Dpnt. de Ciència de la Computació i Intel·ligència Artificial
Dpto. de Ciencia de la Computación e Inteligencia Artificial
Universitat d'Alacant
Universidad de Alicante

TEMA 6: Arboles de Decisión

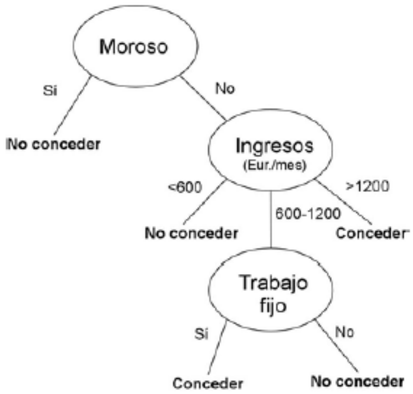
Toma de decisiones en sistemas probabilísticos

Dpnt. de Ciència de la Computació i Intel·ligència Artificial
Dpto. de Ciencia de la Computación e Inteligencia Artificial
Universitat d'Alacant
Universidad de Alicante

- Árboles de decisión
  - Planteamiento del problema
  - Ejemplo: Concesión de créditos
  - Entropía y Ganancia de Información
- Algoritmo ID3
  - Algoritmo recursivo
  - Aplicación al ejemplo
  - Consideración de atributos numéricos
  - Atributos con un gran número de valores

Árboles de decisión

- Características:
  - ➢ Estructura para clasificación de vectores de atributos.
  - ➢ Establece en qué orden testar los atributos para conseguir la clasificación del vector de entrada.
  - ➢ Para componer dicho orden se eligen primero aquellos atributos que mejor ganancia de información prometen a efectos de descubrir la clase del vector de entrada.
  - ➢ Es interesante aprenderlos a partir de un conjunto de vectores



Ejemplo "Concesión de créditos"

Cliente	Moroso	Antigüedad (años)	Ingresos (Eur./mes)	Trab.fijo	Conceder
1	sí	>5	600-1200	sí	no
2	no	<1	600-1200	sí	sí
3	sí	1-5	>1200	si	no
4	no	>5	>1200	no	sí
5	no	<1	>1200	sí	sí
6	sí	1-5	600-1200	si	no
7	no	1-5	>1200	sí	sí
8	no	<1	<600	sí	no
9	no	>5	600-1200	no	no
10	no	1-5	<600	no	no

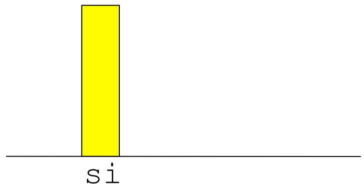
- Aprendizaje:
  - ➢ ¿Por qué atributo comenzar primero?
  - ➢ Esquema voraz: Elegir uno y filtrar recursivamente.

Entropía

- Definición:
  - Medida del grado de incertidumbre asociado a una distribución de probabilidad.
  - En una distribución uniforme, todos los valores son igualmente probables Pi = 1/N y por tanto la entropía es máxima, lo cual indica máxima incertidumbre.
  - Por el contrario, en una distribución pico en la que Pi = 1 y Pj=0, para todo j≠i la entropía es mínima lo cual indica mínima incertidumbre o sea máxima información.

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

-0.5log2(0.5) - 0.5log2(0.5) = 1



-1.0log2(1.0) - 0.0log2(0.0) = 0

Entropía condicionada

- Definición:
  - Entropía de la distribución de Y condicionada a X.
  - Una entropía condicionada menor que E(Y) indica que el conocimiento de X mejora la información que se dispone sobre Y

$E(Y \mid X) = \sum_j Prob(X = v_j)\, E(Y \mid X = v_j)$

v_j	$Prob(X = v_j)$	$E(Y \mid X = v_j)$
Math	0.5	1
History	0.25	0
CS	0.25	0

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

E(Y) = 1

E(Y|X) = 0.5*1 + 0.25*0 + 0.25*0 → $\boxed{E(Y|X) = 0.5}$

Ganancia de información

- Definición:
  - Medida de cuanto ayuda el conocer el valor de una variable aleatoria X para conocer el verdadero valor de otra Y.
  - En nuestro caso, X es un atributo de un ejemplo dado mientras que Y es la clase a la que pertenece el ejemplo.
  - Una alta ganancia implica que el atributo X permite reducir la incertidumbre de la clasificación del ejemplo de entrada.

$$IG(Y \mid X) = E(Y) - E(Y \mid X)$$

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

E(Y) = 1

E(Y|X) = 0.5

IG(Y | X) = 1 – 0.5 = 0.5

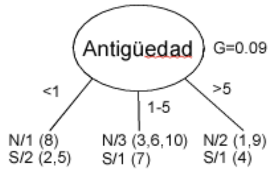
Algoritmo recursivo

Algoritmo $\mathrm{ID3}(ejemplos, atributos)$ {
  Si $atributos = \emptyset$ **o** $\textsc{MismaClase}(ejemplos)$ {
    $C \leftarrow \textsc{ClaseMayoritaria}(ejemplos)$
    $N \leftarrow \textsc{CrearNodoHoja}(C)$
  }
  Sino {
    $a_{max} \leftarrow \text{máx}_{\forall A \in atributos}\, \mathrm{G}(ejemplos, A)$
    $N \leftarrow \textsc{CrearNodo}(a_{max})$
    Para cada $v_i \in \textsc{Valores}(a_{max})$ {
      $ejemplos_{v_i} \leftarrow$ {elementos de ejemplos con valor v_i para a_{max}}
      $\textsc{AñadirHijo}(N, \mathrm{ID3}(ejemplos_{v_i}, atributos - a_{max}))$
    }
  }
  Devolver N
}

Aplicación al ejemplo

- Entropía inicial:
  - Aplicando la ecuación de entropía a los datos de entrada del ejemplo tenemos:

```
        E(S)=
-0.4log2(0.4)-0.6log2(0.6)
         = 0.971
```



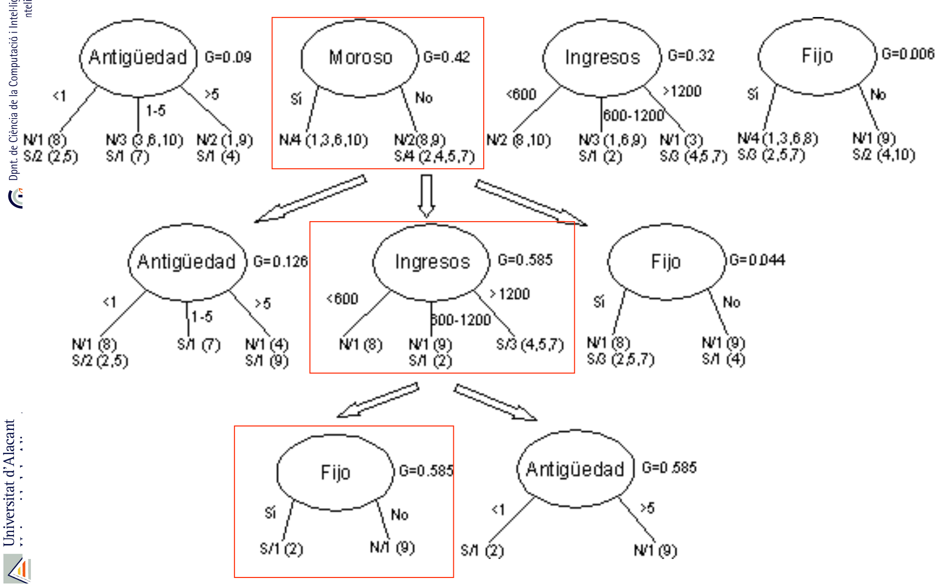
- Para cada atributo (Antigüedad, Moroso, Ingresos, Fijo), calculamos la ganancia de información que obtenemos al seleccionar cada uno de ellos

```
Prob(S<1)=0.3,Prob(S1-5)=0.4,Prob(S>5)=0.3

E(S<1)  = -2/3log2(2/3)-1/3log2(1/3)= 0.9183
E(S1-5)= -1/4log2(1/4)-3/4log2(3/4)= 0.811
E(S>5)  = -1/3log2(1/3)-2/3log2(2/3)= 0.9183

      E(S<1)*0.3 = 0.2755
      E(S1-5)*0.4 = 0.3244
      E(S>5)*0.3 = 0.2755
H(Conceder | Antigüedad) =
      0.2755 + 0.3244 + 0.2755 = 0.8754
Ganancia = 0.971 - 0.8754 = 0.09
```


Aplicación al ejemplo



Extensiones del algoritmo

- Extensiones:
  - Atributos numéricos: ID3 sólo trabaja con atributos discretos. Si se usan atributos continuos hay que descomponerlos en rangos. Para ello se ordenan los ejemplos según el valor y se toman como puntos límite los puntos medios de aquellos en que se cambie de 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	sí	no	sí	sí	sí	sí

  - 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):

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

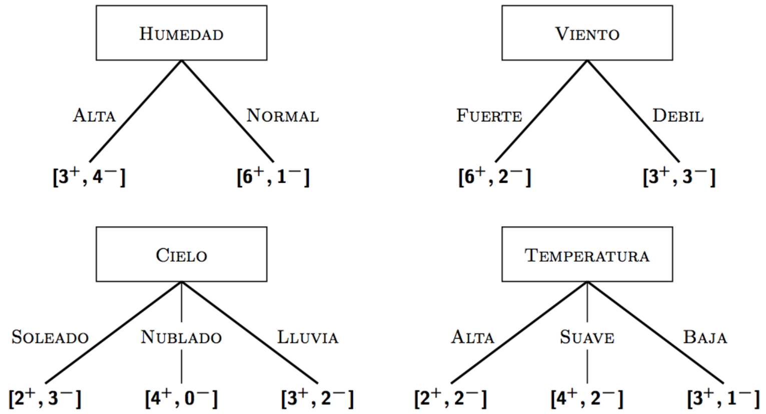
  - Sobre-entrenamiento. Comprobación de capacidad de generalización.

Ejercicios

Objetivo: Dado el conjunto de entrenamiento, aprender el concepto "Días en los que se juega al tenis" obteniendo el árbol de decisión mediante el algoritmo ID3

Ej.	Cielo	Temperatura	Humedad	Viento	JugarTenis
D_1	Soleado	Alta	Alta	Débil	-
D_2	Soleado	Alta	Alta	Fuerte	-
D_3	Nublado	Alta	Alta	Débil	+
D_4	Lluvia	Suave	Alta	Débil	+
D_5	Lluvia	Baja	Normal	Débil	+
D_6	Lluvia	Baja	Normal	Fuerte	-
D_7	Nublado	Baja	Normal	Fuerte	+
D_8	Soleado	Suave	Alta	Débil	-
D_9	Soleado	Baja	Normal	Débil	+
D_{10}	Lluvia	Suave	Normal	Débil	+
D_{11}	Soleado	Suave	Normal	Fuerte	+
D_{12}	Nublado	Suave	Alta	Fuerte	+
D_{13}	Nublado	Alta	Normal	Débil	+
D_{14}	Lluvia	Suave	Alta	Fuerte	-

Ejercicios

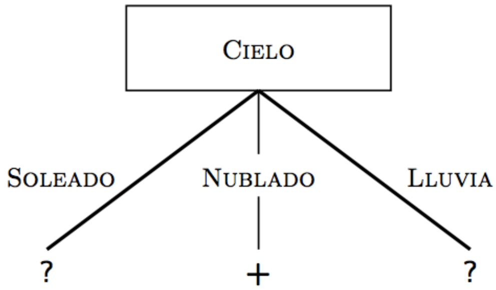


Ejercicios

- Entropía inicial: $\mathbf{Ent([9^+, 5^-]) = 0{,}94}$
- Selección del atributo para el nodo raíz:
  - $\mathbf{Ganancia(D,}\textsc{Humedad}\mathbf{) =}$
    $\mathbf{0{,}94 - \frac{7}{14} \cdot Ent([3^+, 4^-]) - \frac{7}{14} \cdot Ent([6^+, 1^-]) = 0{,}151}$
  - $\mathbf{Ganancia(D,}\textsc{Viento}\mathbf{) =}$
    $\mathbf{0{,}94 - \frac{8}{14} \cdot Ent([6^+, 2^-]) - \frac{6}{14} \cdot Ent([3^+, 3^-]) = 0{,}048}$
  - $\mathbf{Ganancia(D,}\textsc{Cielo}\mathbf{) =}$
    $\mathbf{0{,}94 - \frac{5}{14} \cdot Ent([2^+, 3^-]) - \frac{4}{14} \cdot Ent([4^+, 0^-])}$
    $\mathbf{- \frac{5}{14} \cdot Ent([3^+, 2^-]) = 0{,}246}$ (mejor atributo)
  - $\mathbf{Ganancia(D,}\textsc{Temperatura}\mathbf{) =}$
    $\mathbf{0{,}94 - \frac{4}{14} \cdot Ent([2^+, 2^-]) - \frac{6}{14} \cdot Ent([4^+, 2^-])}$
    $\mathbf{- \frac{4}{14} \cdot Ent([3^+, 1^-]) = 0{,}02}$
- El atributo seleccionado es CIELO

Ejercicios

Árbol parcialmente construido:



Sistemas Inteligentes

Ejercicios

- Selección del atributo para el nodo CIELO=SOLEADO
- $\mathbf{D}_{\text{SOLEADO}} = \{\mathbf{D}_1, \mathbf{D}_2, \mathbf{D}_8, \mathbf{D}_9, \mathbf{D}_{11}\}$ con entropía $\mathbf{Ent([2^+, 3^-]) = 0{,}971}$
  - $\mathbf{Ganancia(D_{\text{SOLEADO}}, \text{HUMEDAD}) =}$ $\mathbf{0{,}971 - \frac{3}{5} \cdot 0 - \frac{2}{5} \cdot 0 = 0{,}971}$ (mejor atributo)
  - $\mathbf{Ganancia(D_{\text{SOLEADO}}, \text{TEMPERATURA}) =}$ $\mathbf{0{,}971 - \frac{2}{5} \cdot 0 - \frac{2}{5} \cdot 1 - \frac{1}{5} \cdot 0 = 0{,}570}$
  - $\mathbf{Ganancia(D_{\text{SOLEADO}}, \text{VIENTO}) =}$ $\mathbf{0{,}971 - \frac{2}{5} \cdot 1 - \frac{3}{5} \cdot 0{,}918 = 0{,}019}$
- El atributo seleccionado es HUMEDAD

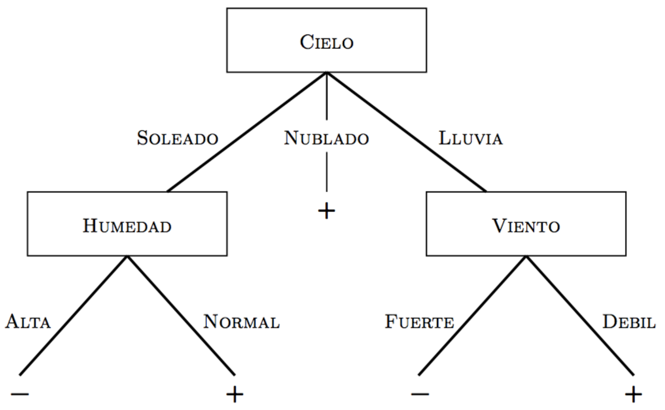
Dpnt. de Ciència de la Computació i Intel·ligència Artificial
Dpto. de Ciencia de la Computación e Inteligencia Artificial
Universitat d'Alacant
Universidad de Alicante

Ejercicios

- Selección del atributo para el nodo CIELO=LLUVIA:
- $D_{\text{LLUVIA}} = \{D_4, D_5, D_6, D_{10}, D_{14}\}$ con entropía $Ent([3^+, 2^-]) = 0{,}971$
  - $Ganancia(D_{\text{LLUVIA}}, \text{HUMEDAD}) = 0{,}971 - \frac{2}{5} \cdot 1 - \frac{3}{5} \cdot 0{,}918 = 0{,}820$
  - $Ganancia(D_{\text{LLUVIA}}, \text{TEMPERATURA}) = 0{,}971 - \frac{3}{5} \cdot 0{,}918 - \frac{2}{5} \cdot 1 = 0{,}820$
  - $Ganancia(D_{\text{LLUVIA}}, \text{VIENTO}) = 0{,}971 - \frac{3}{5} \cdot 0 - \frac{2}{5} \cdot 0 = 0{,}971$ (mejor atributo)
- El atributo seleccionado es VIENTO

Ejercicios

- Árbol finalmente aprendido:



Bibliografía

- Escolano et al. Inteligencia Artificial. Thomson-Paraninfo 2003. Capítulo 4.
- Mitchel, Machine Learning. McGraw Hill, Computer Science Series. 1997
- Cover, Thomas, Information Theory. Wiley & Sons, New York 1991