

Aprendizaje

1. Preliminares
2. Algoritmos genéticos y redes neuronales
3. **Inducción de árboles clasificadores**
4. Inducción de reglas
5. Minería de datos

Inducción de árboles de clasificación

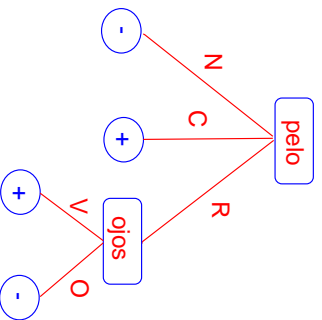
- Árboles de clasificación
- Procedimiento CLS
- Valor informativo de un atributo
- Construcción del árbol
- Algoritmo ID3 y derivados

Árboles de clasificación: ejemplo

Clasificadores:

- Una red neuronal de tipo MLP, una vez entrenada, clasifica objetos caracterizados por atributos numéricos
- Un árbol de clasificación, una vez construido, clasifica objetos caracterizados por atributos nominales (o binarios)

Ejemplo:



- Si (v(pelo) = «negro») entonces «-»
- Si (v(pelo) = «castaño») entonces «+»
- Si (v(pelo) = «rubio») y (v(ojos) = «verdes») entonces «+»
- Si (v(pelo) = «rubio») y (v(ojos) = «oscuros») entonces «-»

Inducción de árboles de clasificación

- Años 60: modelos informales de psicología cognitiva: elección de atributos más significativos para formar conceptos
- Hunt et al. (1966): CLS (Concept Learning System)
- CHAID (Chi-squared Automatic Interaction Detection) (Hartigan, 1975)
- CART (Classification And Regression Trees) (Friedman, 1977; Briemen *et al.* 1984)
- ID3 (Iterative Dichotomizer) (Quinlan, 1979)
- C4.5 (Quinlan, 1993) y C5.0 (comercial)



Procedimiento CLS

Dado $E = \{e\}$, con $e = \{\langle \vec{x}_e, f(\vec{x}_e) \rangle\}$
 $= \{\langle v(A_1), v(A_2), \dots, v(A_n), C_j \rangle\}$

genera árbol de clasificación (clasificador) *mínimo*
(nodos: atributos; arcos: valores)

procedimiento CLS(E)

si todo e está en C_j

terminar con hoja " C_j "

si no

crear nodo = atributo "mejor";

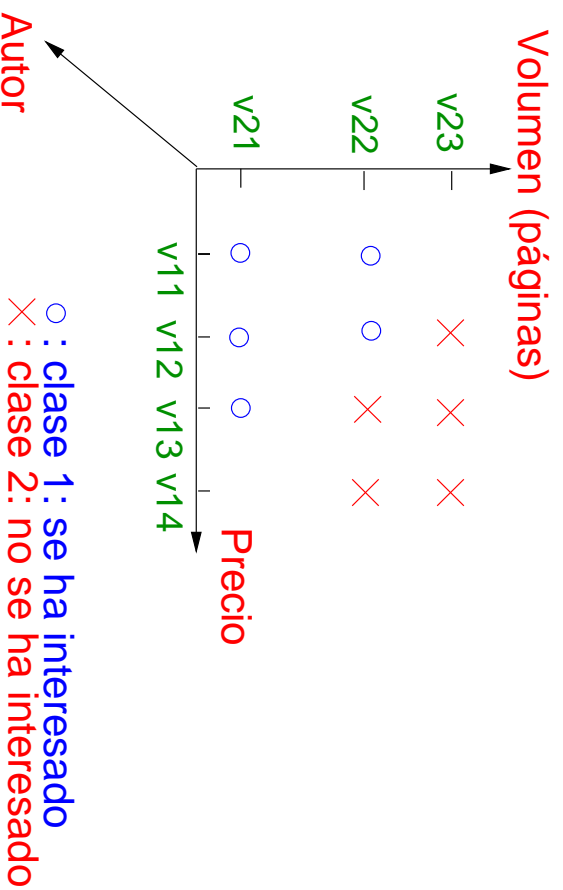
particionar E por valores atributo;

para cada partición

CLS(partición)

¿Cuál es el mejor atributo?

El que más información da



¿Es mejor Precio o Volumen?

Entropía

¿Por qué es mejor Volumen?

Porque da más información (reduce más la incertidumbre)

Medida de la incertidumbre: entropía

$$H = - \sum_k (p_k \times \lg_2 p_k)$$

p_k : probabilidad de que un ejemplo esté en la clase k :

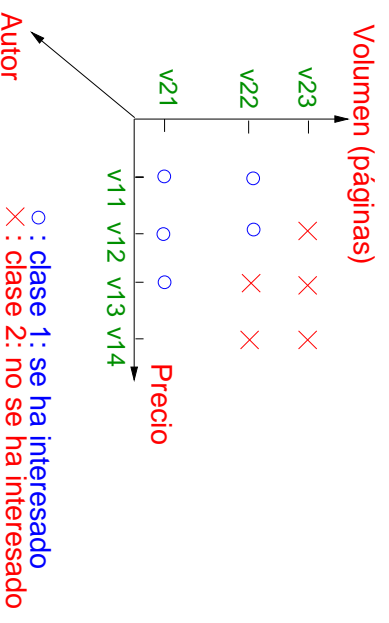
$$p_k = \frac{n_k}{\sum_k n_k}$$

con n_k = número de ejemplos en la clase k

Si hay dos clases, $0 \leq H \leq 1$

Entropía en el ejemplo

Entropía inicial:
 $H_0 = -\left(\frac{5}{10} \cdot \lg \frac{5}{10}\right) - \left(\frac{5}{10} \cdot \lg \frac{5}{10}\right) = 1$ bit



Entropía *media* tras conocer el valor de Precio:

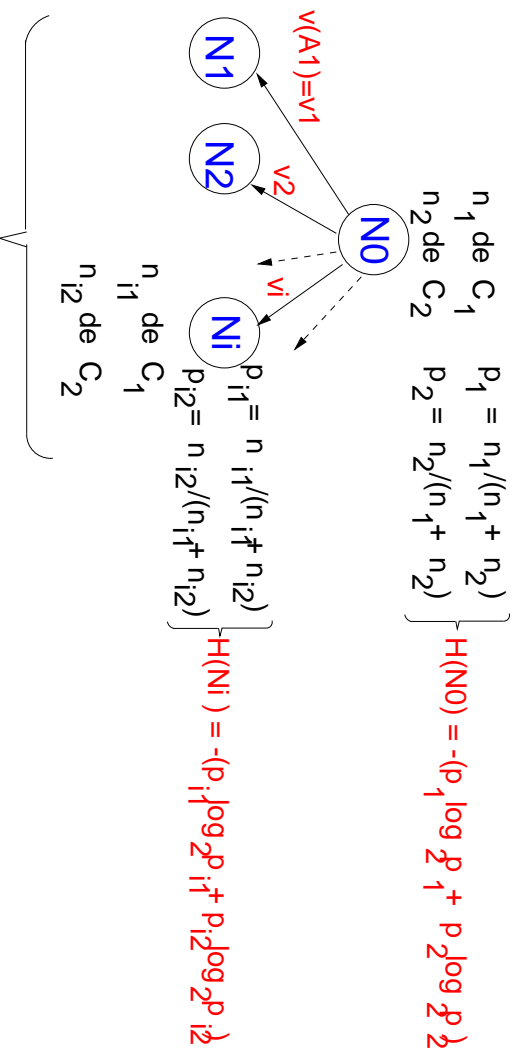
$$H_p = \frac{2}{10} \cdot 0 + \frac{3}{10} \cdot \left(-\frac{1}{3} \lg \frac{1}{3} - \frac{2}{3} \lg \frac{2}{3}\right) + \frac{3}{10} \cdot \left(-\frac{2}{3} \lg \frac{2}{3} - \frac{1}{3} \lg \frac{1}{3}\right) + \frac{2}{10} \cdot 0 = 0,55$$
 bits

Entropía *media* tras conocer el valor de Volumen:

$$H_v = \frac{3}{10} \cdot 0 + \frac{4}{10} \cdot \left(-\frac{1}{2} \lg \frac{1}{2} - \frac{1}{2} \lg \frac{1}{2}\right) + \frac{3}{10} \cdot 0 = 0,4$$
 bits

- Ganancia de información por Precio: $H_0 - H_p = 0,45$ bits
- Ganancia de información por Volumen: $H_0 - H_v = 0,6$ bits

Valor informativo de un atributo



$$H(N_0|A1) = \sum_i \Pr(v(A1) = v_i) * H(N_i) = \sum_i [(n_{i1} + n_{i2}) / (n_1 + n_2)] * H(N_i)$$

Reducción de entropía = información ganada por A1 =

$$H(N_0) - H(N_0|A1)$$

Construcción del árbol (1)

Atrib.	Valores	Clases
talla	Alto, Bajo	+, -
pelo	Negro, Cast., Rubio	
ojos	Verdes, Oscuros	

Inicialmente:

$$p^+ = 3/8; p^- = 5/8;$$

$$H(N_0) = -(\frac{3}{8} \cdot \log_2 \frac{3}{8} + \frac{5}{8} \cdot \log_2 \frac{5}{8}) = 0,954 \text{ bits}$$

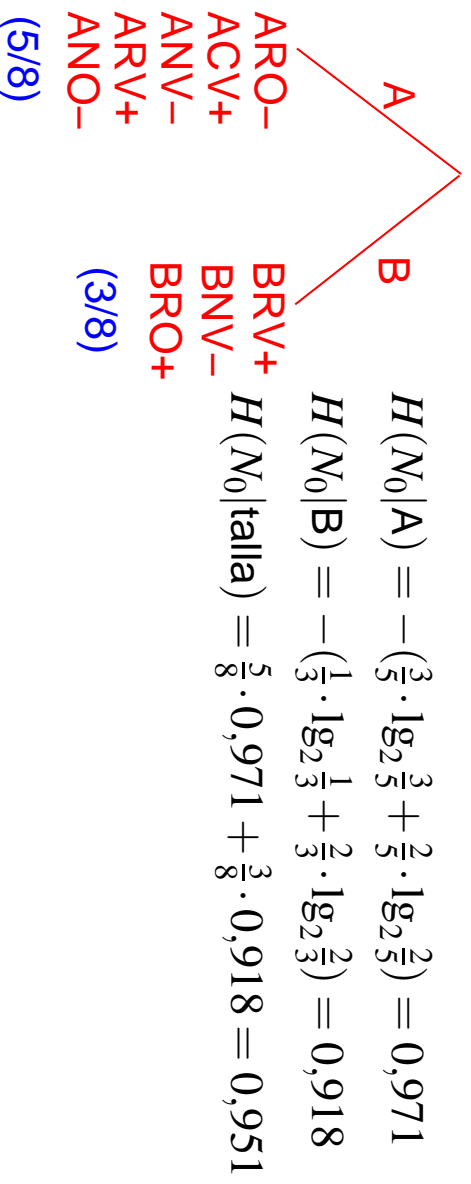
talla	pelo	ojos	clase
B	R	V	+
A	R	O	-
A	C	V	+
B	N	V	-
A	N	V	-
A	R	V	+
A	N	O	-
B	R	O	-

Ejemplo de Quinlan (1983). Trivial, pero ilustra aplicaciones reales: descubrir criterio de selección, factores que influyen en comportamiento (violencia, hábitos de compra...)

Construcción del árbol (2)

Valor informativo de «talla»:

talla

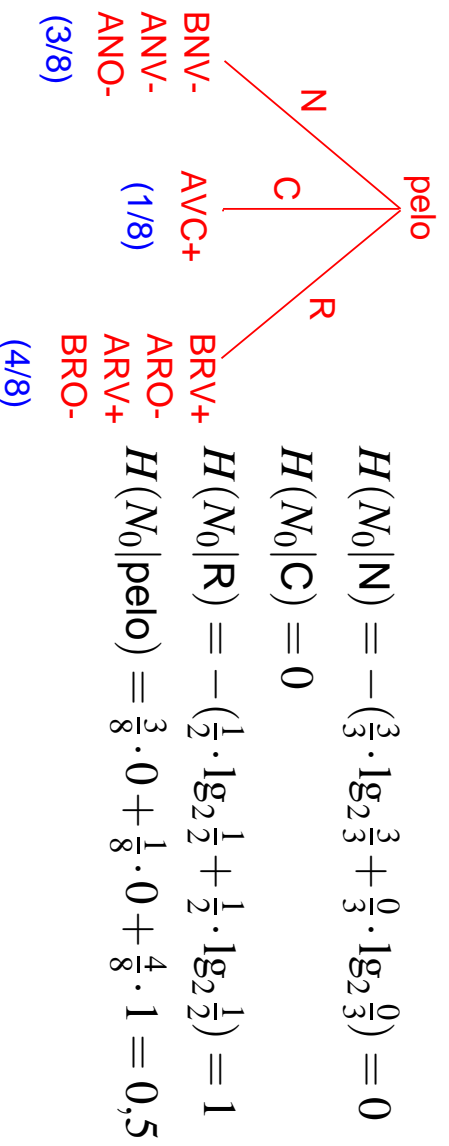


Ganancia de información por «talla»:

$$H(N_0) - H(N_0|\text{talla}) = 0,954 - 0,951 = 0,003 \text{ bits}$$

Construcción del árbol (3)

Valor informativo de «pelo»:



Ganancia de información por «pelo»:

$$H(N_0) - H(N_0|\text{pelo}) = 0,954 - 0,5 = 0,454 \text{ bits}$$

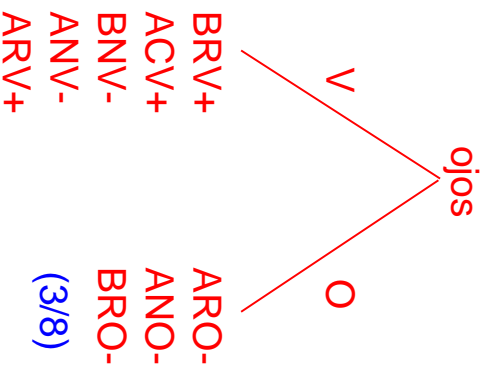
Construcción del árbol (4)

Valor informativo de «ojos»:

$$H(N_0|V) = -\left(\frac{2}{5} \cdot \lg_2 \frac{2}{5} + \frac{3}{5} \cdot \lg_2 \frac{3}{5}\right) = 0,971$$

$$H(N_0|O) = 0$$

$$H(N_0|\text{ojos}) = \frac{5}{8} \cdot 0,971 + \frac{3}{8} \cdot 0 = 0,607$$



(5/8)

(3/8)

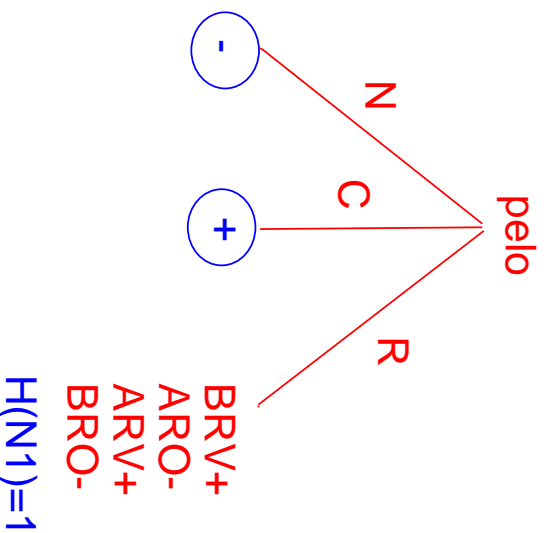
Ganancia de información por «ojos»:

$$H(N_0) - H(N_0|\text{ojos}) = 0,954 - 0,657 = 0,347 \text{ bits}$$

Construcción del árbol (5)

Mejor atributo hasta ahora: **pelo**

Árbol provisional:

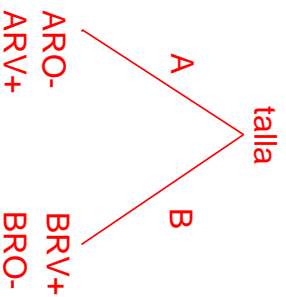


$$H(N|1)=1$$

Construcción del árbol (6)

Dado el valor de «pelo» quedan 4 ejemplos, 2 «+» y 2 «-»

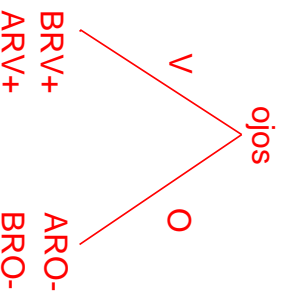
■ talla:



$$H(N|talla) = 1$$

Ganancia: 0

■ ojos:

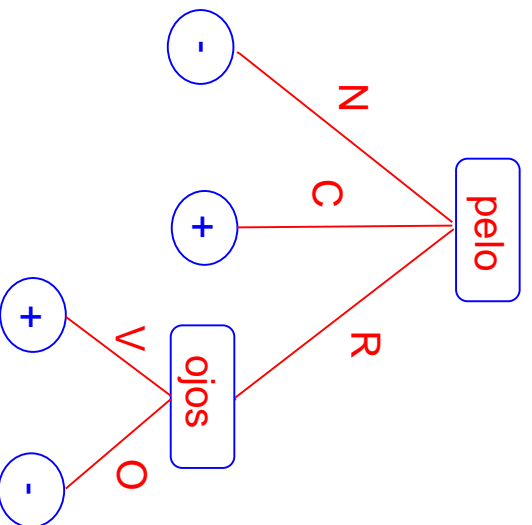


$$H(N|ojos) = 0$$

Ganancia: 1

Construcción del árbol (7)

Árbol resultante:



Reglas:

- Si (v(pelo) = «negro») entonces «-»
- Si (v(pelo) = «castaño») entonces «+»
- Si (v(pelo) = «rubio») y (v(ojos) = «verdes») entonces «+»
- Si (v(pelo) = «rubio») y (v(ojos) = «oscuros») entonces «-»

ID3: generalización para K clases

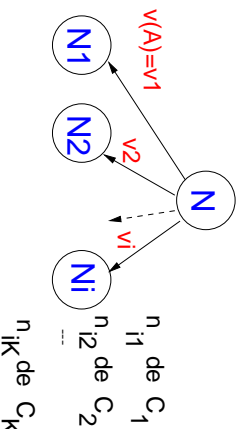
$$p_k = \frac{n_k}{n_1 + n_2 + \dots + n_K}; \quad H(N) = - \sum_{k=1}^K (p_k \lg_2 p_k); \quad 0 \leq H(N) \leq \lg_2 K$$

Partición por el atributo A con valores v1, v2, ...:

$$\begin{matrix} n_1 \text{ de } C_1 & p_1 = n_1 / (n_1 + n_2 + \dots + n_K) \\ n_2 \text{ de } C_2 & p_2 = n_2 / (n_1 + n_2 + \dots + n_K) \\ \dots & \dots \\ n_K \text{ de } C_K & p_K = n_K / (n_1 + n_2 + \dots + n_K) \end{matrix}$$

$$p_{ik} = \frac{n_{ik}}{n_{i1} + n_{i2} + \dots + n_{iK}}$$

$$H_i = - \sum_k (p_{ik} \lg_2 p_{ik})$$



Valor medio de la entropía después de saber el valor de A:

$$H(N|A) = \sum_{i=1}^n Pr(v(A) = v_i) \cdot H_i = \sum_{i=1}^n \left(\frac{\sum_k n_{ik}}{\sum_i \sum_k n_{ik}} \right) \cdot H_i$$

Reducción media de la entropía = información ganada por A =

$$H(N) - H(N|A)$$

¿Es la ganancia la mejor medida?

En el ejemplo de Quinlan supongamos un cuarto atributo: un número de identificación

¿Qué árbol se induce?

$$H(N_0|1) = H(N_0|2) = \dots = H(N_0|8) = 0$$

Es decir, ¡ $H(N_0|ID) = 0$!

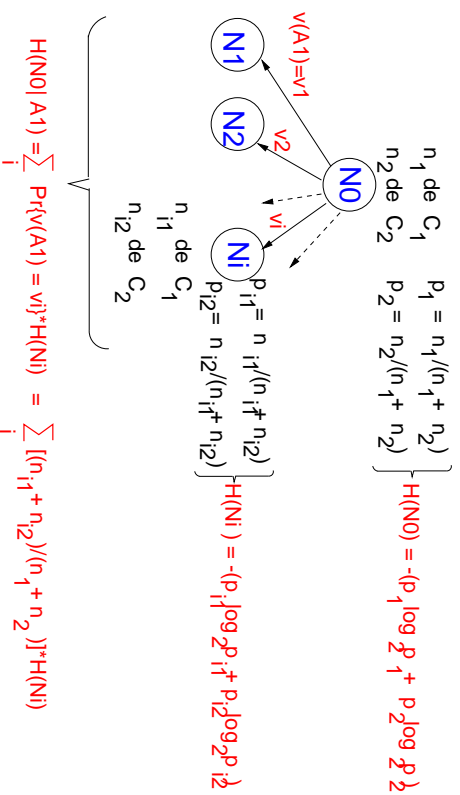
ID	talla	pelo	ojos	clase
1	B	R	V	+
2	A	R	O	-
3	A	C	V	+
4	B	N	V	-
5	A	N	V	-
6	A	R	V	+
7	A	N	O	-
8	B	R	O	-

Causa: con la ganancia se da preferencia a atributos con muchos valores posibles que conducen a nodos con pocos ejemplos.

Una solución: ponderar la ganancia del atributo teniendo en cuenta el número de nodos resultantes y el número de ejemplos en cada uno, independientemente de la clasificación de estos ejemplos en cada nodo

Entropía de un atributo

Medida de la cantidad de información necesaria para determinar a qué partición (rama del árbol) se asigna un ejemplo: a más ramas y a menos ejemplos por rama, más entropía.



Entropía de A_1 en N_0 : $H(A_1) = -\sum_i (p_i \times \lg_2 p_i)$

donde: $p_i = n_i/n$, con $n_i = n_{i1} + n_{i2}$ y $n = \sum_i n_i$

Ganancia ponderada, o tasa de ganancia

$$G_P(A) = \frac{H(N) - H(N|A)}{H(A)}$$

En el ejemplo, en N_0

$H(\text{ID}) = -(\frac{1}{8} \cdot \lg_2 \frac{1}{8}) \times 8 = 3$ (8 ramas, 1 ejemplo por rama)

$H(\text{talla}) = H(\text{ojos}) = -(\frac{5}{8} \cdot \lg_2 \frac{5}{8} + \frac{3}{8} \cdot \lg_2 \frac{3}{8}) = 0,954$

(dos ramas, 5 + 3 ejemplos)

$H(\text{pelo}) = -(\frac{3}{8} \cdot \lg_2 \frac{3}{8} + \frac{1}{8} \cdot \lg_2 \frac{1}{8} + \frac{4}{8} \cdot \lg_2 \frac{4}{8}) = 1,406$

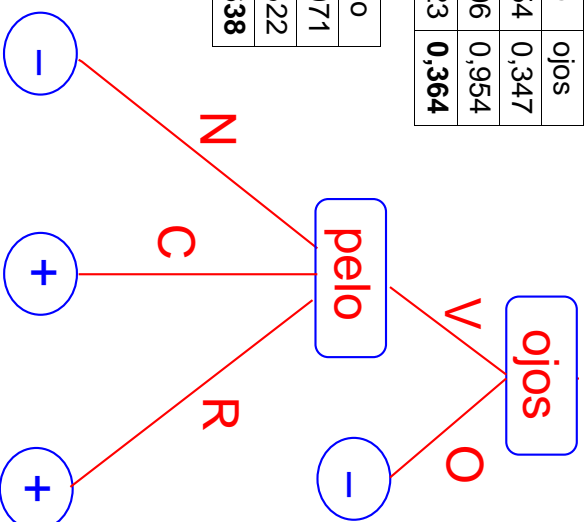
	ID	talla	pelo	ojos
Ganancia, G	0,954	0,003	0,454	0,347
Entropía, H	3,0	0,954	1,406	0,954
$G_P = G/H$	0,318	0,003	0,323	0,364

Y el mejor atributo es...

Árbol resultante en el ejemplo con la ganancia ponderada

	ID	talla	pelo	ojos
Ganancia, G	0,954	0,003	0,454	0,347
Entropía, H	3	0,954	1,406	0,954
$G_p = G/H$	0,318	0,003	0,323	0,364

	ID	talla	pelo
Ganancia, G	0,971	0,020	0,971
Entropía, H	2,322	0,951	1,522
$G_p = G/H$	0,418	0,0021	0,638



Mejoras sobre ID3

- Datos con «ruido» evitando sobreajuste
- Atributos con valores continuos
- Ejemplos incompletos
- Costes de los atributos

Muchas implementadas en C4.5 (Quinlan, 1993) y derivados

Sobreajuste (*overfitting*) (1)

Una hipótesis $h \in \mathcal{H}$ *sobreajusta* los datos del conjunto de entrenamiento (ejemplos) si existe una hipótesis alternativa $h' \in \mathcal{H}$ tal que:

- h tiene menor error que h' sobre los ejemplos, pero
- h' tiene menor error que h sobre el universo (\mathcal{E})

Causa: Unos pocos ejemplos con ruido hacen crecer artificialmente al árbol

Sobreajuste (*overfitting*) (2)

Heurístico: poda

- *temprana* (durante la construcción del árbol)
- *tardía* (después de construido)

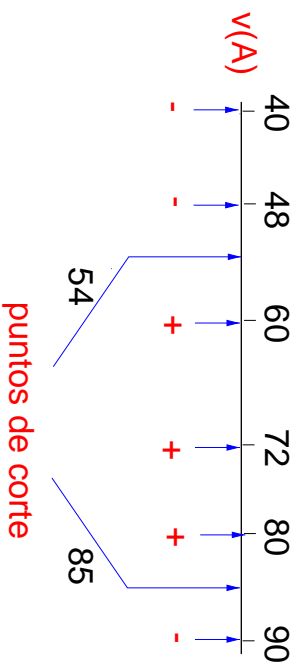
Criterios para la poda:

- un segundo conjunto de entrenamiento (conjunto de validación)
- pruebas estadísticas para estimar si la expansión (o la poda) de un nodo puede mejorar la precisión sobre \mathcal{E}
- medidas de la complejidad de la enumeración de los ejemplos y de la descripción del árbol (*Minimum Description Length*)

Atributos con valores continuos

Discretización

- Dependiente del conocimiento de base del dominio
Por ejemplo: Edad \rightsquigarrow {niño, joven, adulto, viejo}
Temperatura \rightsquigarrow {baja, normal, alta}
- Automática. Por ejemplo:



$$v(A) \rightsquigarrow \{v(A) > 54, v(A) > 85\}$$

Ejemplos incompletos

Estimación del valor de un atributo en un ejemplo en el que falta:

- valor = «desconocido»
- el más frecuente entre los ejemplos que sí lo tienen (atributos nominales)
- un valor medio, ponderado o no (atributos numéricos)

Costes de medida de atributos

1. *Definición de una función de coste, $C(A)$*

2. *Ponderación de la función Ganancia*
(muy dependiente del dominio):

- $\frac{G^2(A)}{C(A)}$

(Tan, 1993: robótica)

- $\frac{2^{G(A)} - 1}{(C(A) + 1)^\alpha}$

(Núñez, 1991: diagnóstico médico)

Software para inducción de árboles

- C4.5: <http://www.rulequest.com/Personal/Integrado> (con distintas variantes) en muchas herramientas de minería de datos
- ITI (Incremental Tree Inducer)
<http://www-lrn.cs.umass.edu/iti/index.html>
- Enlaces a programas comerciales y libres en:
<http://www.kdnuggets.com/software/classification-decision-tree.html>