

2022

20 / 10 / 2022

# Resumen sesión teoría 20/10

SISTEMAS INTELIGENTES  
ADRIAN UBEDA TOUATI 50771466R

## Contenido

Arboles de decisión.....	2
Entropía condicionada .....	2
Algoritmo ID3.....	3
Extensión del algoritmo .....	4

## Arboles de decisión

### Introducción

Un árbol de decisiones es una estructura que clasifica los vectores de atributos, establece en qué orden testear los atributos para conseguir la clasificación del vector de entrada.

Desde los ejemplos se extrae el árbol de decisión

Entropía: Medida del grado de incertidumbre asociado a una distribución de probabilidad.

Si la entropía es 0 todo este concentrado en un punto, cuando más entropía, más distribución en la concentración.

$$E(S) = \sum_{i \in C} -p_i \log_2 p_i$$

Cuanto mayor es el valor de la entropía, menos información tenemos.

### Entropía condicionada

Muestra sin sesgo, significa que lo que dice el conjunto, nos informa de la realidad en todo su tamaño.

$v_i$	$\text{Prob}(X = v_i)$	$E(Y   X = v_i)$
Math	0.5	1
History	0.25	0
CS	0.25	0

Por lo que habrá siempre 50% de matemáticos en la residencia, sean 5 alumnos o 1000

X	Y
Math	Yes
Hist.	No
CS	Yes
Math	No
Math	No
CS	Yes
Hist.	No
Math	Yes

Como las respuestas entre los alumnos de Hist y entre los alumnos de CS no varían, la entropía para las 2 materias es 0, por lo que es en una distribución pico

La ganancia de la información es la diferencia entre  $E(Y)$  y  $E(Y | X)$

$$IG(Y | X) = E(Y) - E(Y | X)$$

En el caso de la película Gladiador:

$$E(Y) = 1$$

$$E(Y|X) = 0.5$$

$$IG(Y | X) = 1 - 0.5 = 0.5$$

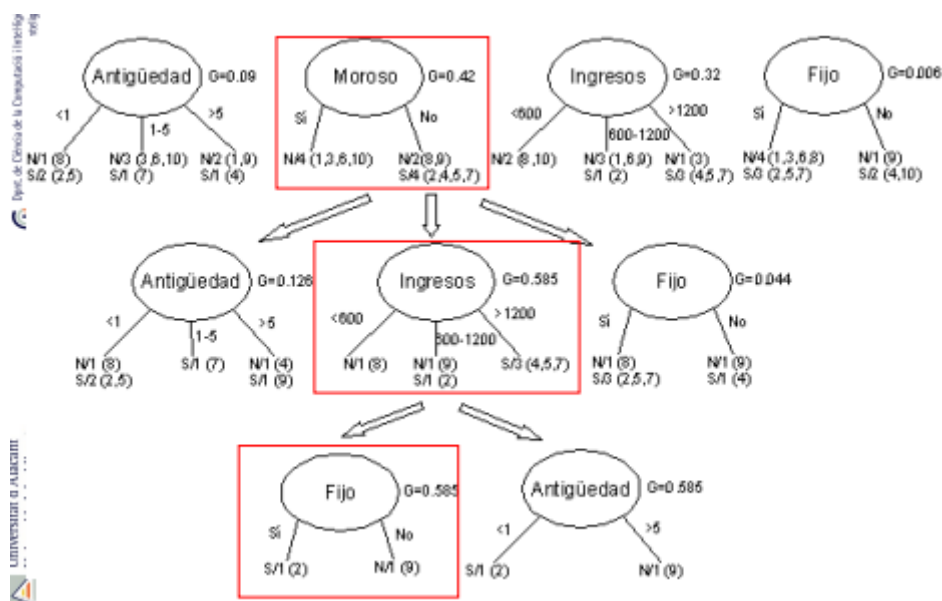
### Algoritmo ID3

```

Algoritmo ID3(ejemplos, atributos) {
  Si atributos =  $\emptyset$  o MISMACLASE(ejemplos) {
     $C \leftarrow$  CLASEMAYORITARIA(ejemplos)
     $N \leftarrow$  CREARNODOHOJA( $C$ )
  }
  Sino {
     $a_{max} \leftarrow \max_{A \in \text{atributos}} G(\text{ejemplos}, A)$ 
     $N \leftarrow$  CREARNODO( $a_{max}$ )
    Para cada  $v_i \in \text{VALORES}(a_{max})$  {
       $\text{ejemplos}_{v_i} \leftarrow \{\text{elementos de ejemplos con valor } v_i \text{ para } a_{max}\}$ 
      AÑADIRHIJO( $N$ , ID3( $\text{ejemplos}_{v_i}$ ,  $\text{atributos} - a_{max}$ ))
    }
  }
  Devolver  $N$ 
}

```

Para formar el árbol de decisión, primero debemos ver que atributo posee más información, y después preguntar en los siguientes nodos sobre los demás atributos no seleccionados. En el caso del ejemplo, daría igual que última pregunta hiciésemos, puesto que dan el mismo nivel de información



Si un nodo tiene entropía de 0 entonces es un nodo hoja, no tendrá más hijos, igualmente, si no quedan más atributos, será un nodo hoja también

En el caso de procesar el algoritmo y ver que no es capaz de llegar a hojas, significa que los datos proporcionados no son fiables, y por lo tanto tienen sesgos, tendríamos que cambiar el muestreo.

### Extensión del algoritmo

Es importante saber que el algoritmo ID3 trabaja con atributos discretos, En el caso que se usen atributos continuos hay que descomponerlos en rangos.

Ej:

clase.

				825	950	1150					
Ejemplo	8	10	6	2	1	9	3	5	4	7	
Ingresos	450	530	650	800	850	1050	1250	1400	1600	3000	
Crédito	no	no	no	no	si	no	si	si	si	si	

En el caso de atributos con gran número de valores se forman grupos pequeños de ejemplos que pueden ser homogéneos por casualidad. Debe introducirse un elemento corrector que penalice atributos con un elevado número de valores (ganancia normalizada)

$$G_N(S, A) = \frac{G(S, A)}{\sum_{v_i \in V(A)} -p_{v_i} \log_2 p_{v_i}}$$