Métodos de Minería de Datos

ALGORITMOS SUPERVISADOS

JAIME MIRANDA

Departamento de Ingeniería Industrial

Universidad de Chile

CONCEPTOS BÁSICOS

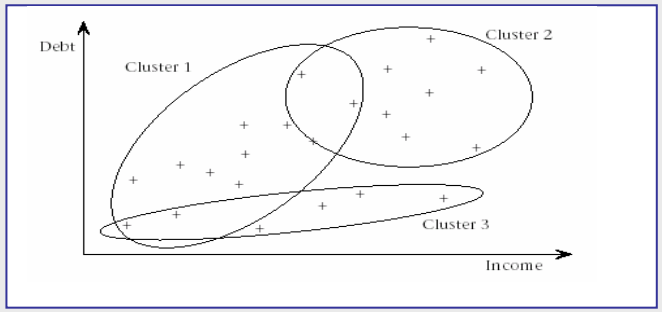
→ Medida de distancia

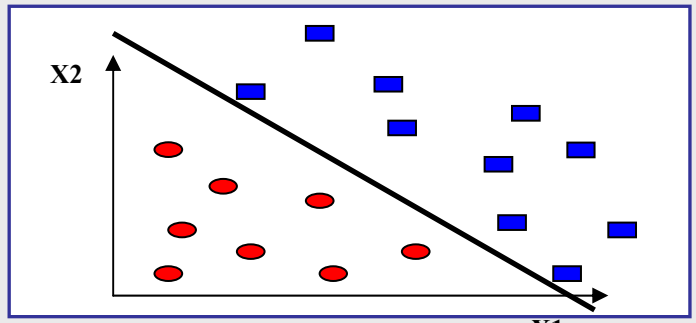
→ Prototipo o centro de clase más cercana

→ 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

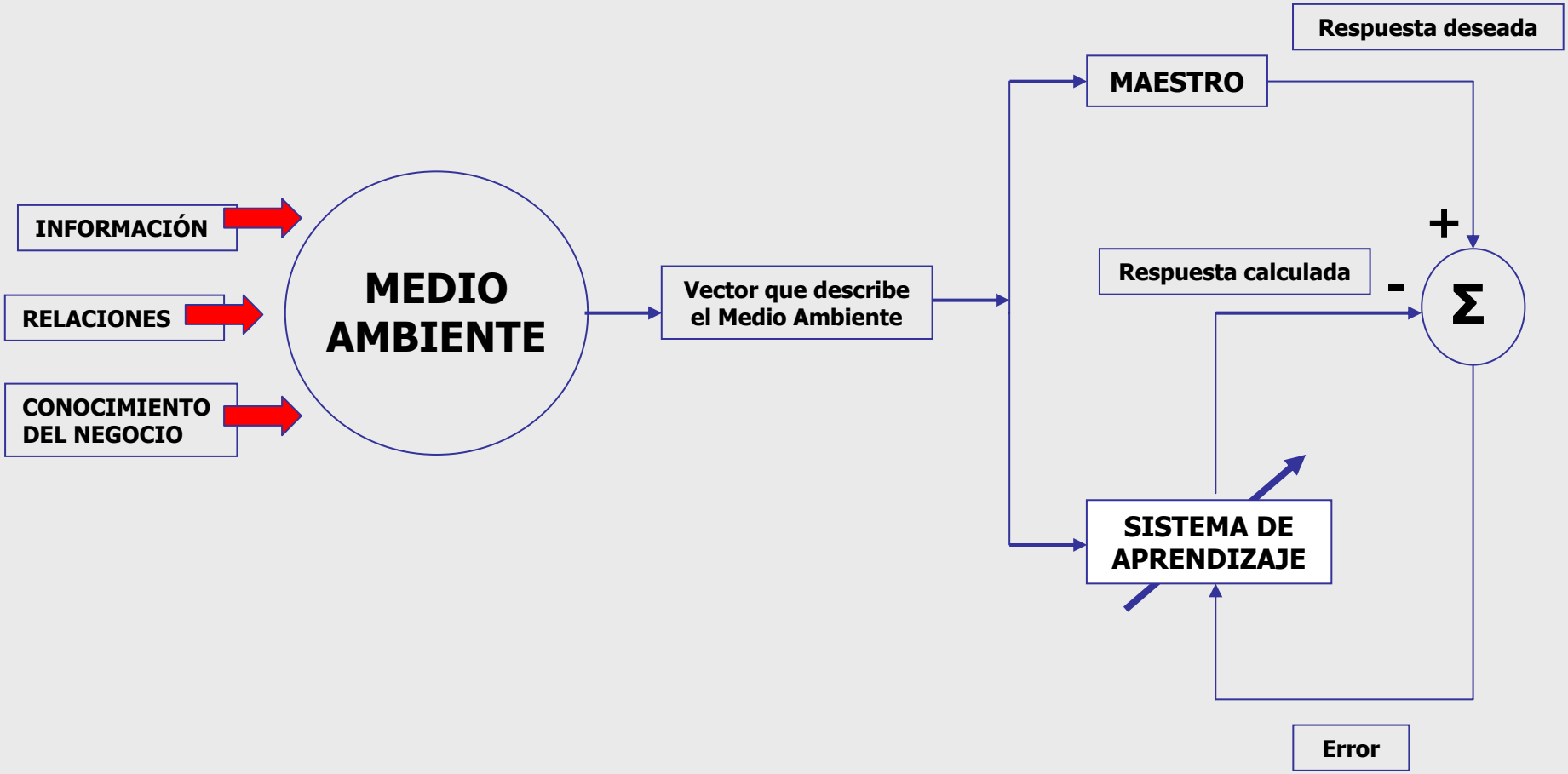
$$p(C_K / X_L) = \frac{p(X_L / C_K) * p(C_K)}{p(X_L)}$$

Donde:

C_k denota la clase k del atributo elegido como base.

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

DIAGRAMA DE APRENDIZAJE SUPERVISADO



MODELOS SUPERVISADOS

ÁRBOLES DE DECISIÓN

Jaime Miranda

Departamento de Ingeniería Industrial

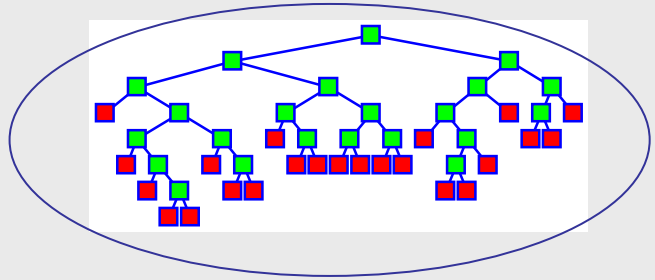
Universidad de Chile

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	

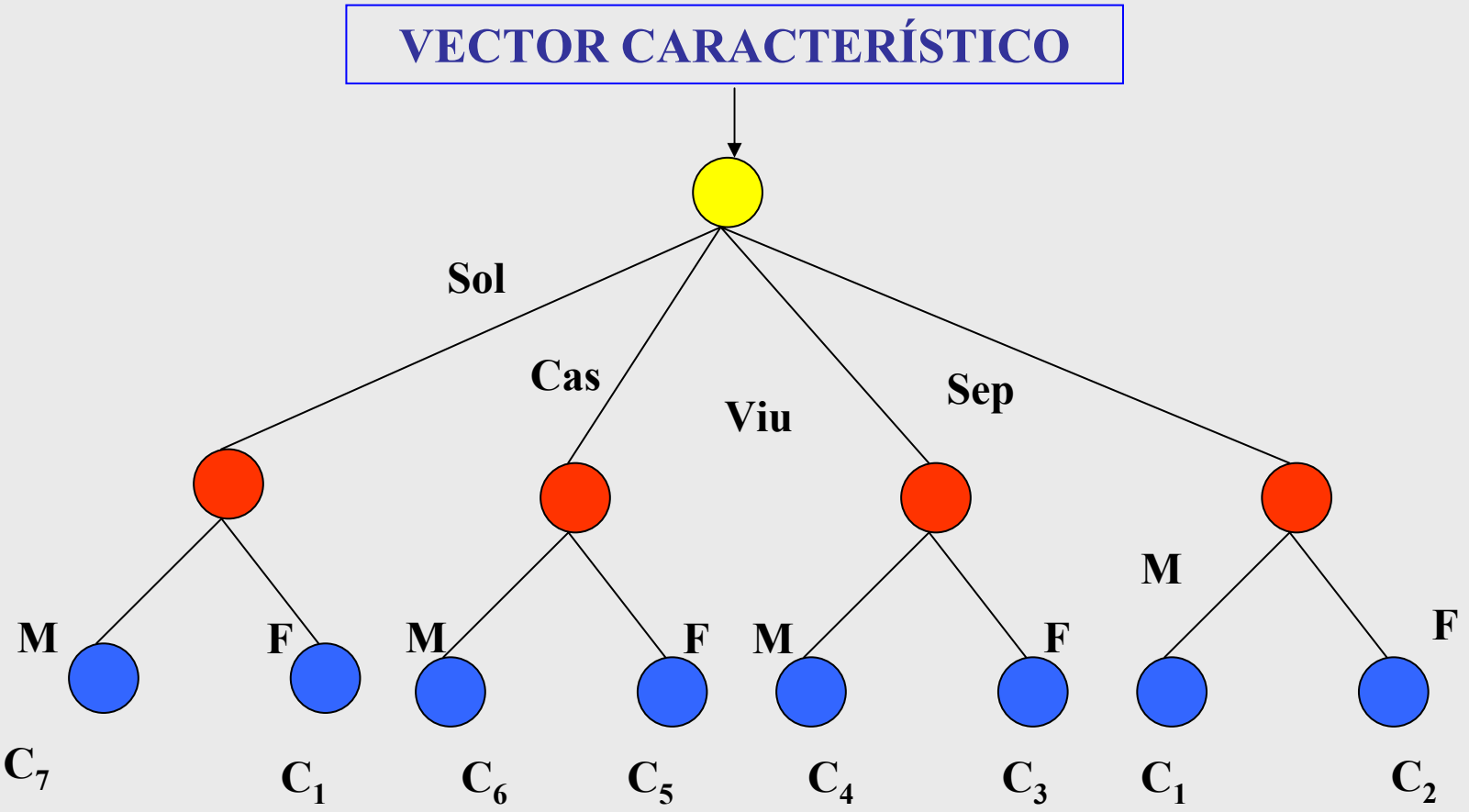
DESCRIPCIÓN :

→ 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...

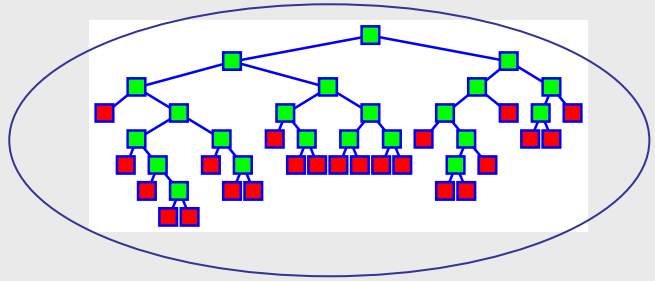


CARACTERÍSTICAS GENERALES

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

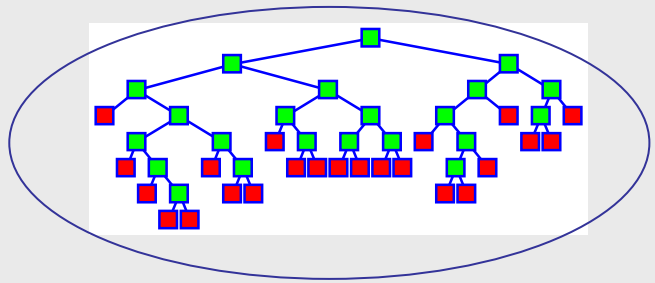
CLASIFICACIÓN

→ 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

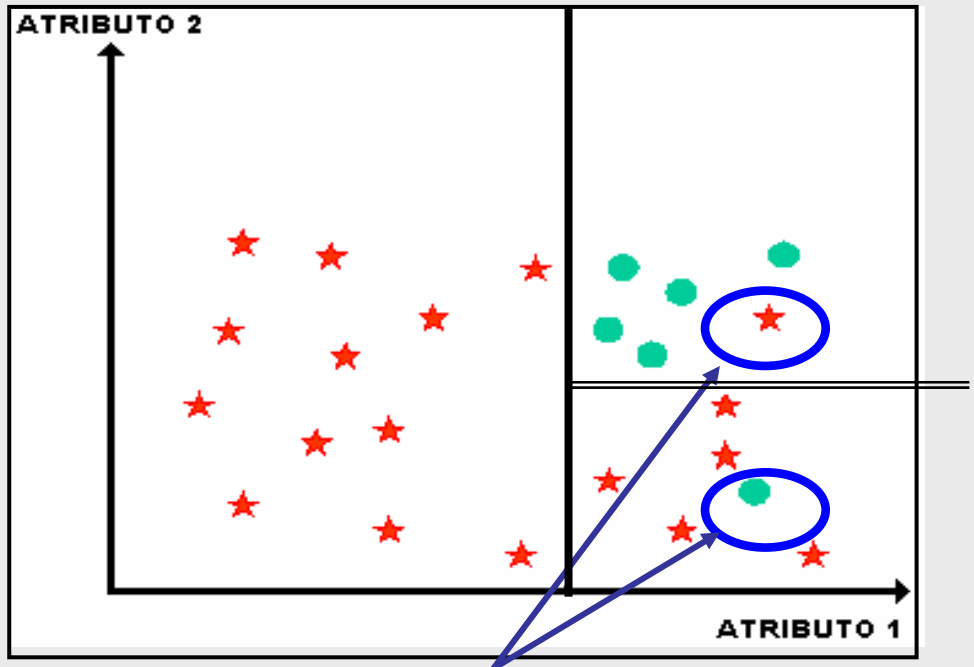
→ Usado principalmente en series temporales



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

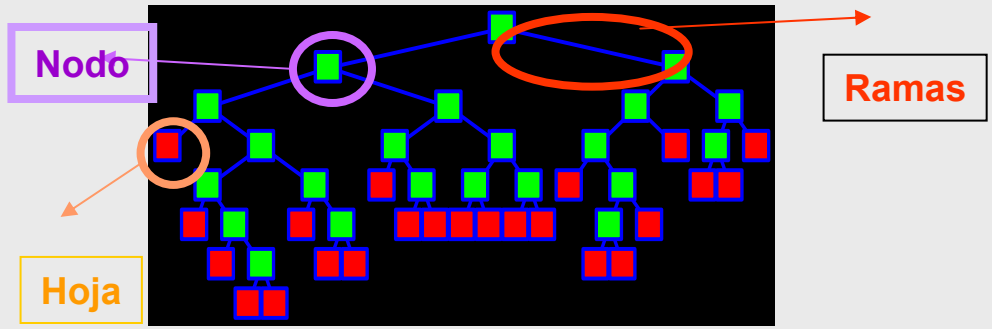
→ es un punto de unión, donde se representa un lugar en el que se debe tomar una decisión.

RAMA

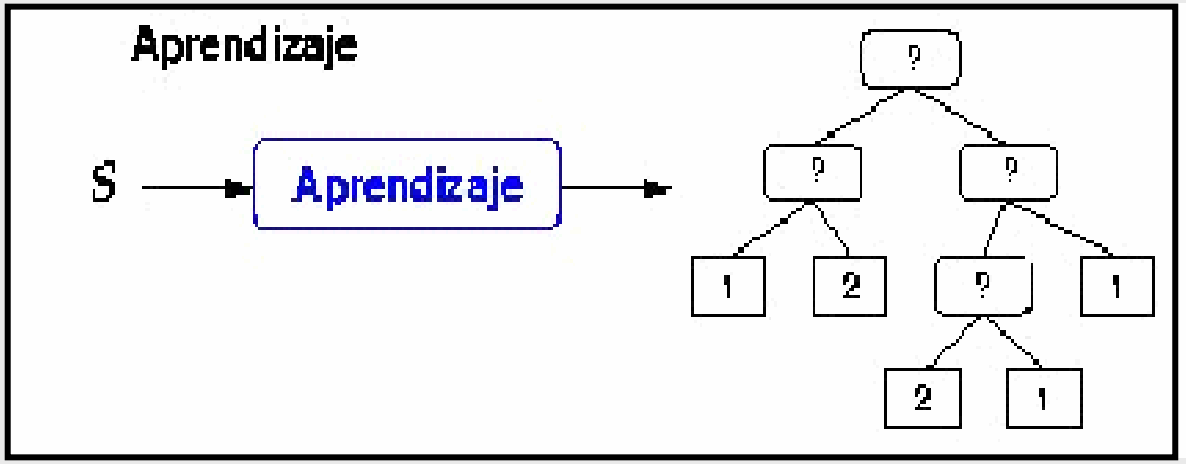
→ representa un arco de conexión entre nodos.

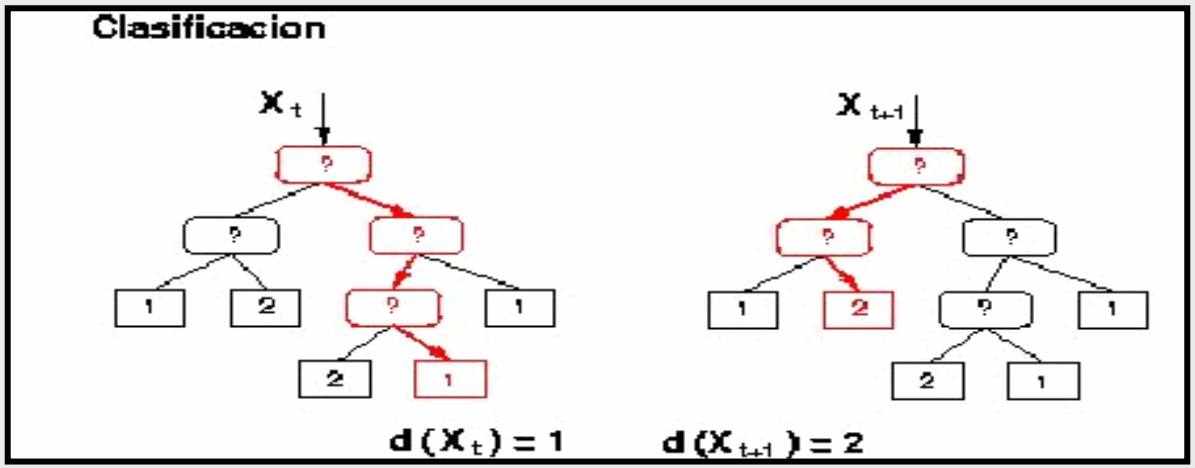
HOJA

→ es un nodo terminal (sin hijos)



ETAPAS ENTRENAMIENTO





RESPECTO A LAS DIVISIONES...

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

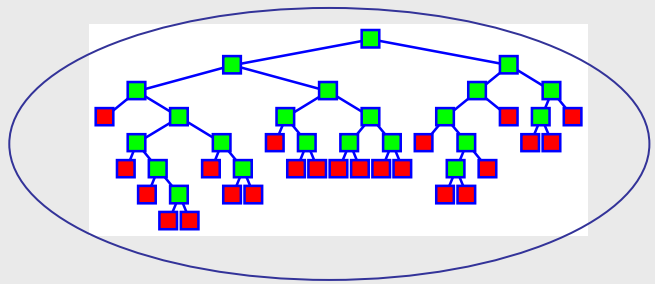
→ Gini Index

→ 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

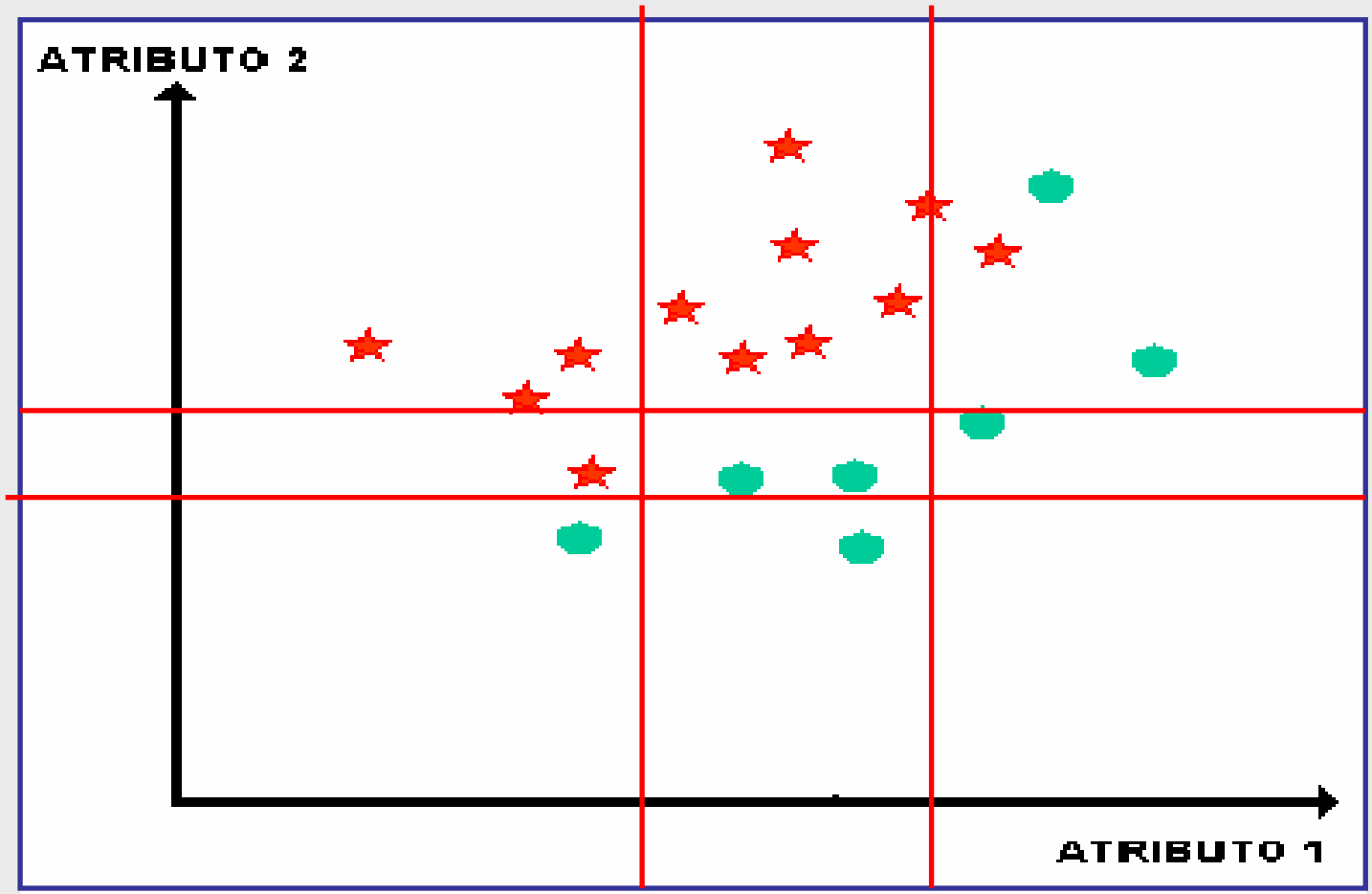
Probabilidad de pertenecer a la clase i/j dado que estoy en v

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

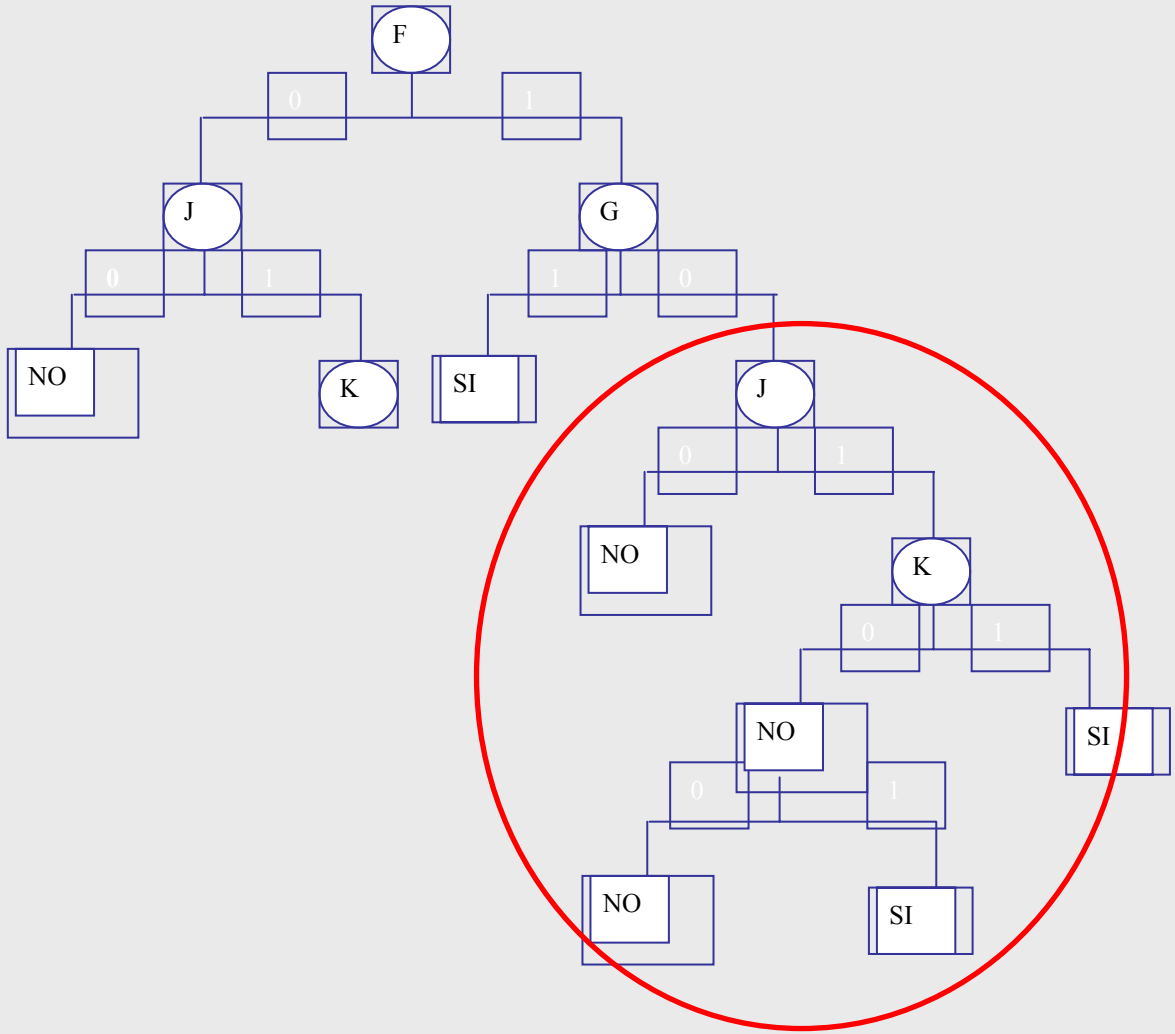
→ A medida que se baja en el árbol el atributo posee menor índice de GINI

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

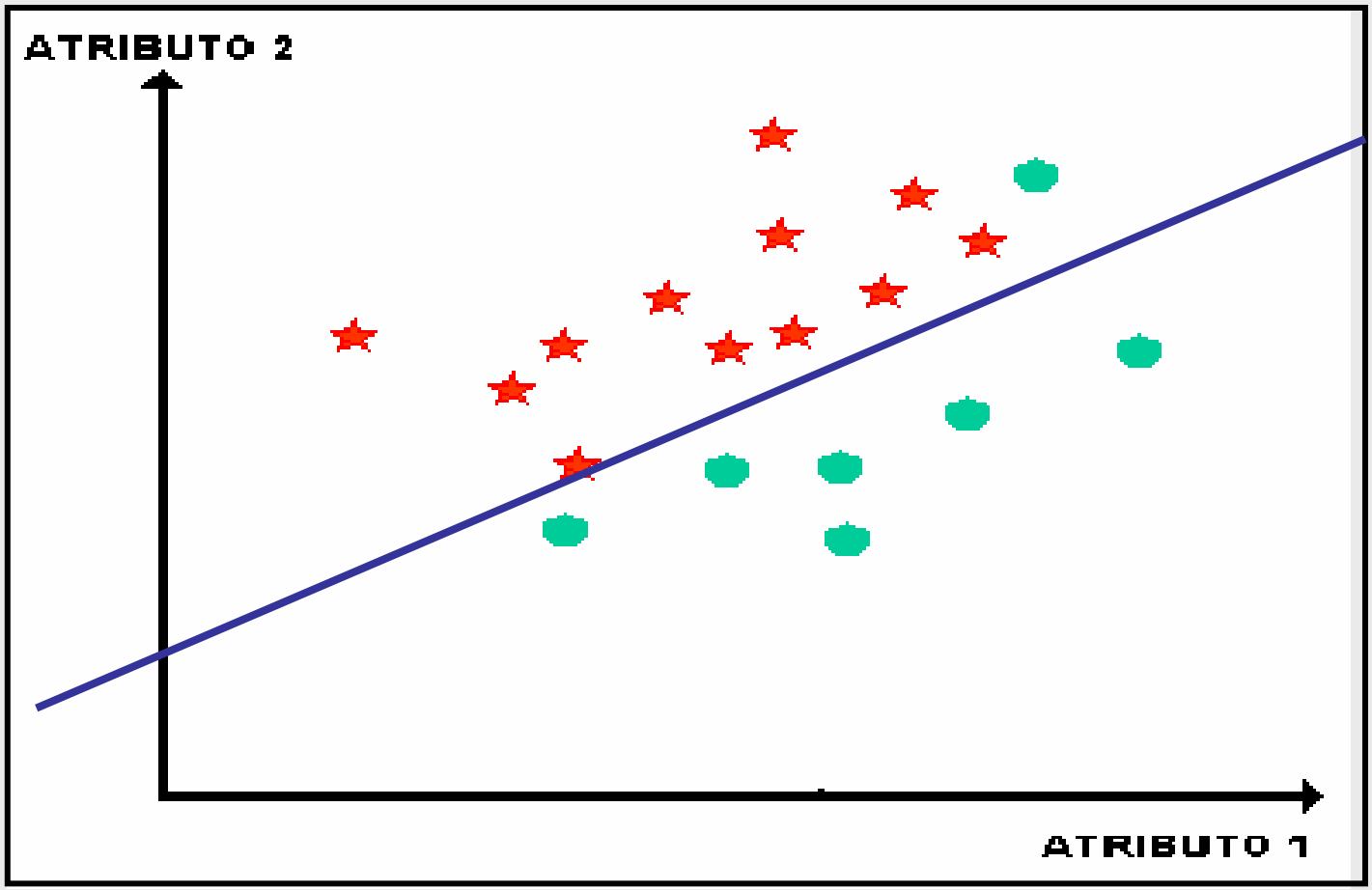
SOBREAJUSTE DEL MODELO



SOBREAJUSTE EN FORMA GRAFICA



SOLUCIÓN LINEAL



FORTALEZAS

EXPLICABILIDAD

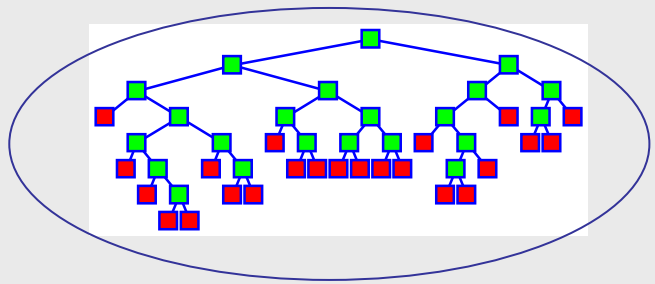
→ Es intuitivo y da claras reglas de decisión

→ 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

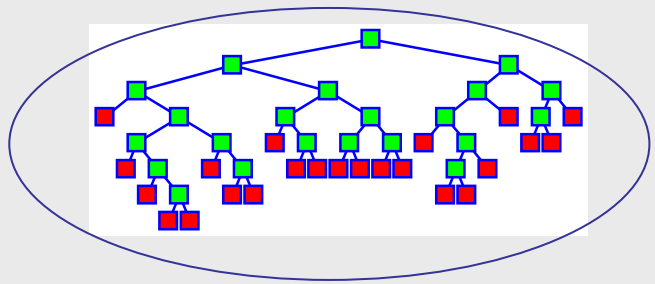
ATRIBUTOS CON NUMEROSOS VALORES

→ 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.

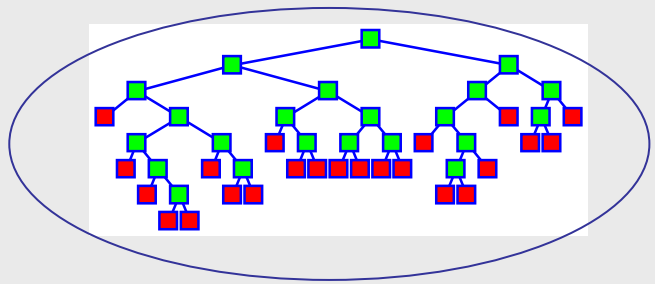
POSIBILIDADES DISCRETAS

→ 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