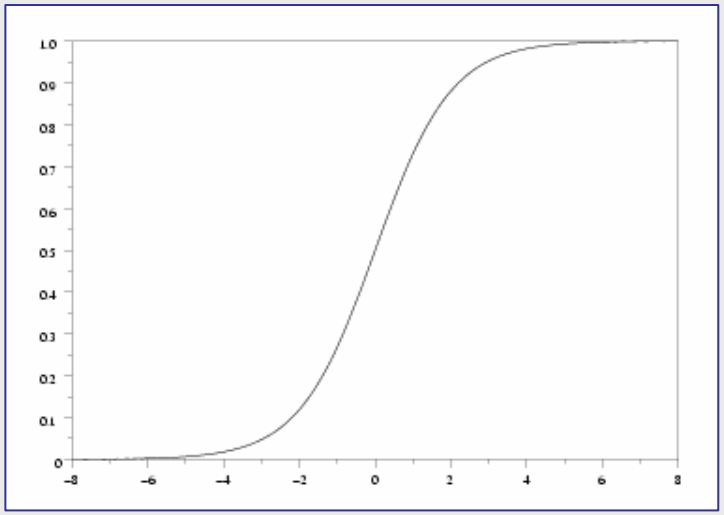
$$\text{sigm}(x) = \frac{1}{1 + e^{-x}}$$

→ Tangente hiperbólica (-1,1)

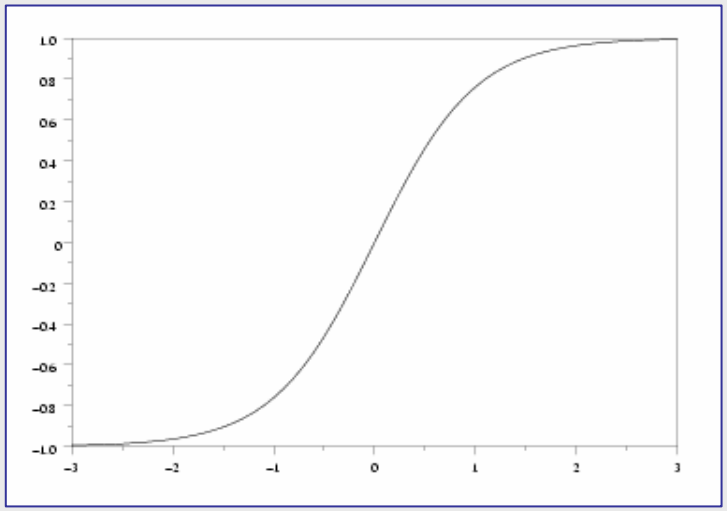
$$\text{sigm}(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{x}}$$

FUNCIONES DE ACTIVACIÓN: FORMAS GRÁFICAS

Logística



Tangente Hiperbólica



RETROPORPAGACIÓN DEL ERROR

→ Se explora el grado de influencia de los pesos en el error

→ Estos se ajustan desde la capa de salida hacia la capa de entrada

→ Necesita que las funciones de activación sean diferenciables

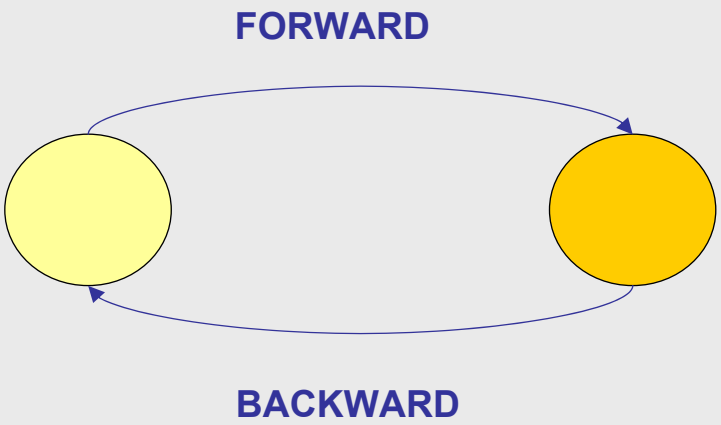
Error $\varepsilon_k = (d_k - z_k)$

Salida Deseada **Salida Calculada**

RETROPORPAGACIÓN EN DOS PASADAS

→ FORWARD: Con los pesos fijos se calcula la respuesta de las distintas unidades (capa oculta y de salida) y se determina el error.

→ BACKWARD: La señal del error es propagada hacia atrás usando los pesos de la red y ajustando los pesos.



TASA DE APRENDIZAJE (μ)

→ Es la encargada de la velocidad en que son modificados los pesos en cada una de las iteraciones del algoritmo

→ Toma valores entre 0 y 1

→ Valor alto

- Rápida minimización del error
- Soluciones poco precisas e inestables

→ Valor bajo

- Mayor precisión en la búsqueda de la minimización del error
- Mayor cantidad de épocas pata ajustar el modelo
- Sobreajuste por alto número de iteraciones del algoritmo

$$\Delta w(k) = \mu \cdot f`(e) \cdot x(k)$$

Tasa de aprendizaje

MÉTODO DE APRENDIZAJE (4)

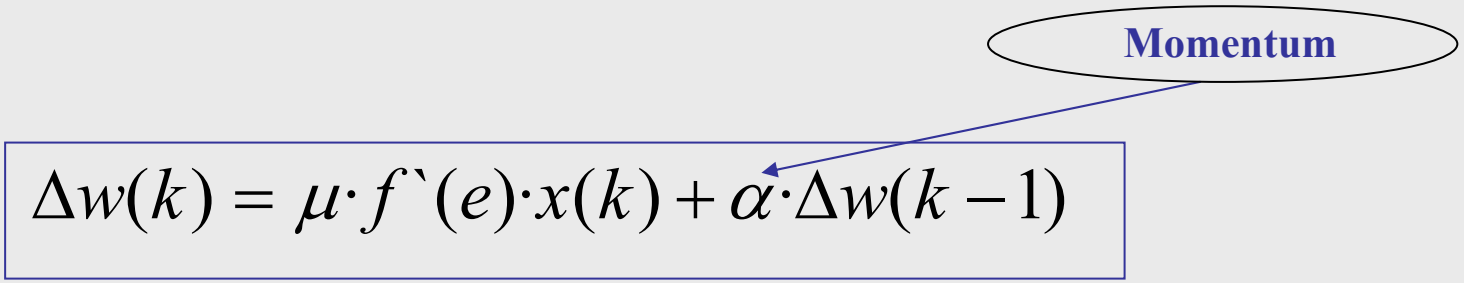
MOMENTUM

→ Trata de aumentar la tasa de aprendizaje sin producir inestabilidad

→ Trata de aumentar la velocidad de convergencia

→ Ayuda en los “baches” provocado por los mínimos locales

→ Amortigua las oscilaciones del error e durante el aprendizaje

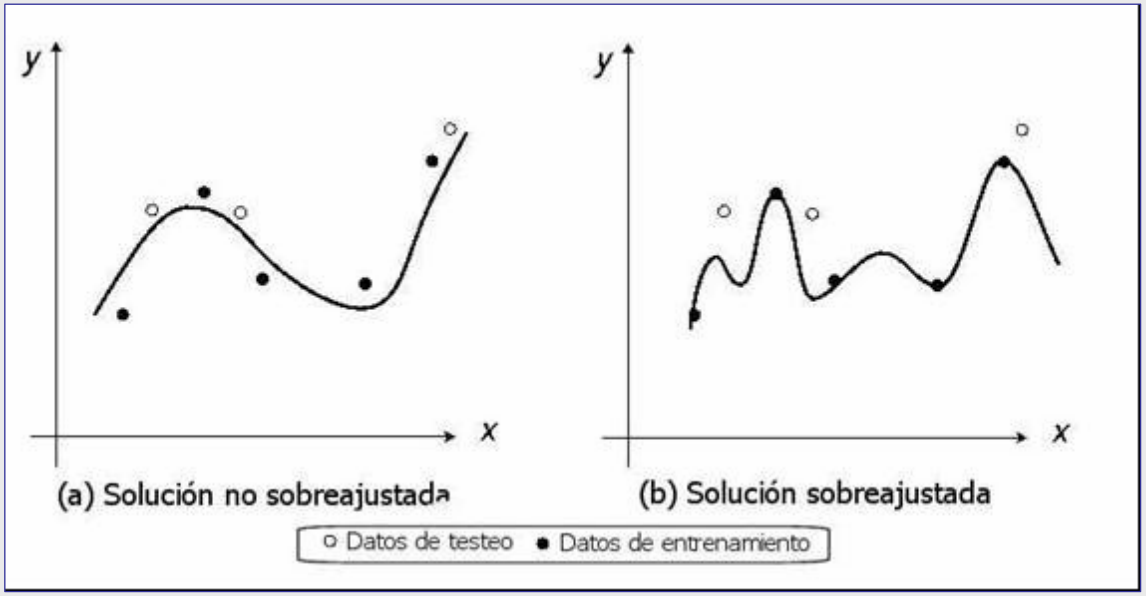


SOBREAJUSTE DE LA RED

Ocurre cuando la red se aprende los patrones “de memoria”

Produce una mala generalización del modelo a nuevos ejemplos

Se ajusta sólo a los datos de entrenamiento

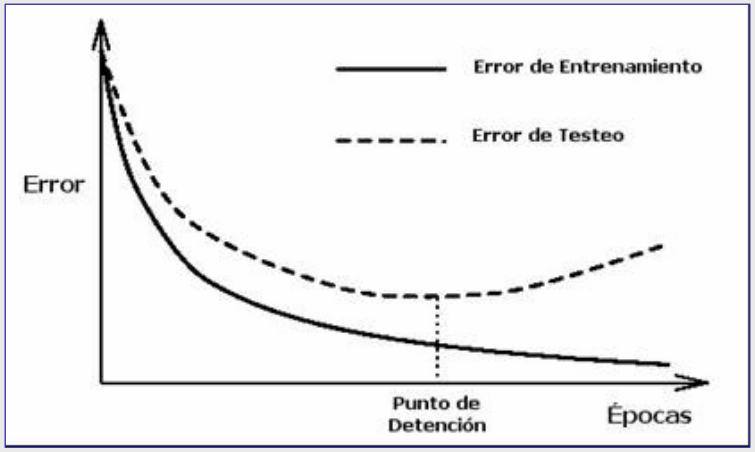


SOBREAJUSTE DE LA RED (2)

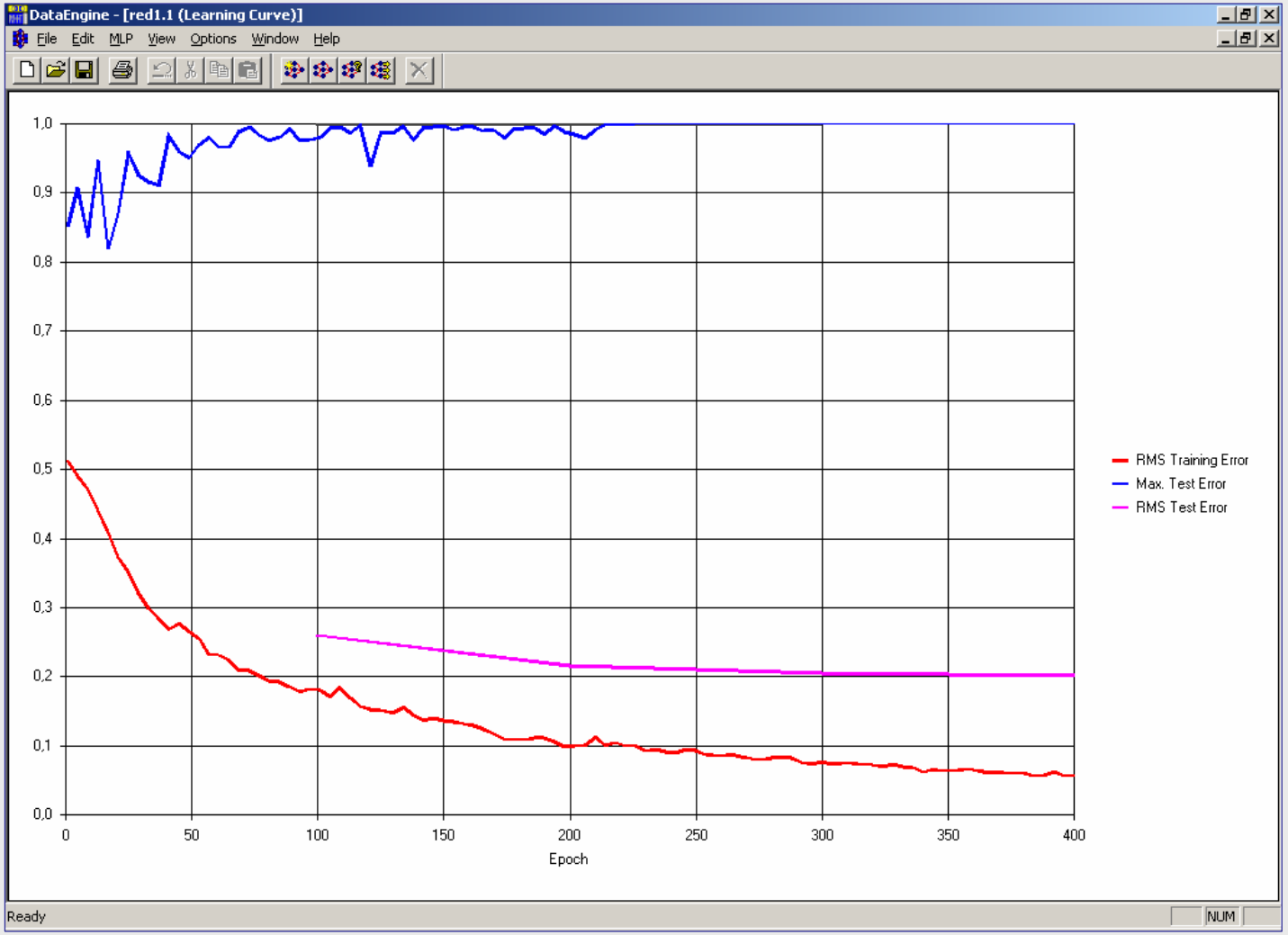
SOLUCIÓN: DETENCIÓN TEMPRANA

Busca encontrar el punto de equilibrio entre el número de épocas usado para el entrenamiento y el sobreajuste de la red

Lo mide a través de una comparación entre el error en el conjunto de entrenamiento y el conjunto de test



SOBREAJUSTE



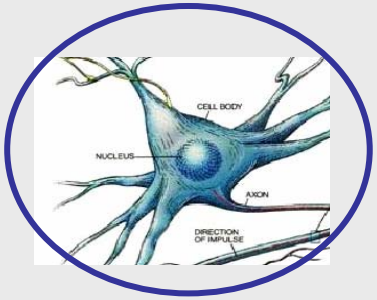
FORTALEZAS

Tiene una alta aplicabilidad

→ Data Mining supervisado y no supervisado

→ Variadas áreas de estudio

- Industrial
- Geología
- Biotecnología y genética



Aplicabilidad a soluciones complejas

→ Modelación no lineal

→ Gran adaptabilidad a los inputs de la red

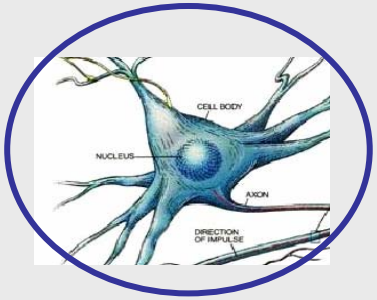
Trabaja tanto con variables categóricas (transformadas) como continuas

DEBILIDADES

Los datos deben ser preprocesados

→ Uso de normalizaciones a intervalos definidos

→ EJ: Normalización intervalo [0,1]



Los resultados son valores continuos

→ Es difícil introducir variables categóricas a pesar de ser transformadas

Poseen una enorme cantidad de parámetros

→ Hay enormes cantidades de combinaciones

→ Problema combinatorial respecto al número de combinaciones posibles entre los valores de los parámetros de su arquitectura

→ Diferencias pequeñas en la información de entrada pueden causar enormes cambios en la salida

MODELOS SUPERVISADOS

ÁRBOLES DE DECISIÓN

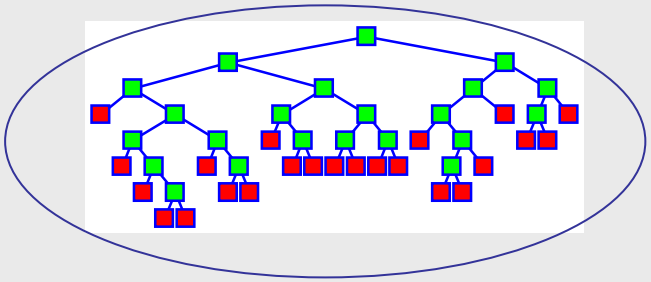
Jaime Miranda
Departamento de Ingeniería Industrial
Universidad de Chile

ÁRBOLES DE DECISIÓN

Aplicaciones

→ Segmentación de clientes

→ Generación de reglas de clasificación en general



Fortalezas

→ Fácil interpretación y entendimiento

→ Genera un ranking automático de variables

→ Rápida convergencia del algoritmo

Debilidades

→ Si poseen mucha “profundidad” son difíciles de interpretar

→ Posibilidades discretas: relacionado a variables con muchas categorías

UN PEQUEÑO EJEMPLO...

Determinación de la renta usado variables sociodemográficas

	Ck	X1	X2	X3	X4
k	Nivel de Renta	Estado Civil	Nivel Educacional	Rango Edad	Sexo
1	0-200	Soltero	Ed. Media	20-29	Hombres
2	201-400			30-39	
3	401-600	Casado	EUN	40-49	
4	601-800			50-59	Mujeres
5	801-1000	Separado	Tecnico	60-69	
6	1001-2000	Viudo	Universitario	70-79	
7	2001-15000			80-89	

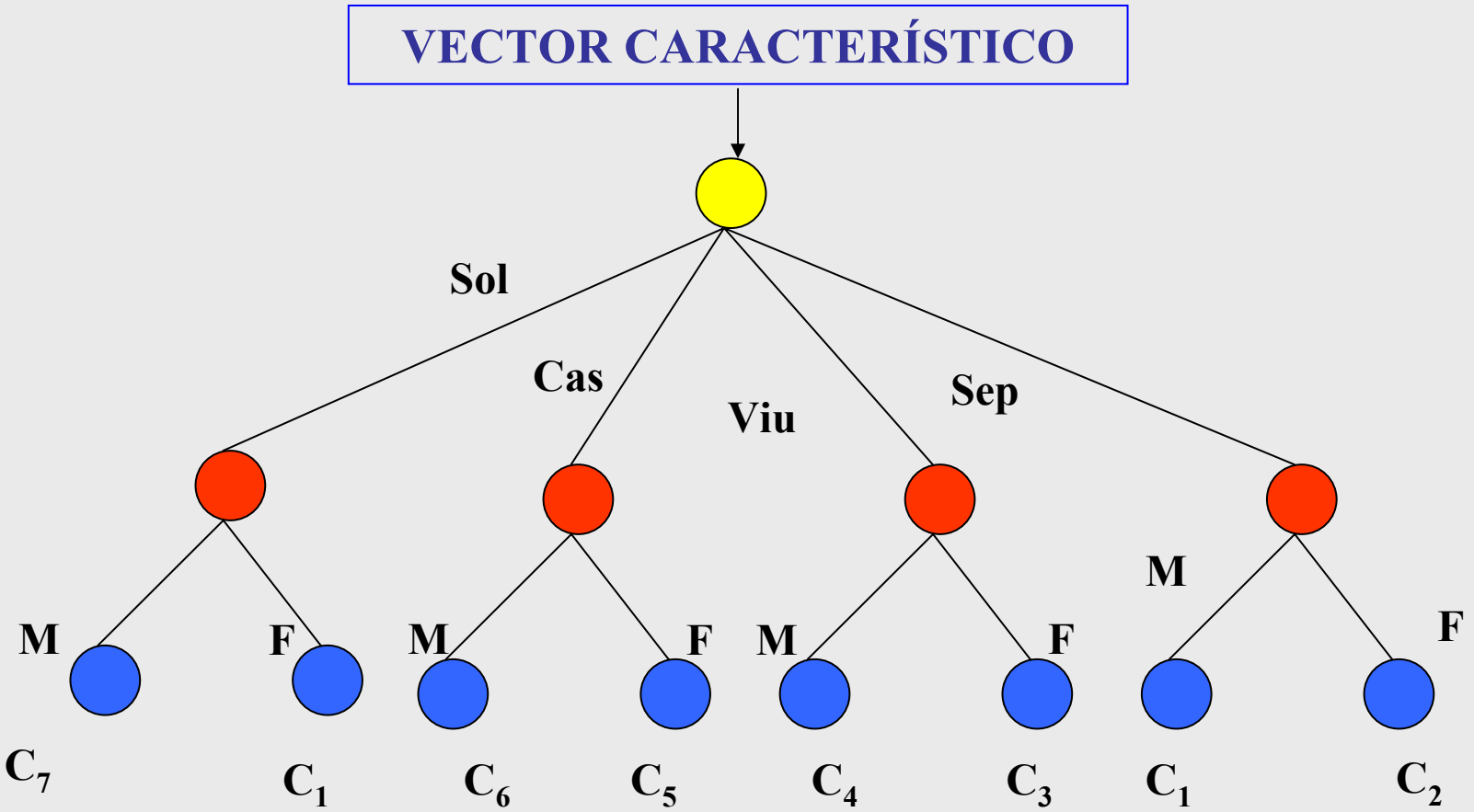
DESCRIPCIÓN :

→ Información de clientes completa: poseen algún valor en todas las variables

→ Un conjunto de clientes que no se les conoce la renta

→ PROBLEMA: Se desea estimar la renta de un conjunto nuevo de clientes

UN PEQUEÑO EJEMPLO...

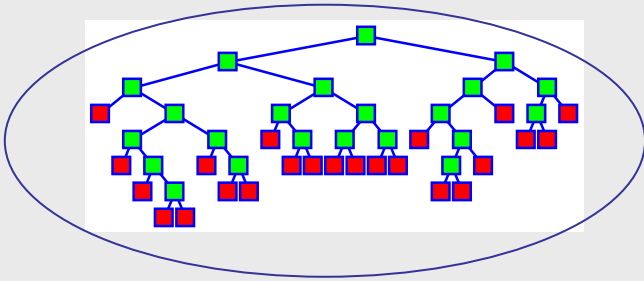


CARACTERÍSTICAS GENERALES

Separan datos en conjuntos de reglas que probablemente respondan a un efecto o variable objetivo

Son un conjunto:

→ Conexo
→ Acíclico
→ Dirigido.



Permite tener :

→ Valores mal clasificados.
→ Valores perdidos.
→ Una ilustración sobre la manera en que se pueden desglosar los problemas y la secuencia del proceso de decisión (subproblemas).

TIPOS ÁRBOLES DE DECISIÓN

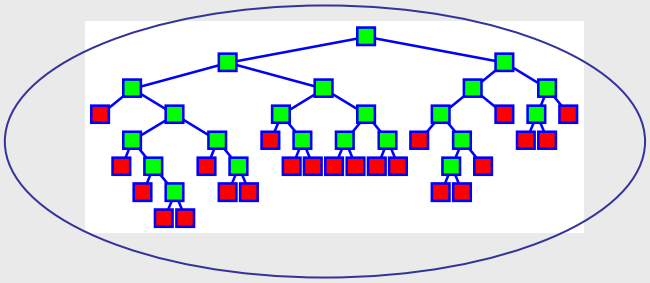
CLASIFICACIÓN

→ Se trata de encontrar el grado de pertenencia de un objeto a una clase especifica

REGRESIÓN

→ Se trata de predecir un valor futuro de una variable en base a su comportamiento pasado

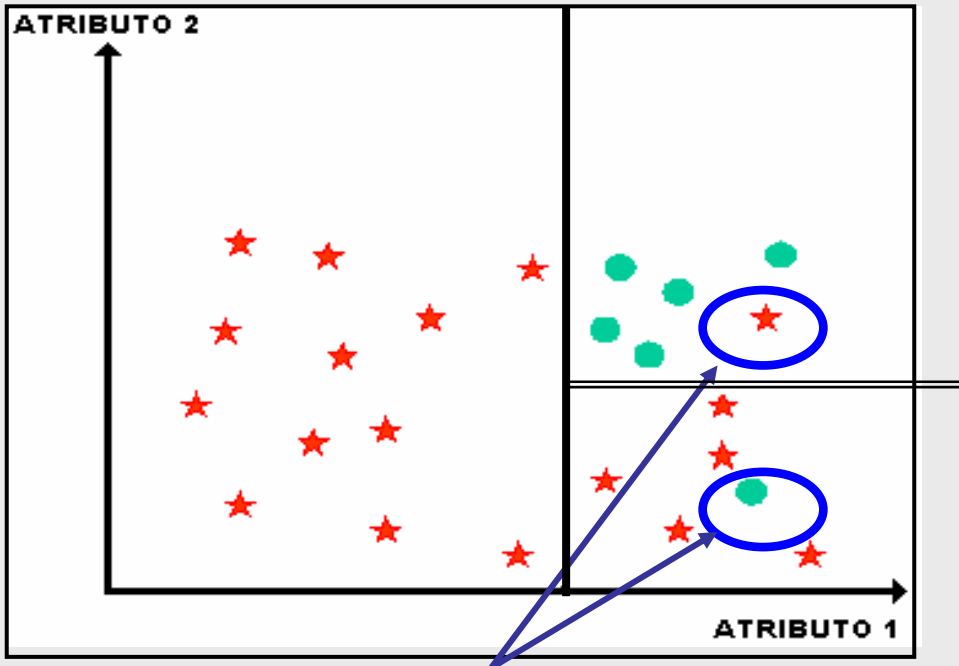
→ Usado principalmente en series temporales



EN FORMA GENERAL …

OBJETIVO

"Obtener modelos que discrimine las instancias de entrada en diferentes clases de equivalencia por medio de los valores de diferentes atributos."



Errores de clasificación

ARQUITECTURA DEL ÁRBOL

NODO

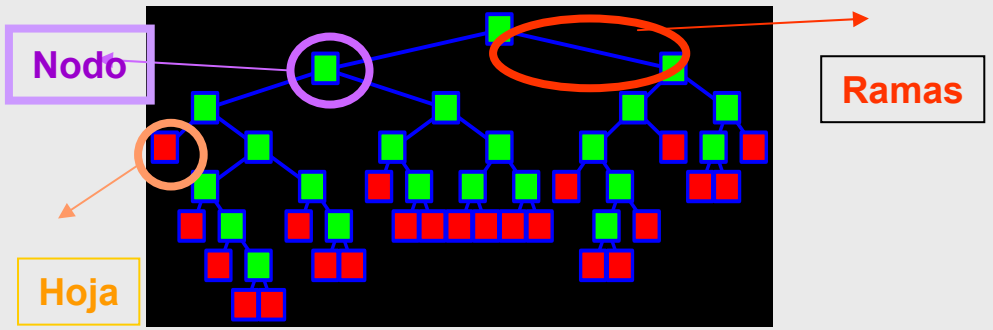
→ es un punto de unión, donde se representa un lugar en el que se debe tomar una decisión.

RAMA

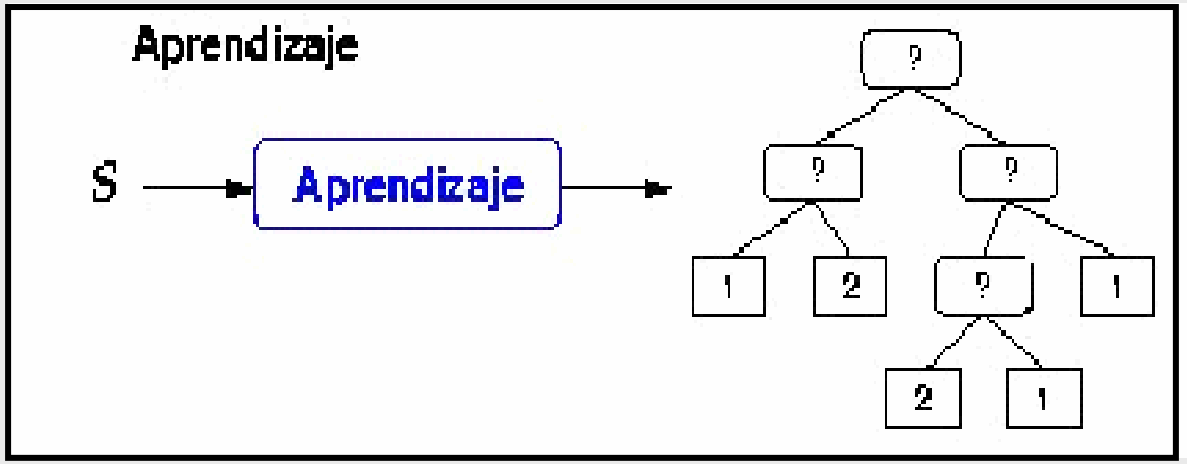
→ representa un arco de conexión entre nodos.

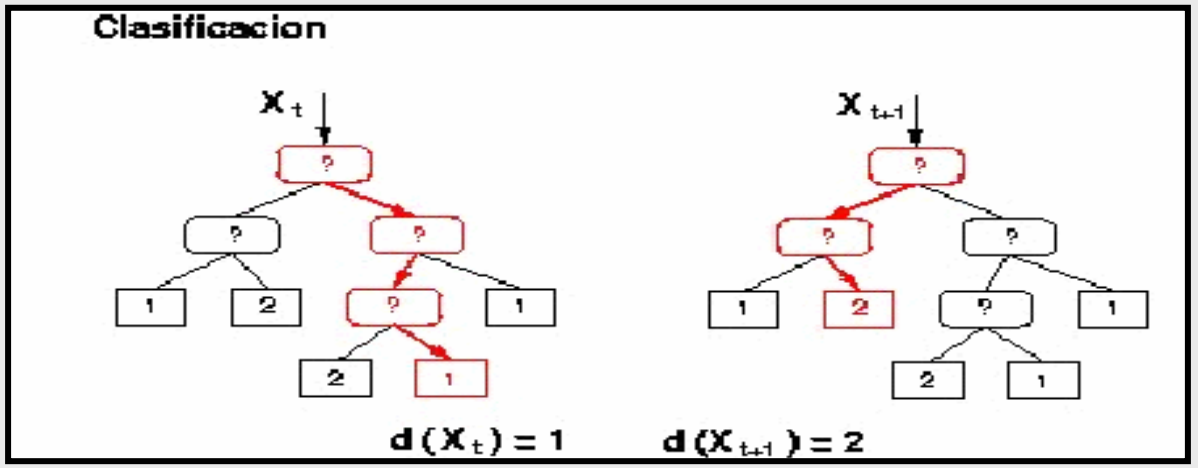
HOJA

→ es un nodo terminal (sin hijos)



ETAPAS ENTRENAMIENTO





RESPECTO A LAS DIVISIONES...

Existen varios criterios para poder hacer la división en el nodo (+ 15)

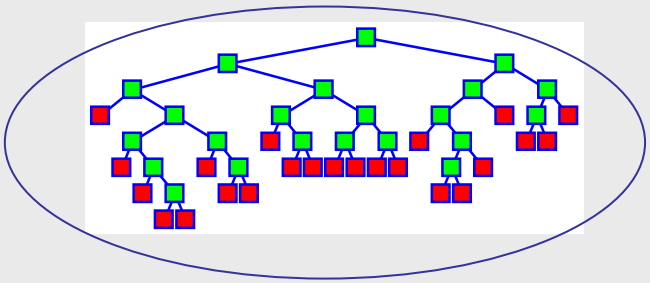
→ Gini Index

→ Twoing

→ CHAID

Cada partición depende de un único atributo

La gran mayoría se basa en medidas de diversidad o “desorden” del nodo



GINI INDEX

$$Gini(V) = \sum_{i=j} p(v) * \sum_{i \neq j} p(i \setminus v) * p(j \setminus v)$$

Probabilidad de estar en el nodo v

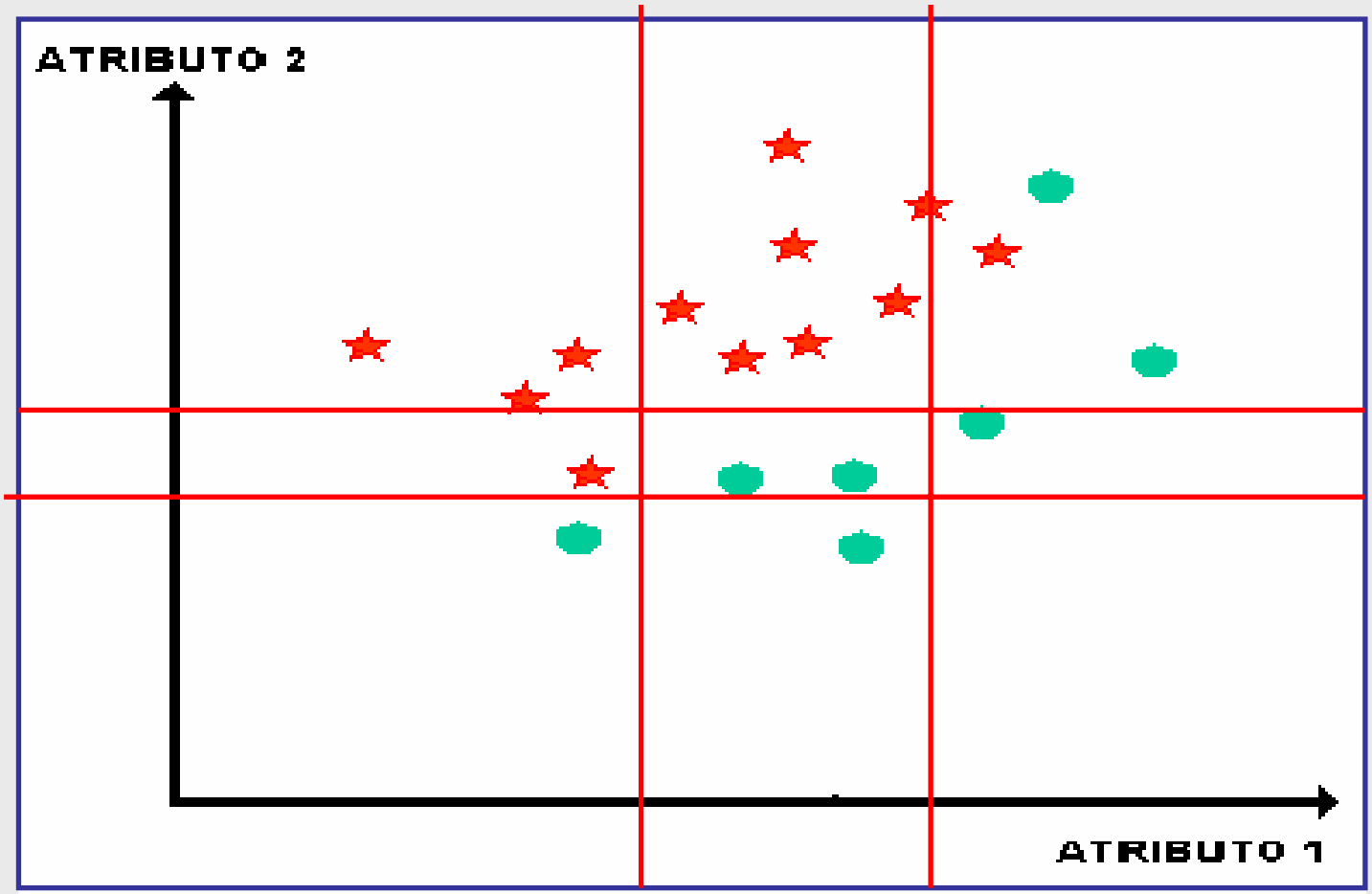
Probabilidad de pertenecer a la clase i/j dado que estoy en v

→ Se elige el atributo que posee el mayor índice de GINI

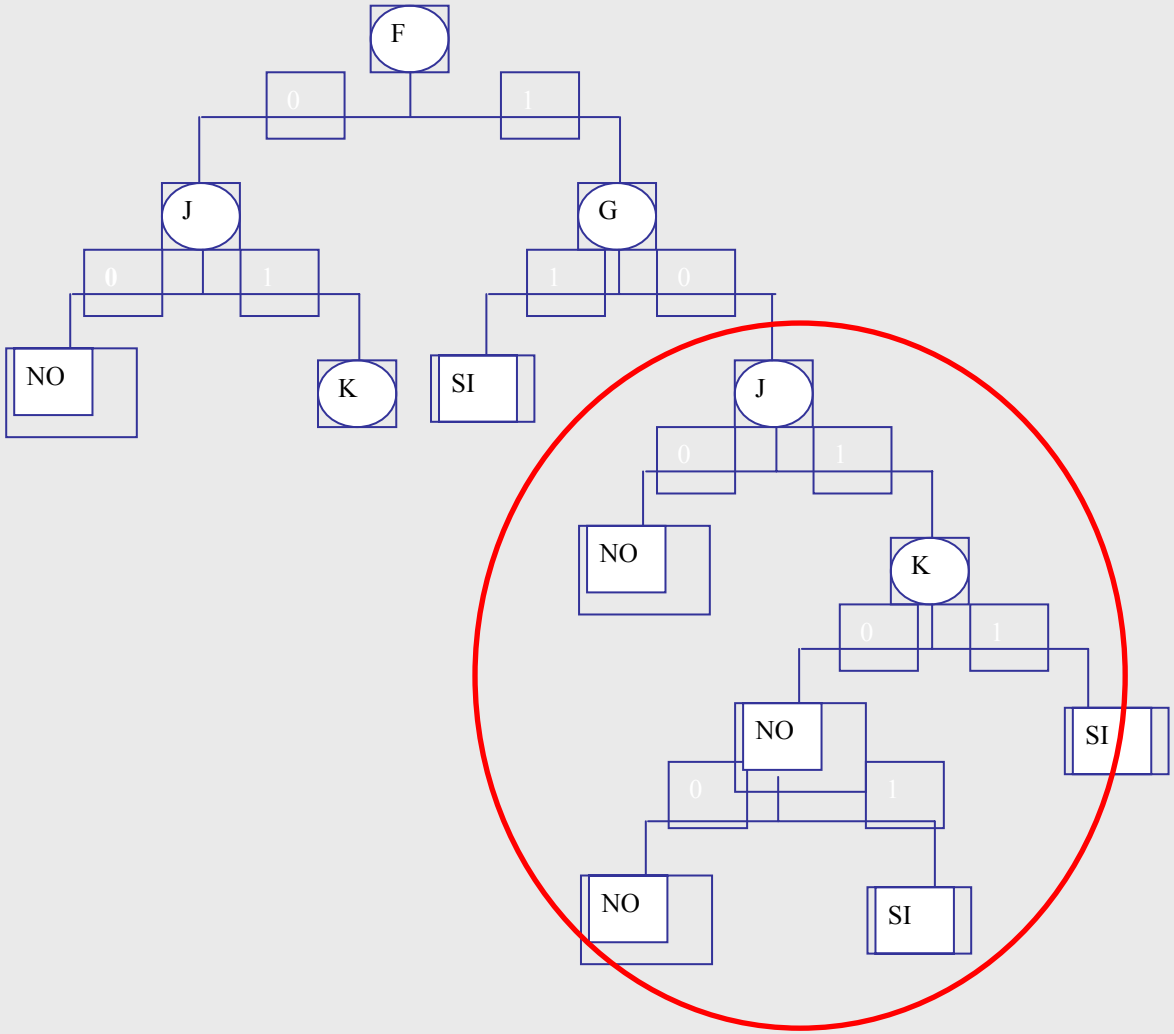
→ A medida que se baja en el árbol el atributo posee menor índice de GINI

→ Este índice ve que tan heterogéneo es el nodo respecto a los elementos que lo conforman

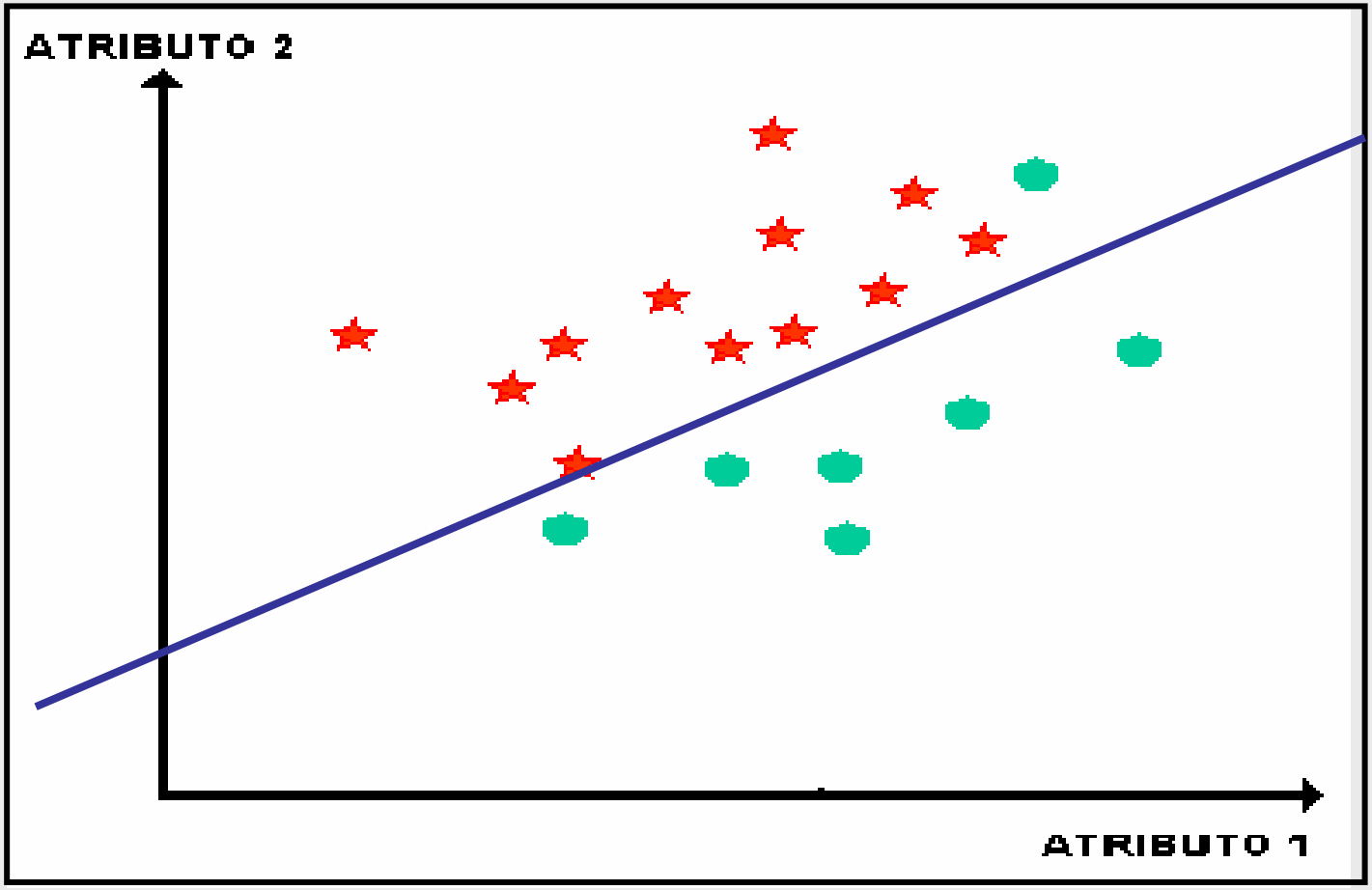
SOBREAJUSTE DEL MODELO



SOBREAJUSTE EN FORMA GRAFICA



SOLUCIÓN LINEAL



FORTALEZAS

EXPLICABILIDAD

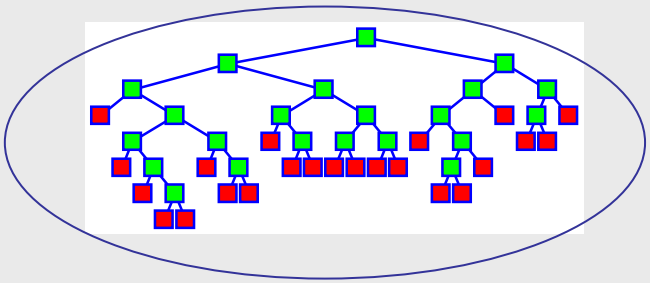
→ Es intuitivo y da claras reglas de decisión

→ Da una buena descripción visual en problemas

FÁCIL IMPLEMENTACIÓN Y CONSTRUCCIÓN

→ Las reglas pueden ser implementadas en cualquier lenguaje lógico

→ No necesitan de un fuerte apoyo computacional



DEBILIDADES

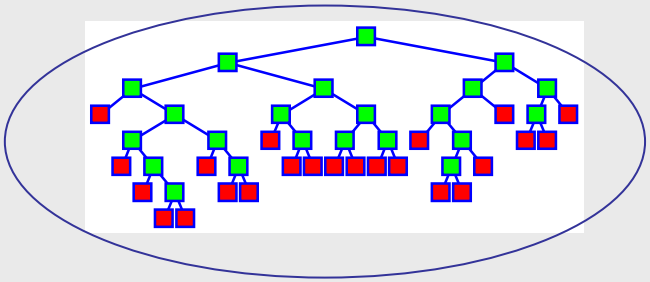
ATRIBUTOS CON NUMEROSOS VALORES

→ Es debido a que inducen particiones más finas, que no sean significativos

TENER DEMASIADOS NIVELES

→ Al tener mayor profundidad la explicabilidad decae

→ Necesidad de altos volúmenes de información



LIMITACIONES Y PROBLEMAS

RUIDO

→ Ejemplos con la misma descripción pero distinta clase

→ Consecuencia: error no nulo en ejemplos de entrenamiento.

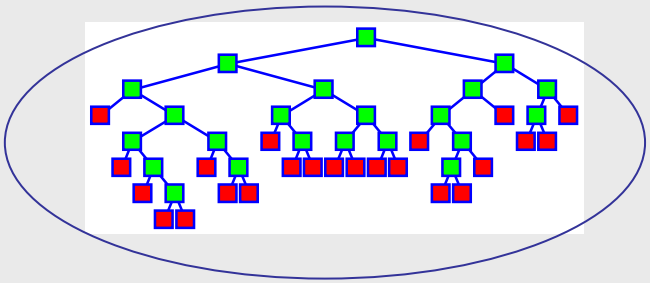
POSIBILIDADES DISCRETAS

→ Solo es posible tener un número finito de “ramas” y no un continuo.

SOBREAJUSTE

→ Uso de atributos no relevantes para ajustar árbol a datos

→ Consecuencia: disminuye capacidad generalización del modelo.



Métodos de Minería de Datos

ALGORITMOS SUPERVISADOS

JAIME MIRANDA

Departamento de Ingeniería Industrial

Universidad de Chile