



MODELOS PREDICTIVOS

Árboles de Decisión

Jaime Miranda

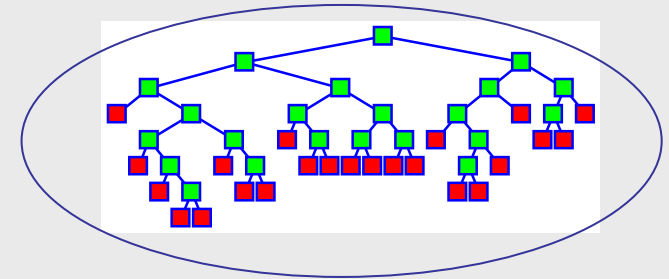
Departamento de Ingeniería Industrial
Universidad de Chile

IN47B

Ingeniería de Operaciones

Aplicaciones

- Segmentación de clientes
- Generación de reglas en general



Fortalezas

- Fácil interpretación y entendimiento
- Genera un ranking automático de variables
- Rápida convergencia del algoritmo

Debilidades

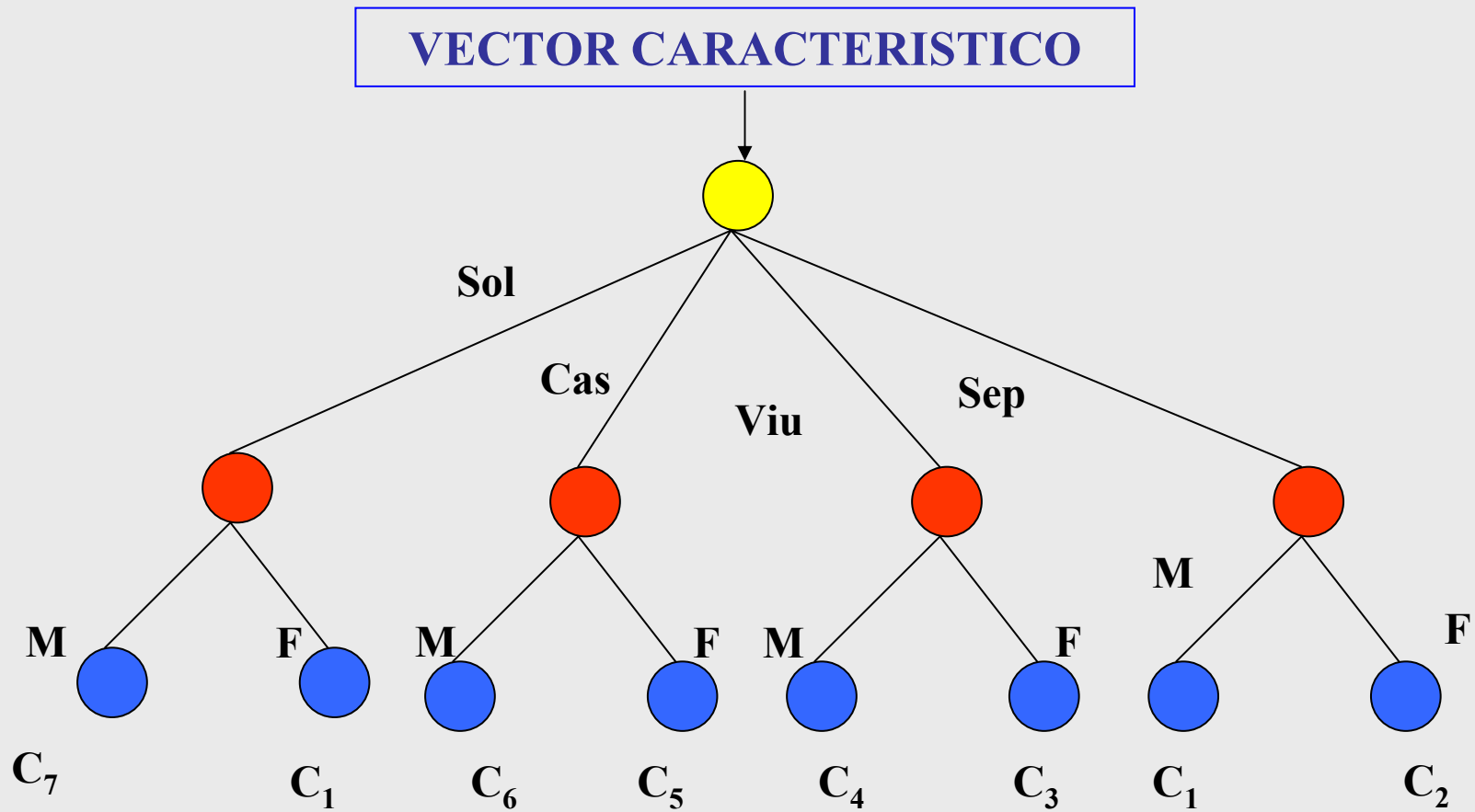
- Si poseen mucha “profundidad” son difíciles de interpretar
- Posibilidades discretas: relacionado a variables con muchas categorías

UN PEQUEÑO EJEMPLO...

Determinación de la renta usando 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	
5	801-1000	Separado	Tecnico	60-69	Mujeres
6	1001-2000			70-79	
7	2001-15000	Viudo	Universitario	80-89	

UN PEQUEÑO EJEMPLO...



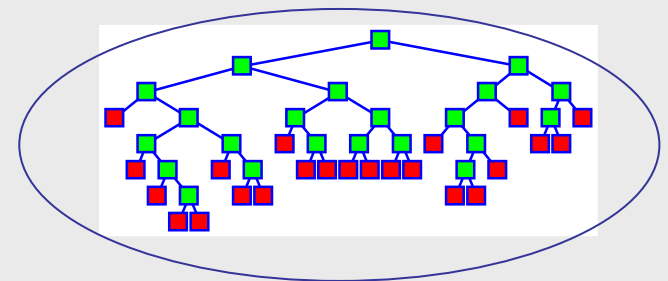
ÁRBOLES DE DECISIÓN (2)

Son un conjunto conexo, acíclico y 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).

Los árboles de decisión separan datos en conjuntos de reglas que probablemente respondan a un efecto o variable objetivo

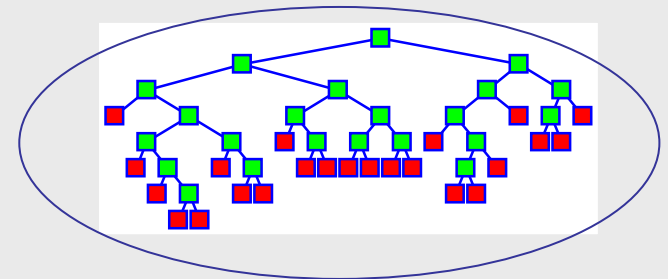


CLASIFICACIÓN

- Se trata de encontrar el grado de pertenencia de un objeto a una clase específica.

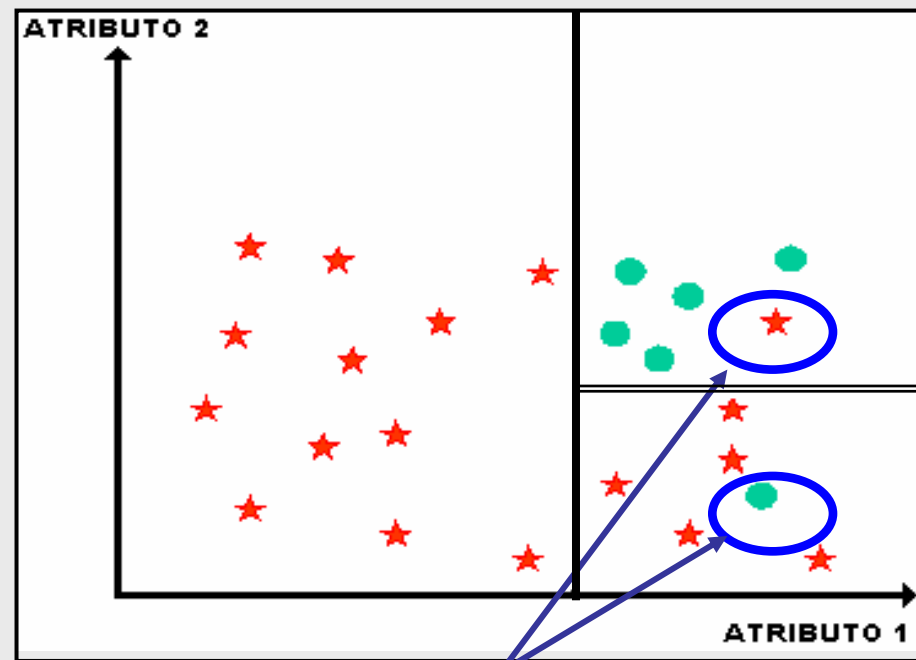
REGRESIÓN

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



OBJETIVO GENERAL

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



**Errores de
clasificación**

Un nodo

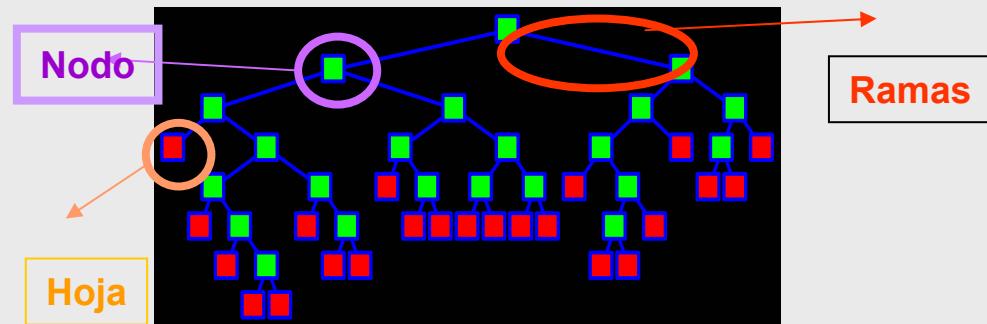
- es un punto de unión, donde se representa un lugar en el que se debe tomar una decisión.

Una rama

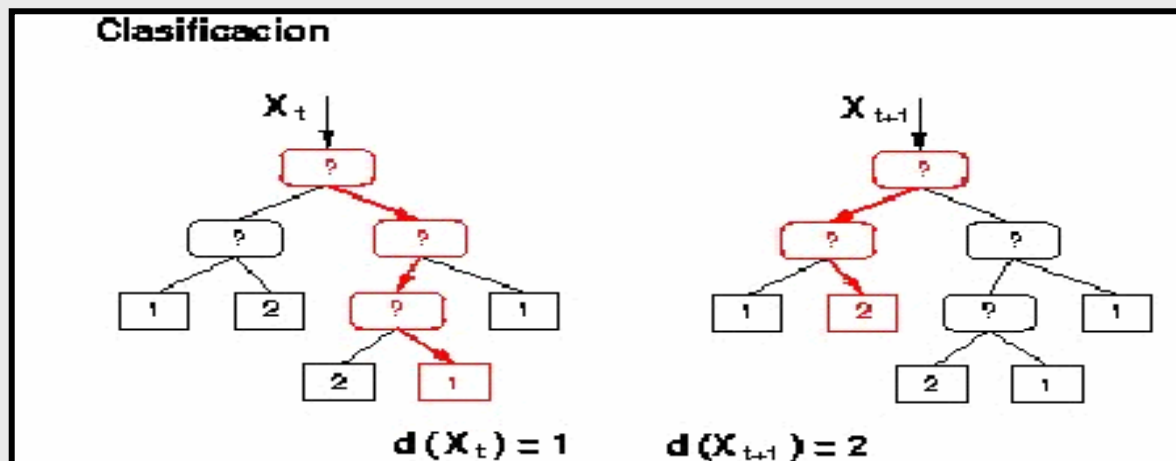
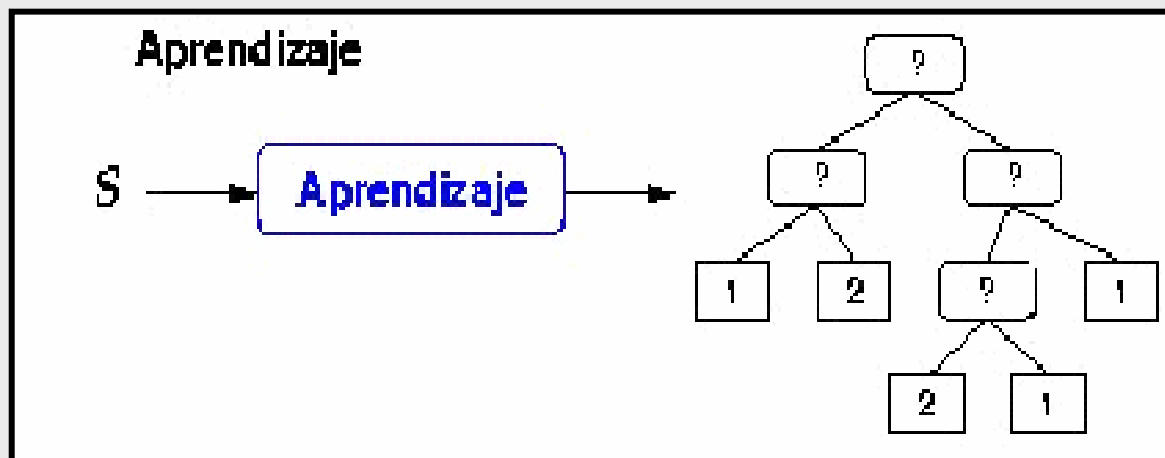
- representa un arco de conexión entre nodos.

Una hoja

- es un nodo terminal (sin hijos)



ETAPAS ENTRENAMIENTO

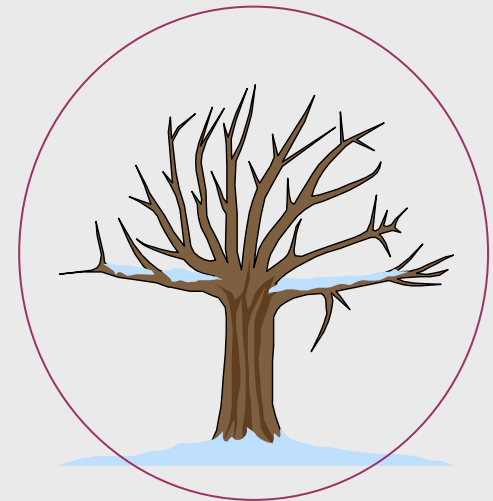


RESPECTO A LAS DIVISIONES...

Cada partición depende de un único atributo.

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

La gran mayoría se basa en la diversidad 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.

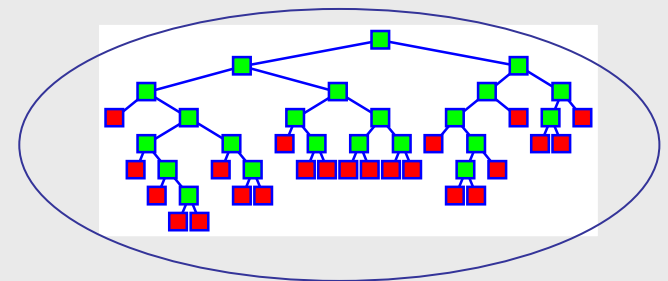
A medida que se baja en el árbol el atributo posee menor índice de GINI.

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

Un árbol de decisión da una buena descripción visual en problemas relativamente simples, pero su complejidad aumenta exponencialmente a medida que se agregan etapas adicionales.

Atributos con numerosos valores

- Debido a que inducen particiones más finas, que no sean significativos.



Ruido

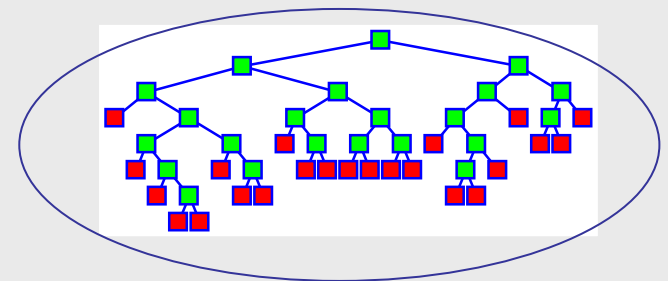
- Ejemplos con la misma descripción pero distinta clase
- Consecuencia: error no nulo en ejemplos de entrenamiento.

Posibilidades discretas

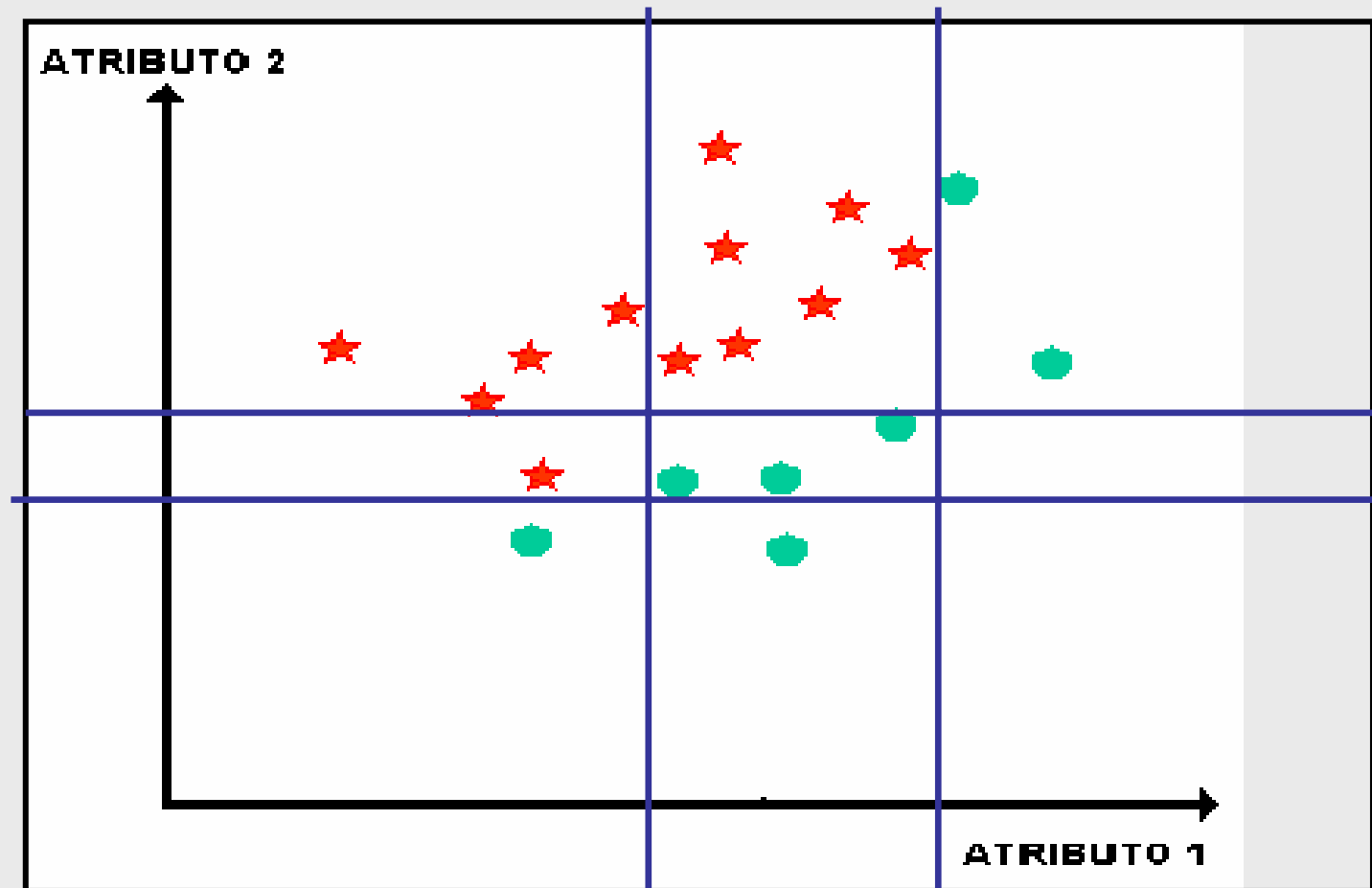
- Solo es posible tener un número finito de “ramas” y no un continuo.

Sobreajuste

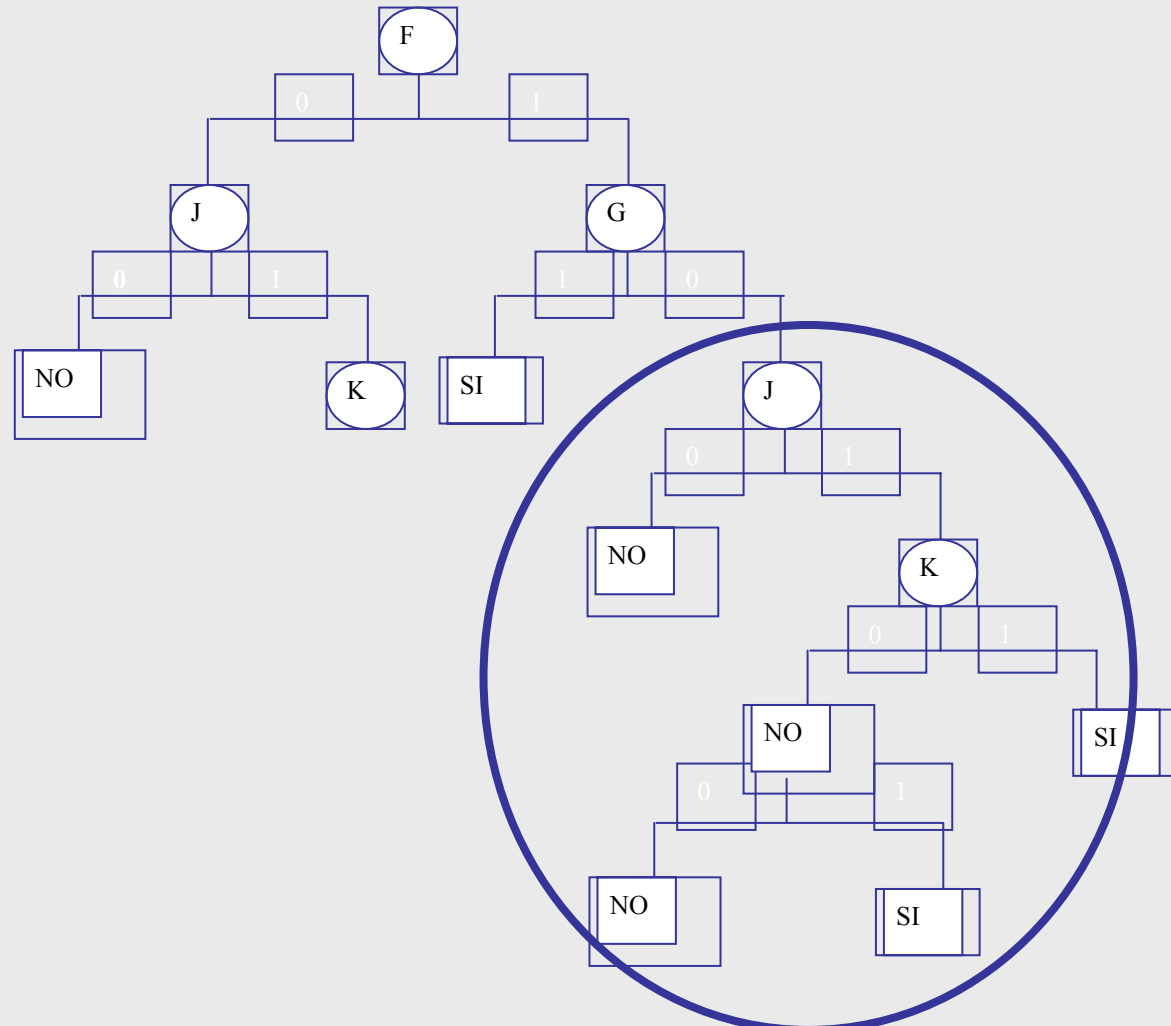
- Uso de atributos no relevantes para ajustar árbol a datos
- Consecuencia: disminuye capacidad generalización del modelo.



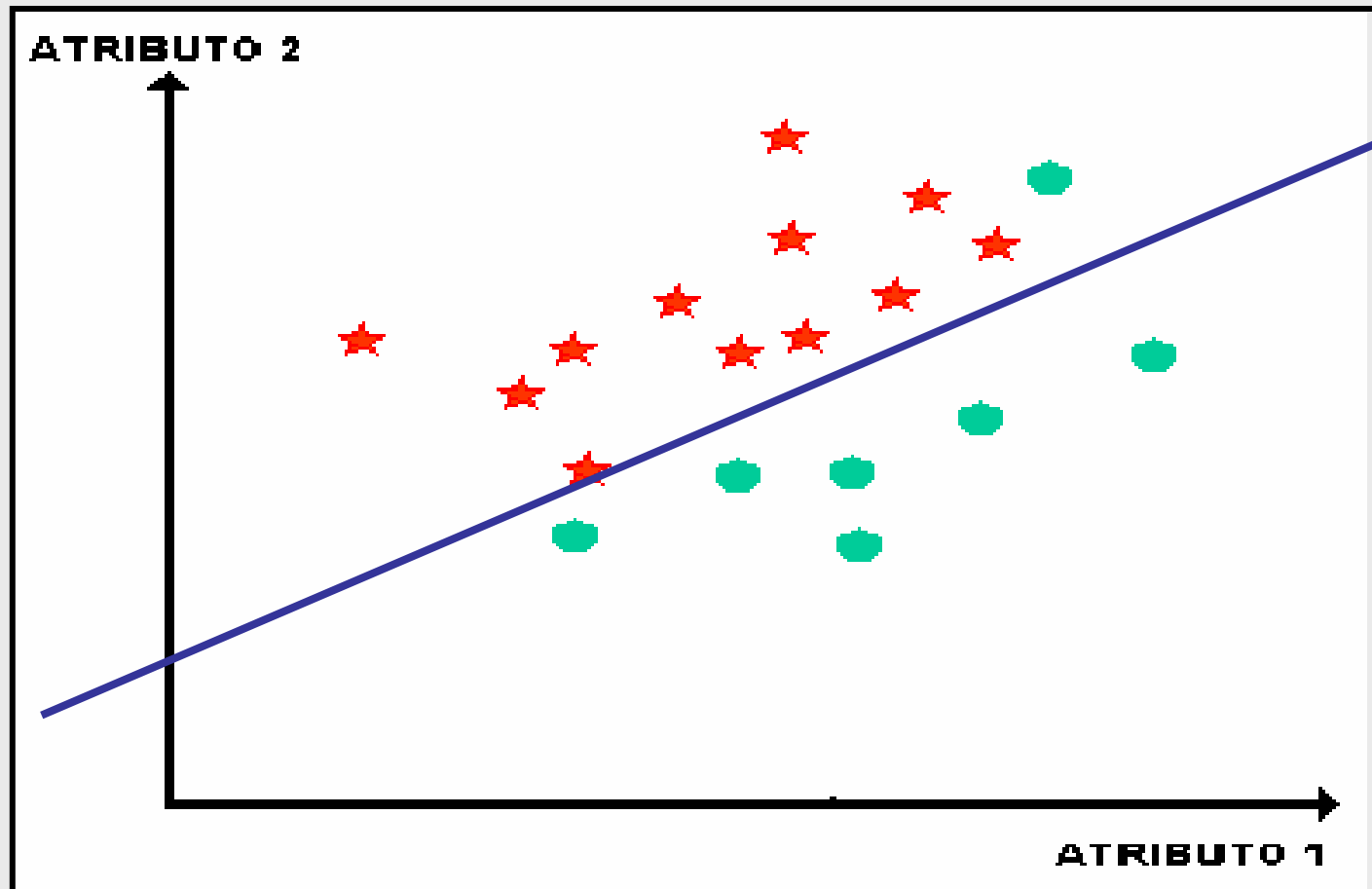
Limitaciones y problemas (4)



EN FORMA GRAFICA



SOLUCIÓN





MODELOS PREDICTIVOS

MAPAS AUTORGANIZATIVOS DE KOHONEN

Jaime Miranda

Departamento de Ingeniería Industrial
Universidad de Chile

IN47B

Ingeniería de Operaciones

Algunas nociones

- Sólo ocupan como información los atributos o variables de entrada.
- No se conocen las clases a priori.
- Se ocupan relaciones de similitud.

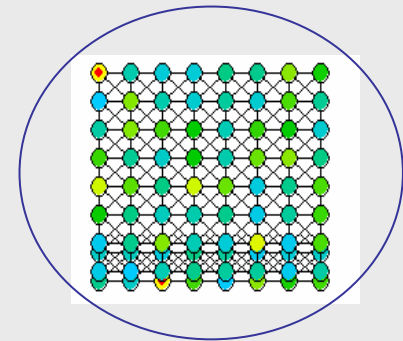
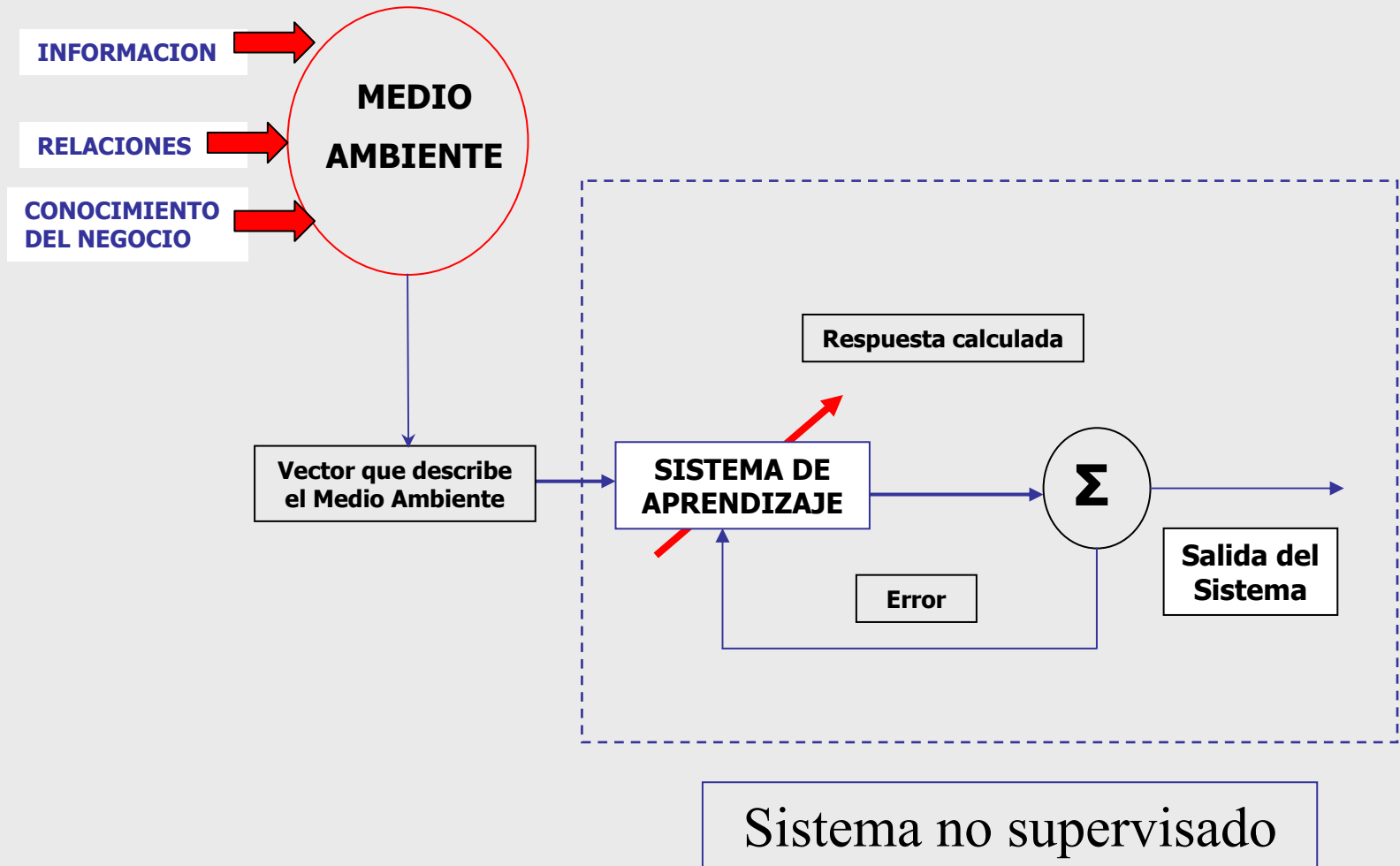
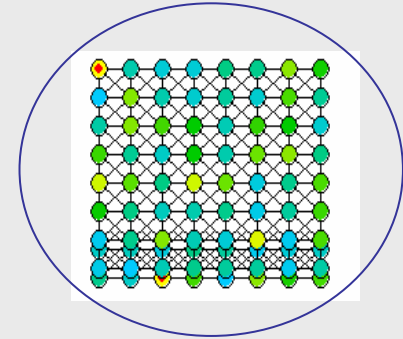


DIAGRAMA DE APRENDIZAJE NO SUPERVISADO



APLICACIONES

- **Marketing**
 - Segmentación de clientes
 - Ofertas focalizadas



FORTALEZAS

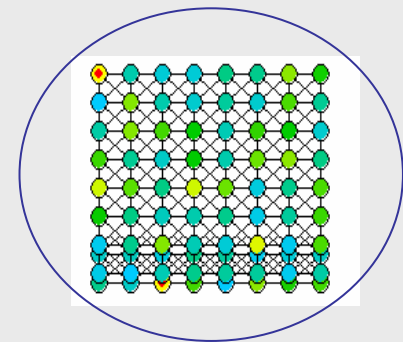
- Reduce la dimensionalidad del espacio.
- Los resultados son mas intuitivos y fáciles de entender e interpretar

DEBILIDADES

- Solo nos da una visión espacial de los resultados
- No trabajan bien con variables categóricas

Proviene de la cuantización vectorial

- Aproximación a la función de densidad de probabilidad $P(x)$ de las variables de entrada (x) usando un número finito de prototipo.
- La idea es relacionar un objeto con su prototipo mas cercano.
- Trata de buscar una aplicación topológicamente ordenada del espacio de entrada.
- Los hiperplanos de separación son lineales en trozos.



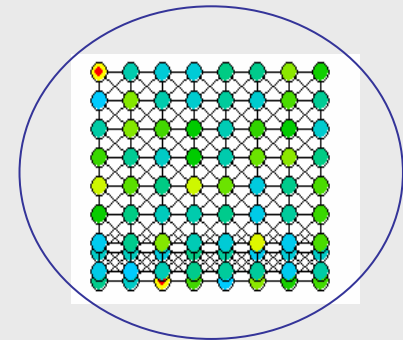
ALGUNAS NOCIONES...

Método no supervisado

Especializado para encontrar relaciones entre subconjuntos

Reduce la dimensionalidad del espacio de entradas

Tiene forma de grilla (2D) o caja (3D)



CLUSTERING

- Aproximación: es posible agrupar datos del conjunto de entrada, atendiendo a diferentes criterios con el fin de asignarle una clase a cada objeto del conjunto de entrada.

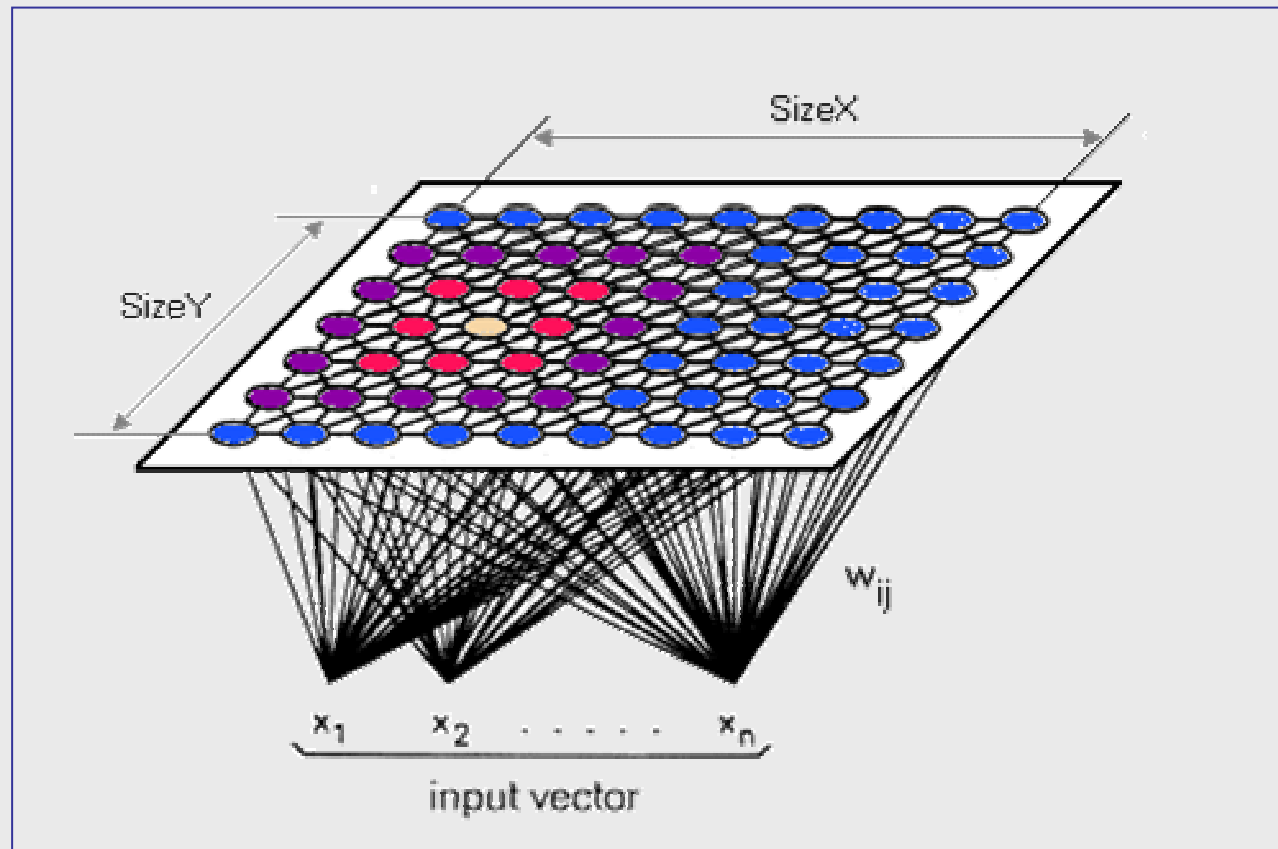
VISUALIZACIÓN

- Este agrupamiento, trata de mostrar en forma grafica las similitudes a través de la mantención de las características topológicas.

CLASIFICACIÓN:

- Una vez calibrado el mapa, o asignada algún tipo de etiqueta a cada clusters, se puede usar para clasificar sobre datos desconocidos.

EN FORMA GRAFICA...



1. Inicializar los prototipos

2. Encontrar neurona ganadora (BEST MATCHING)

$$m_i = m_i(0)$$

3. Modificar los vectores prototipos

$$\|x(t) - m_c(t)\| = \min \{ \|x(t) - m_i(t)\| \}$$

4. Volver a 3 hasta que no ocurran cambios significativos en el mapa **para $i \in N_c$**

$$m_i(t+1) = m_i(t) + \alpha [x(t) - m_i(t)]$$

$$m_i(t+1) = m_i(t)$$

Tasa de aprendizaje



El algoritmo básicamente realiza:

- Definir la estructura de la red (cantidad de neuronas y dimensionalidad).
- Inicializar neuronas aleatoriamente.
- Presentar un ejemplo y determinar la neurona ganadora.
- Modificar la neurona ganadora y las de la vecindad de acuerdo al ejemplo.
- Volver a 3) hasta que no se produzcan cambios significativos.

Dimensiones del mapa

- 2D ó 3D.
- Número de nodos y clases.

Geometría del mapa

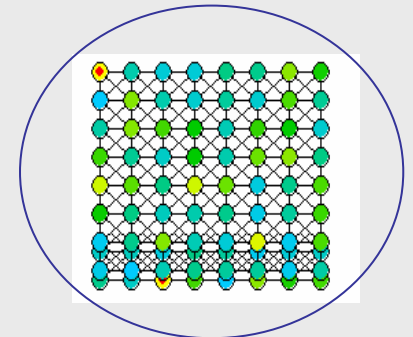
- Definición de la función de vecindad y radios de convergencia.

Tasa de aprendizaje

- Ajuste de los pesos.

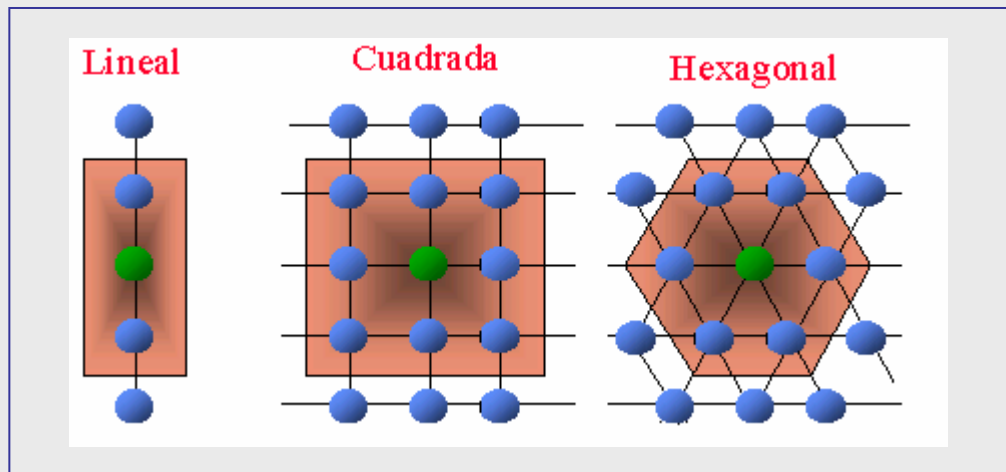
Épocas

- Número de veces que se “muestran” los datos al método.

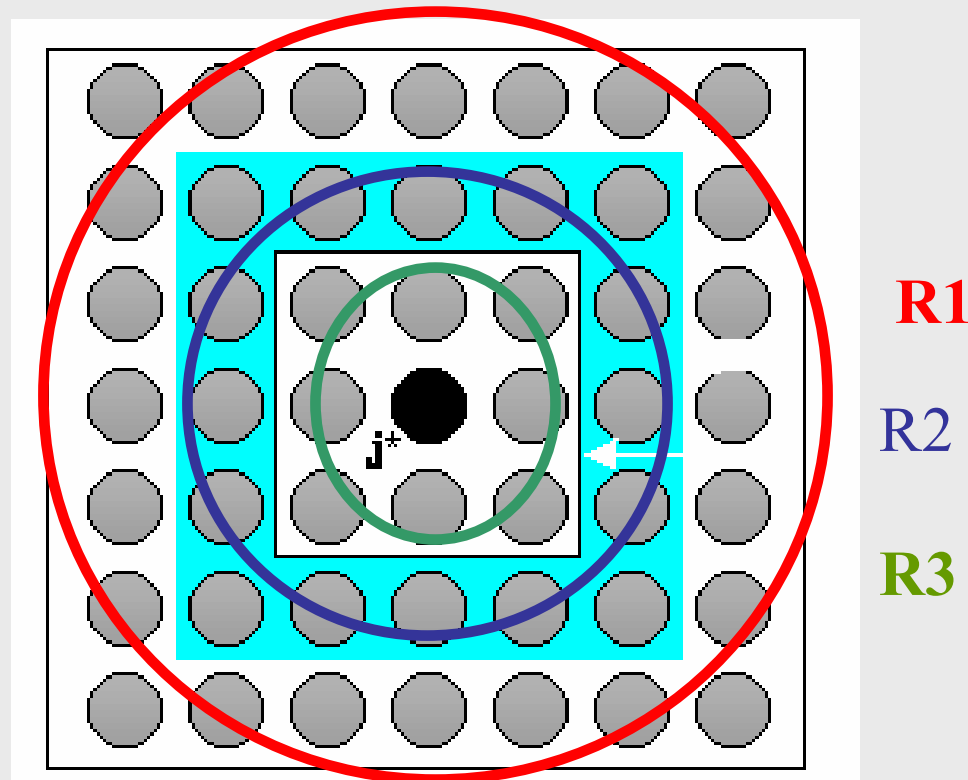


FUNCION DE VECINDAD

Nos dice que neuronas serán modificadas en cada iteración del algoritmo
Esta vecindad decrece a mitad que pasan la épocas.



FUNCIÓN DE VECINDAD CIRCULAR



En cada etapa de la iteración se disminuye el radio de pertenencia R_i

EJEMPLO...

PROBLEMA:

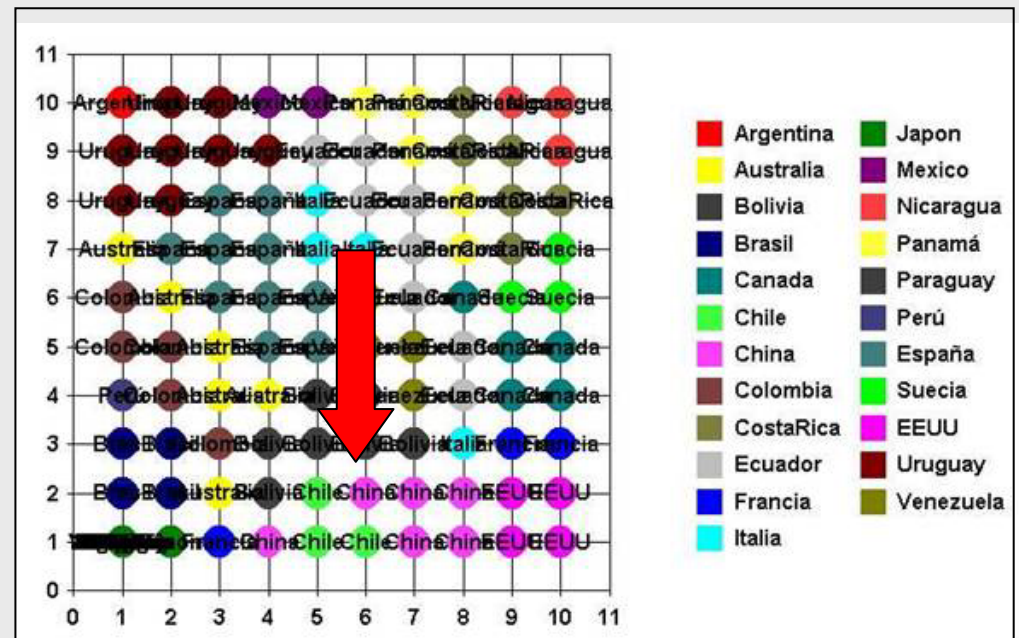
- Clasificar a un grupo de países a través de variables socioeconómicas.
- Conjunto de 23 países.
- Cada país esta caracterizado por 18 atributos (4 grupos).

ATRIBUTOS DE LOS PAISES	
CLASES	DESCRIPCIÓN
1	Consumo energético y condiciones ambientales
2	Desarrollo económico.
3	Desarrollo tecnológico.
4	Población y densidad demográfica.

RESULTADOS

Desarrollo económico

- Países cercanos
- China.
- Francia
- Australia
- Chile
- USA



RESULTADOS (2)

Desarrollo tecnológico

→ Países vecinos.

→ Argentina.

→ Venezuela.

→ Costa Rica

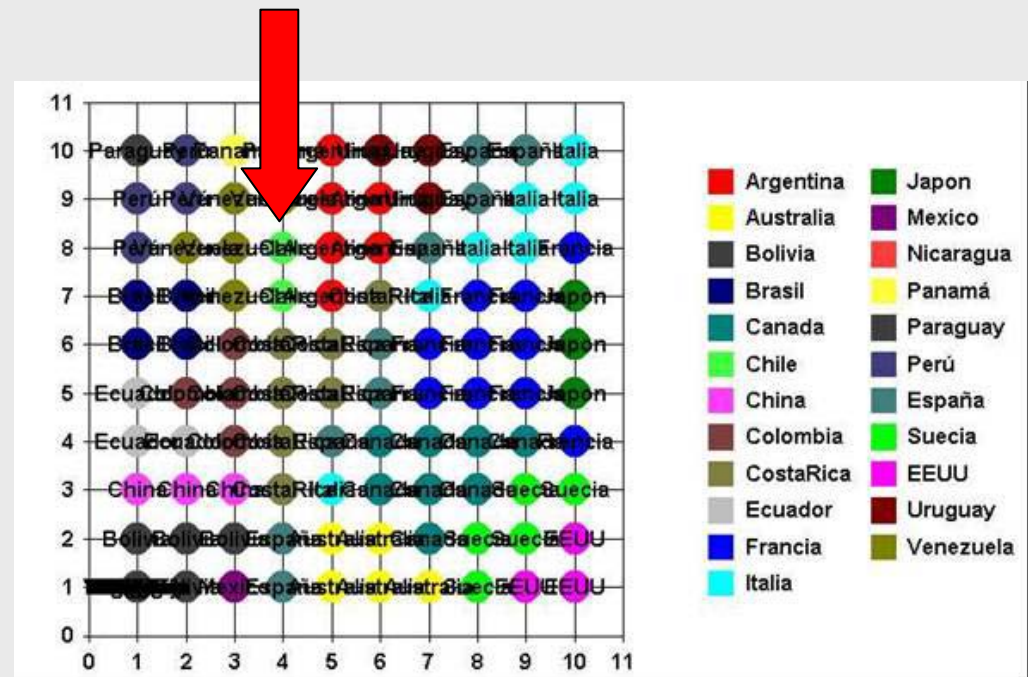
→ Uruguay

→ Otros países:

→ EEUU.

→ Suecia.

→ Australia



VENTAJAS

- Entrenamiento no supervisado.
- No necesita pares de Entrada/Salida, solo los de entrada.
- Disminuye la dimensionalidad del problema a algo “representable” en un espacio visualmente representable.

DESVENTAJAS

- Solo entrega información sobre la ubicación del espacio.
- Es necesario interpretar la información encontrada (Componentes principales).



MODELOS PREDICTIVOS

FUZZY-C-MEANS

Jaime Miranda

Departamento de Ingeniería Industrial
Universidad de Chile

IN47B

Ingeniería de Operaciones

Aplicaciones

→ Marketing

- Segmentación de clientes
- Ofertas focalizadas



Fortalezas

- No asume ninguna distribución estadística entre los datos.
- Los resultados son mas intuitivos y fáciles de entender e interpretar

Debilidades

- Son muy sensitivos a los valores fuera de rango (outliers)
- No trabajan bien con variables categóricas

Método de clustering que utiliza un algoritmo matemático.

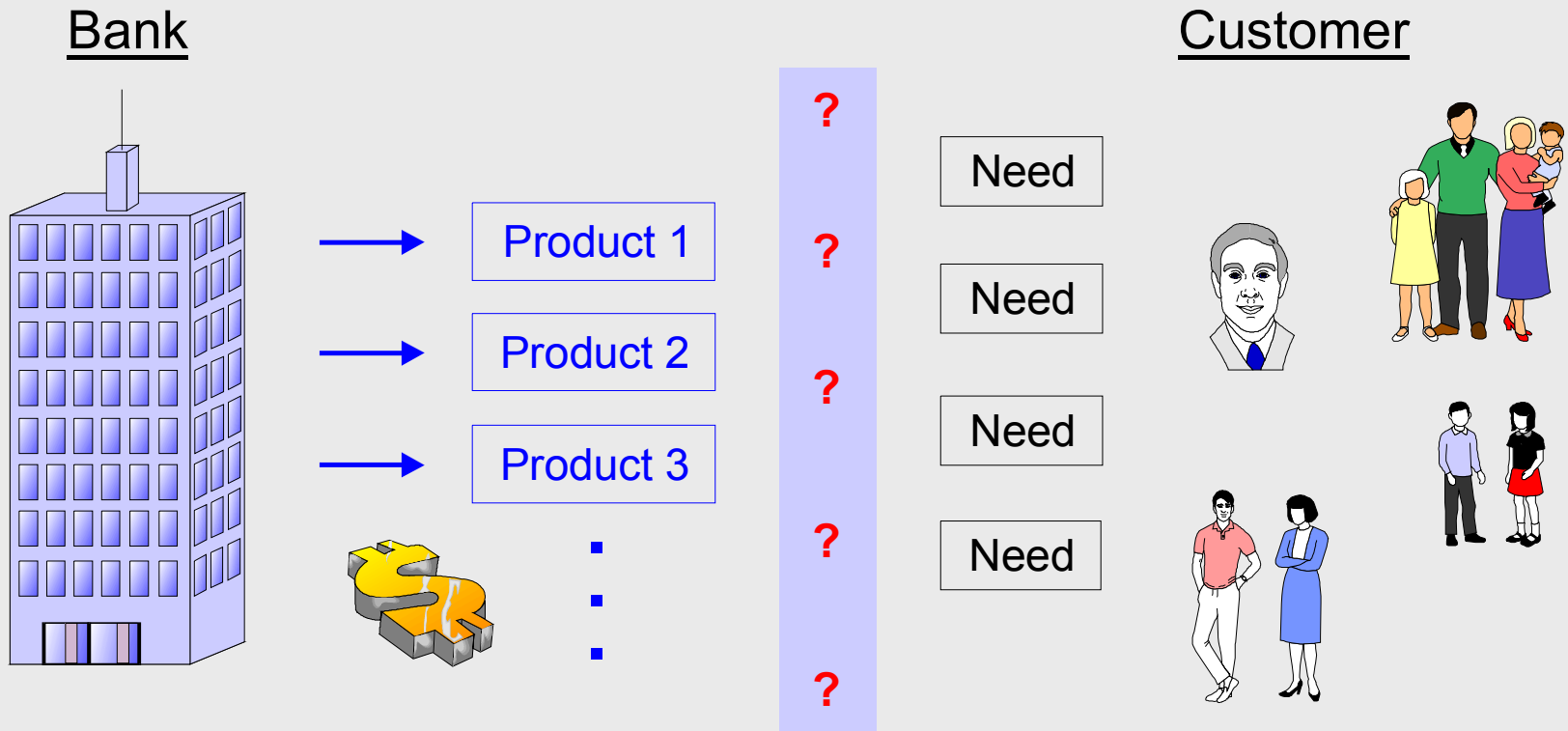
A través de la iteración este algoritmo determina el grado de pertenencia de un objeto a una clase determinada y el centro de esta.

Método no supervisado.

La base de este algoritmo es la minimización de la suma de las distancias de cada objeto al centro de la clase.

Se formula un problema de optimización cuya función objetivo es minimizar la suma de las distancias ponderadas por el grado de pertenencia del objeto.

SEGMENTACION DE CLIENTES



Which product for which (segment of) customers ?

Variables y Parámetros:

- n = Número de objetos.
- c = Numero de clases.
- X_i = Vector con valores de los atributos del objeto i , con $i = (1, \dots, n)$.
- C_j = Centro de la clase j , con $j = (1, \dots, c)$.
- $d^2(X_i, C_j)$ = Distancia entre el objeto i al centro de la clase j .



Variables y Parámetros:

→ u_{ij} = Grado de pertenencia del objeto i a la clase j .

→ $u_{i,j} \in [0,1]$

$$\sum_{j=1}^c u_{ij} = 1$$

$$U = (u_{ij})_{ij}$$

Matriz de pertenencia



Función Objetivo:
$$\min \sum_{i=1}^n \sum_{j=1}^c (u_{ij})^m d^2(x_i, c_j)$$

m = parámetro de difusibilidad

PASO 1

→ Determinar una matriz $U = (u_{ij})_{ij}$ que cumpla :

PASO 2

→ Determinar los centros de las clases:

$$\sum_{j=1}^c u_{ij} = 1$$

$$c_j = \frac{\sum_{i=1}^n u_{ij}^m x_i}{\sum_{i=1}^n u_{ij}^m}$$

PASO 3

→ Actualizar los grados de pertenencia:

$$u_{ij} = \frac{1}{\sum_{k=1}^c \left(\frac{d(x_i, c_j)}{d(x_i, c_k)} \right)^{\frac{2}{m-1}}}$$

PASO 4

→ Criterio de detención

$$\|U^{k+1} - U^k\| < \varepsilon$$

$$U^k = \text{matriz_en_la_iteración_}k$$

Número de clases

- En cuantos segmentos separo mi muestra.
- “Cluster Analysis”.
- Regla de los codos.

Coeficiente de difusividad (m)

- Rango (1 , ∞).

Umbral de detención

- El error de ajuste que el modelo acepta.

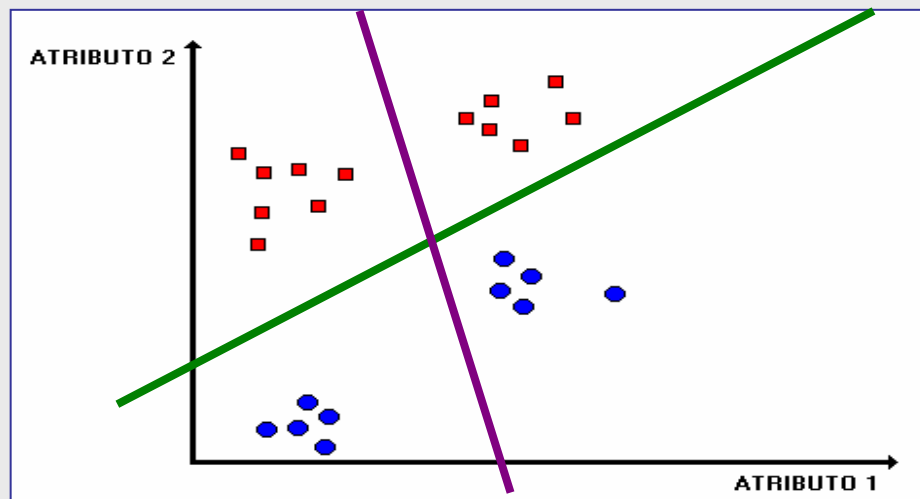
Valores iniciales de los centros

Este método a través de algún criterio especial nos indica el número óptimo de cluster's

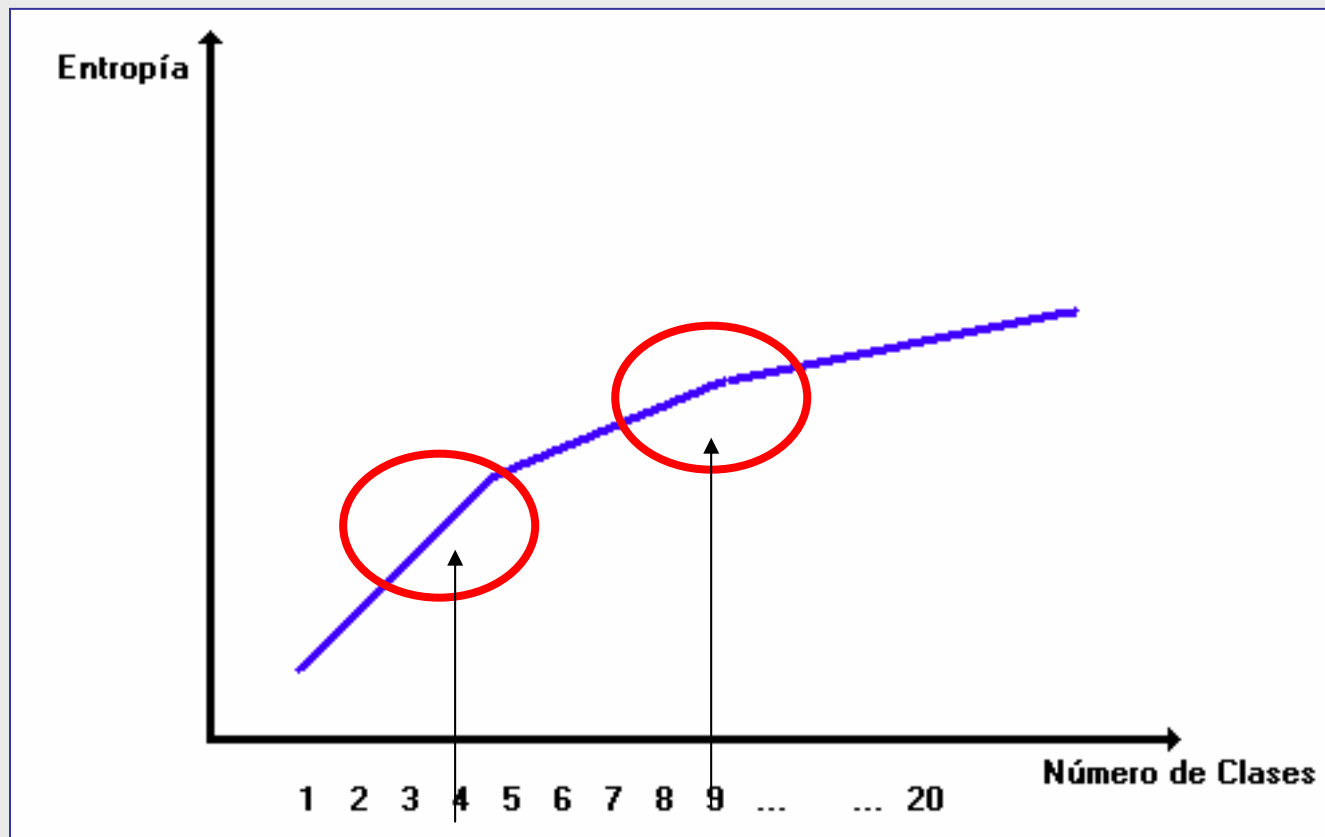
Uso de la entropía (Classification Entropy):

→ Regla de los codos

- mide la discontinuidad de la entropía respecto al numero de clases.



EJEMPLO



PARÁMETRO DE DIFUSIVIDAD (m)

Muestra que tan difusa será la clasificación.

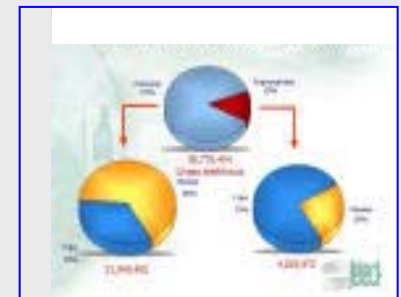
Rango $(1, \infty)$.

Valor mínimo = 1 , mínima difusividad (todos distintos).

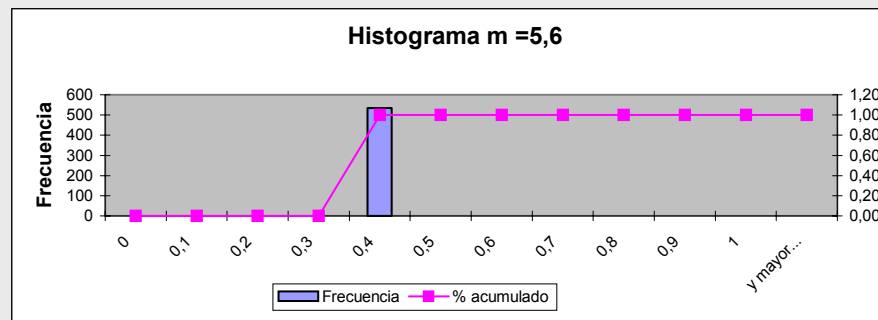
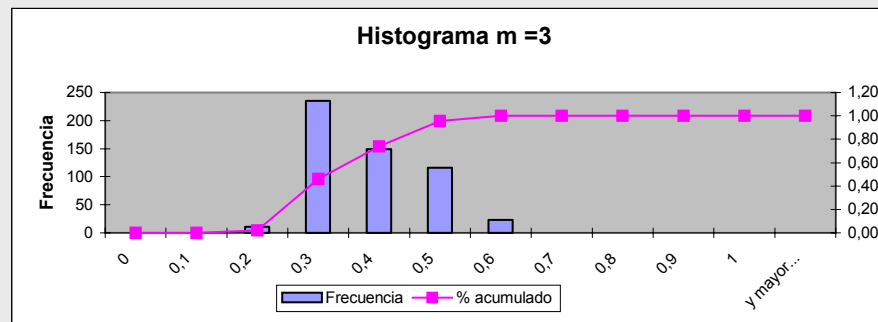
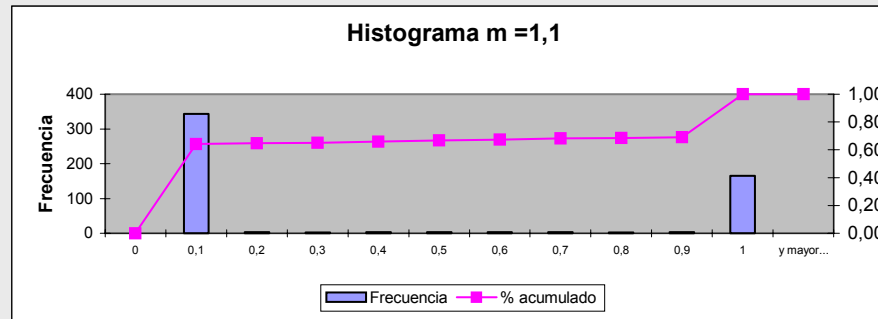
Valor máximo = ∞ máxima difundida (todos iguales).

Valores pequeños da clasificaciones más discretas.

Al aumentar la clasificación se hace más difusa.



PARÁMETRO DE DIFUSIVIDAD (m)





MODELOS PREDICTIVOS

FUZZY-C-MEANS

Jaime Miranda

Departamento de Ingeniería Industrial
Universidad de Chile

IN47B

Ingeniería de Operaciones