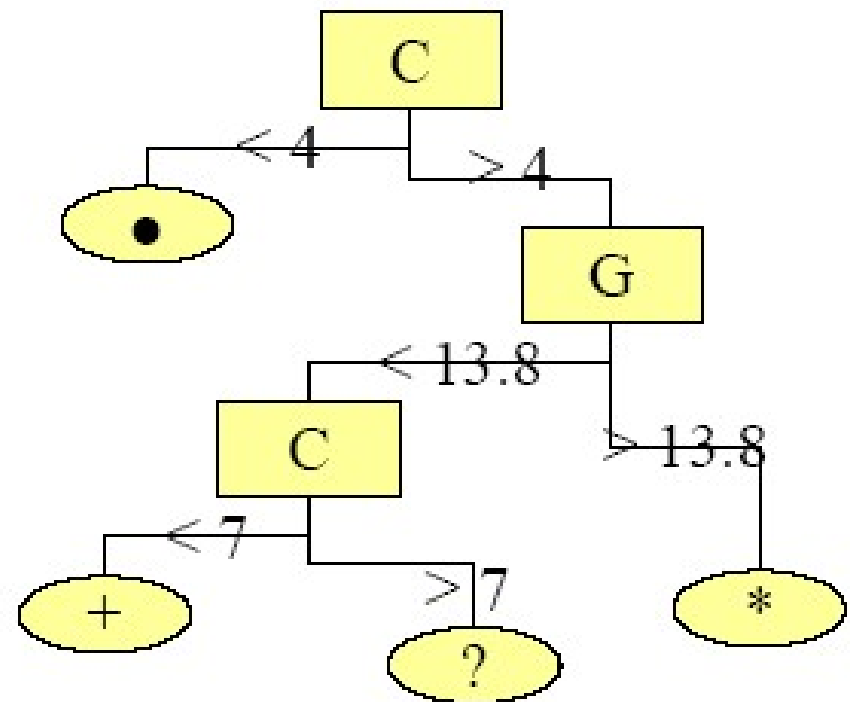
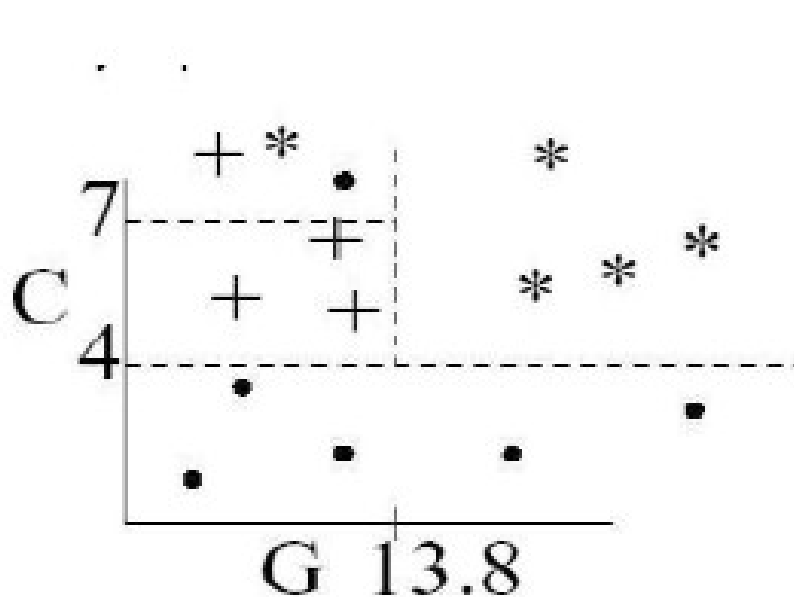


Arboles de Decisión (II)

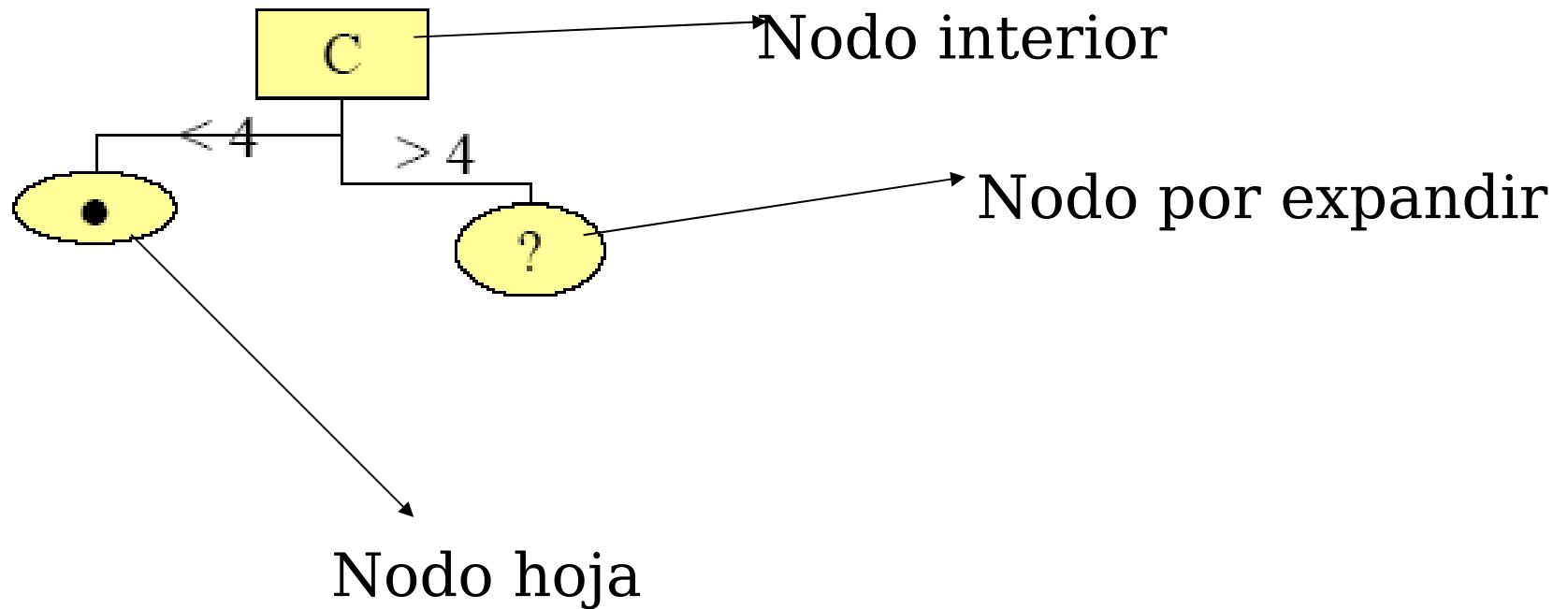
Carlos Hurtado L.

Depto de Ciencias de la
Computación, Universidad de
Chile

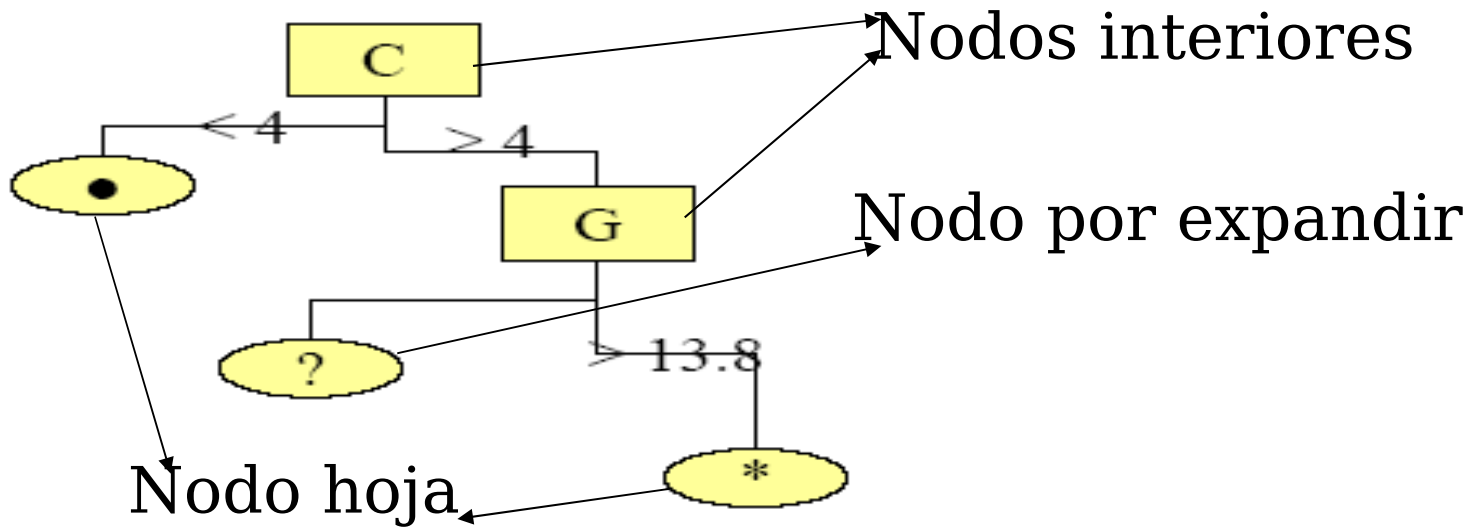
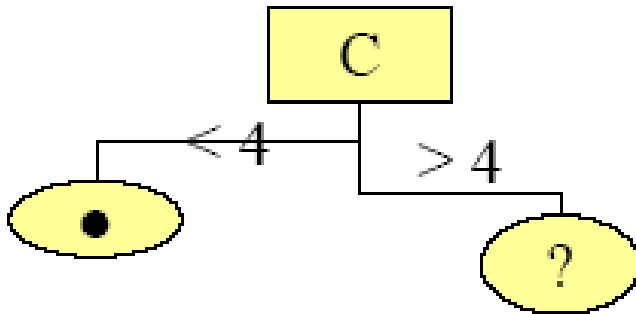
Arboles de Decisión



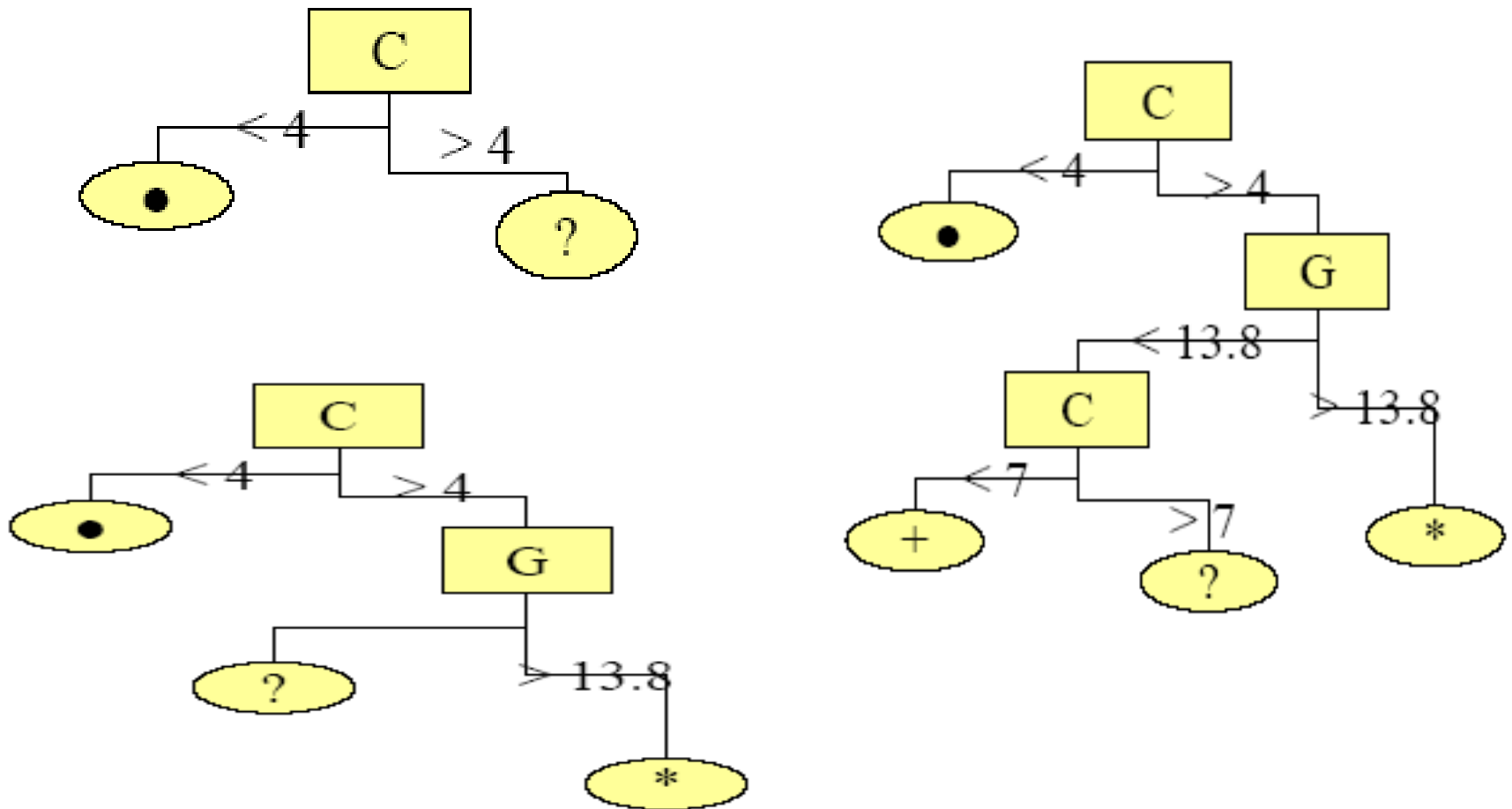
Algoritmo de Hunt (I)



Algoritmo de Hunt (II)



Algoritmo de Hunt (III)



Algoritmo de Hunt

- Main(Conjunto de Datos T)
 - Expandir(T)
- Expandir(Conjunto de Datos S)
 - If (todos los datos están en la misma clase) then return
 - Encontrar el mejor split r
 - Usar r para particionar S en $S1$ y $S2$
 - Expandir($S1$)
 - Expandir($S2$)

¿Cuál es el mejor split?

- Buscamos splits que generen nodos hijos con la menor impureza posible (mayor pureza posible)
- Existen distintos métodos para eval splits.
- **Criterios de Selección:**
 - Índice Gini
 - Entropía (Ganancia de información)
 - Test Chi-cuadrado
 - Proporción de Ganancia de Información

Selección de splits usando índice Gini

- Recordar que cada nodo del árbol define un subconjunto de los datos de entrenamientos
- Dado un nodo t del árbol, $Gini(t)$ mide el grado de “pureza” de t con respecto a las clases
 - Mayor $Gini(t)$ implica menor pureza
 - $Gini(t) = 1 - \text{Probabilidad de sacar dos registros de la misma clase}$

Indice Gini

$Gini(t)$: probabilidad de NO sacar dos t registros de la misma clase del nodo

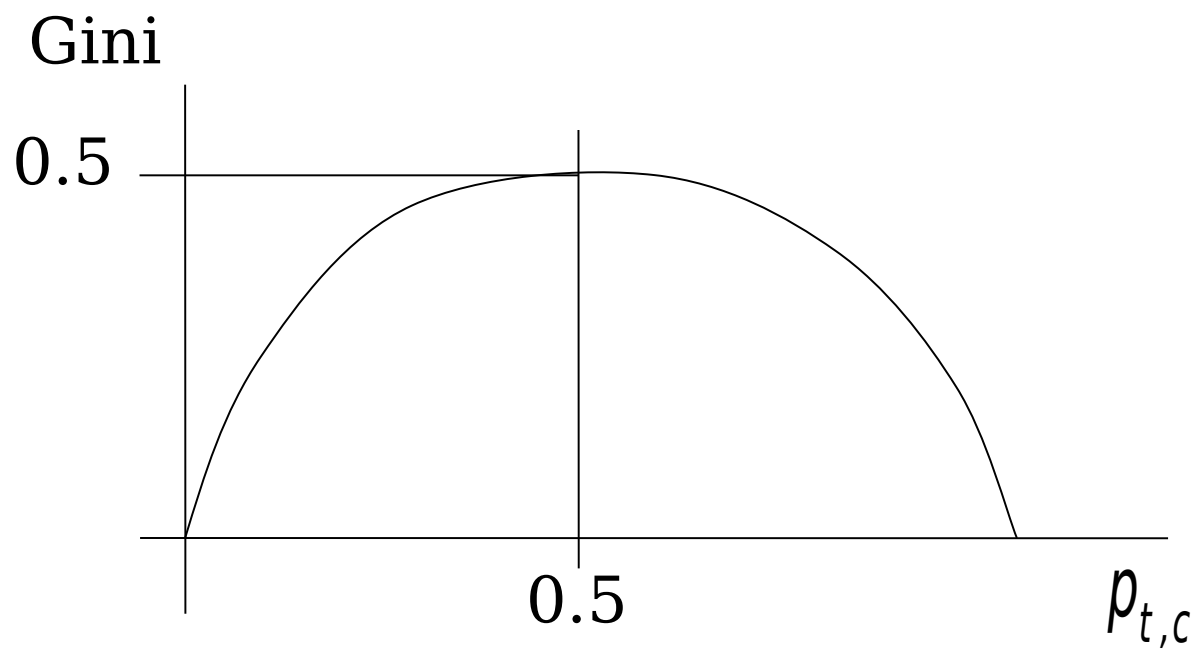
$$Gini(t) = 1 - \sum_{c \in C} p_{t,c}^2$$

donde C es el conjunto de clases y $p_{t,c}$ es la prob. de ocurrencia de la clase c en el nodo t

Indice Gini: ejemplo

C_1	C_2	Gini
0	6	0
1	5	0.278
2	4	0.444
3	3	0.5

Indice Gini



Selección de Splits: GiniSplit

- Criterio de selección: seleccionar el split con menor gini ponderado (GiniSplit)
- Dado un split $S = \{s_1, \dots, s_n\}$ de t

$$GiniSplit(t, S) = \sum_{s \in S} \frac{|s|}{|t|} Gini(s)$$

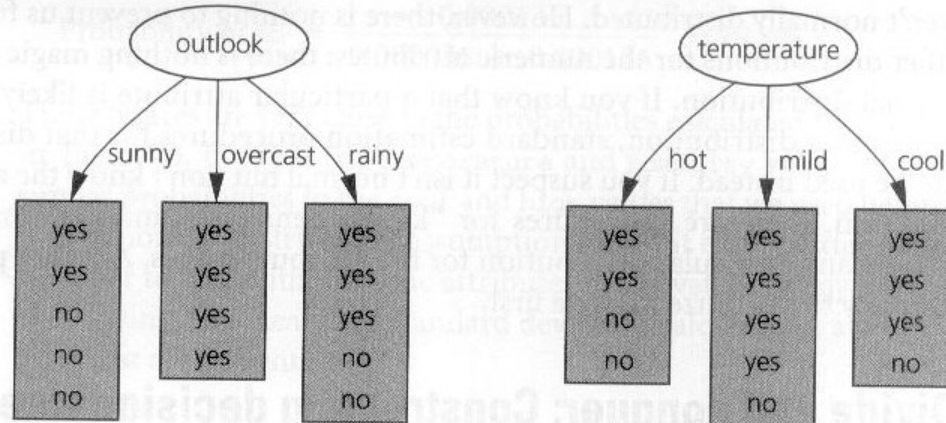
Ejemplo: weather.nominal

Outlook	Temp.	Humidity	Windy	Play
Sunny	Hot	High	FALSE	No
Sunny	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Rainy	Mild	High	FALSE	Yes
Rainy	Cool	Normal	FALSE	Yes
Rainy	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Sunny	Mild	High	FALSE	No
Sunny	Cool	Normal	FALSE	Yes
Rainy	Mild	Normal	FALSE	Yes
Sunny	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Rainy	Mild	High	TRUE	No

Weather.nominal: splits

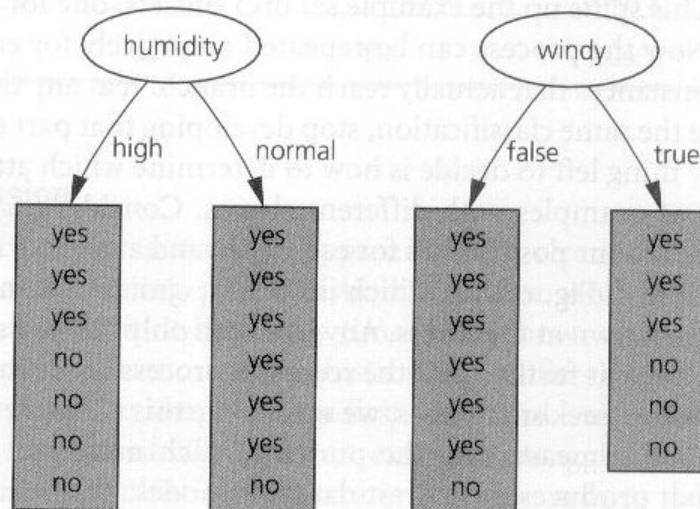
- En este caso todos los atributos son nominales
- En este ejemplo usaremos *splits* simples

Possible splits



(a)

(b)



(c)

(d)

weather.nominal: atributos vs. clases

<i>Outlook</i>			<i>Temperature</i>			<i>Humidity</i>			<i>Windy</i>			<i>Play</i>	
<i>Yes</i>		<i>No</i>	<i>Yes</i>		<i>No</i>	<i>Yes</i>		<i>No</i>	<i>Yes</i>		<i>No</i>	<i>Yes</i>	<i>No</i>
Sunny	2	3	Hot	2	2	High	3	4	FALSE	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	TRUE	3	3		
Rainy	3	2	Cool	3	1								

Figure 18.1 displays four decision trees (a, b, c, d) illustrating the greedy search algorithm for feature selection. Each tree shows a root node (outlook, temperature, humidity, or windy) branching into child nodes based on feature values, leading to leaf nodes with 'yes' or 'no' labels.

(a) **outlook**

- sunny: yes, yes, no, no, no
- overcast: yes, yes, yes, yes
- rainy: yes, yes, yes, no, no

(b) **temperature**

- hot: yes, yes, no, no
- mild: yes, yes, yes, yes, no, no
- cool: yes, yes, yes, no

(c) **humidity**

- high: yes, yes, yes, no, no, no, no
- normal: yes, yes, yes, yes, yes, no

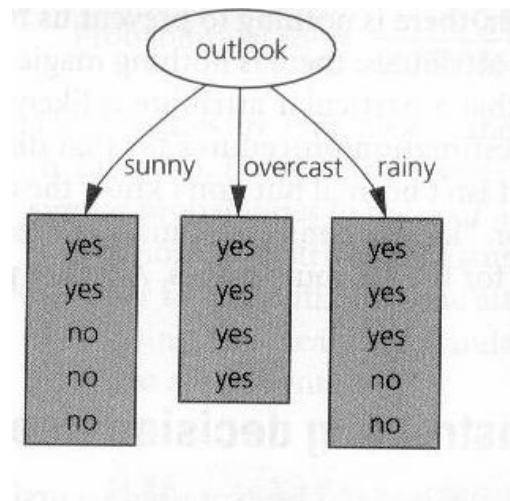
(d) **windy**

- false: yes, yes, yes, yes, yes, yes, no, no
- true: yes, yes, yes, no, no, no



Ejemplo

- Split simple sobre variable Outlook
 - $\text{Gini}(\text{sunny}) = 1 - 0.16 - 0.36 = 0.48$
 - $\text{Gini}(\text{overcast}) = 0$
 - $\text{Gini}(\text{rainy}) = 0.48$
 - $\text{GiniSplit} = (5/14) 0.48 + 0 0.48 + (5/14) 0.48 = 0.35$



Ejercicio: selección de splits usando Gini para weather.nominal

1. Dar el número de splits que necesitamos evaluar en la primera iteración del algoritmo de Hunt, para (a) splits complejos y (b) splits simples.
2. Seleccionar el mejor split usando el criterio del índice Gini
 - Calcular el GiniSplit de cada split
 - Determinar el mejor split

¿Cuál es el mejor split?

- Buscamos splits que generen nodos hijos con la menor impureza posible (mayor pureza posible)
- Existen distintos métodos para eval splits.
- **Criterios de Selección:**
 - Índice Gini
 - **Entropía (Ganancia de información)**
 - Test Chi-cuadrado
 - Proporción de Ganancia de Información

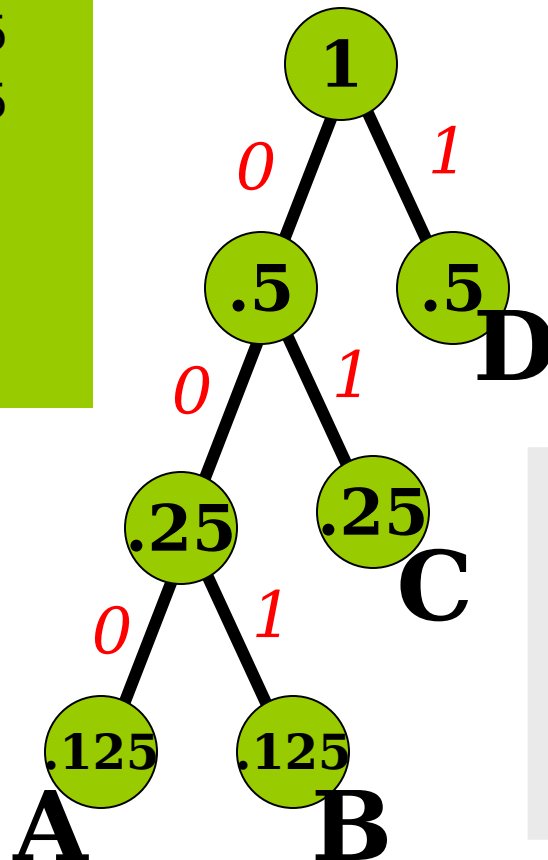
Teoría de Información

- Supongamos que transmitimos en un canal (o codificamos) símbolos en $\{x_1, x_2, \dots, x_n\}$.
- A primera vista, necesitamos $\log n$ bits por mensaje.
- Si el emisor y el receptor conocen la distribución de prb. $P(X=x_i)=p_i$, podríamos elaborar un código que requiera menos bits por mensaje.

Códigos de Huffman

Símb. Prob.

A .125
B .125
C .25
D .5



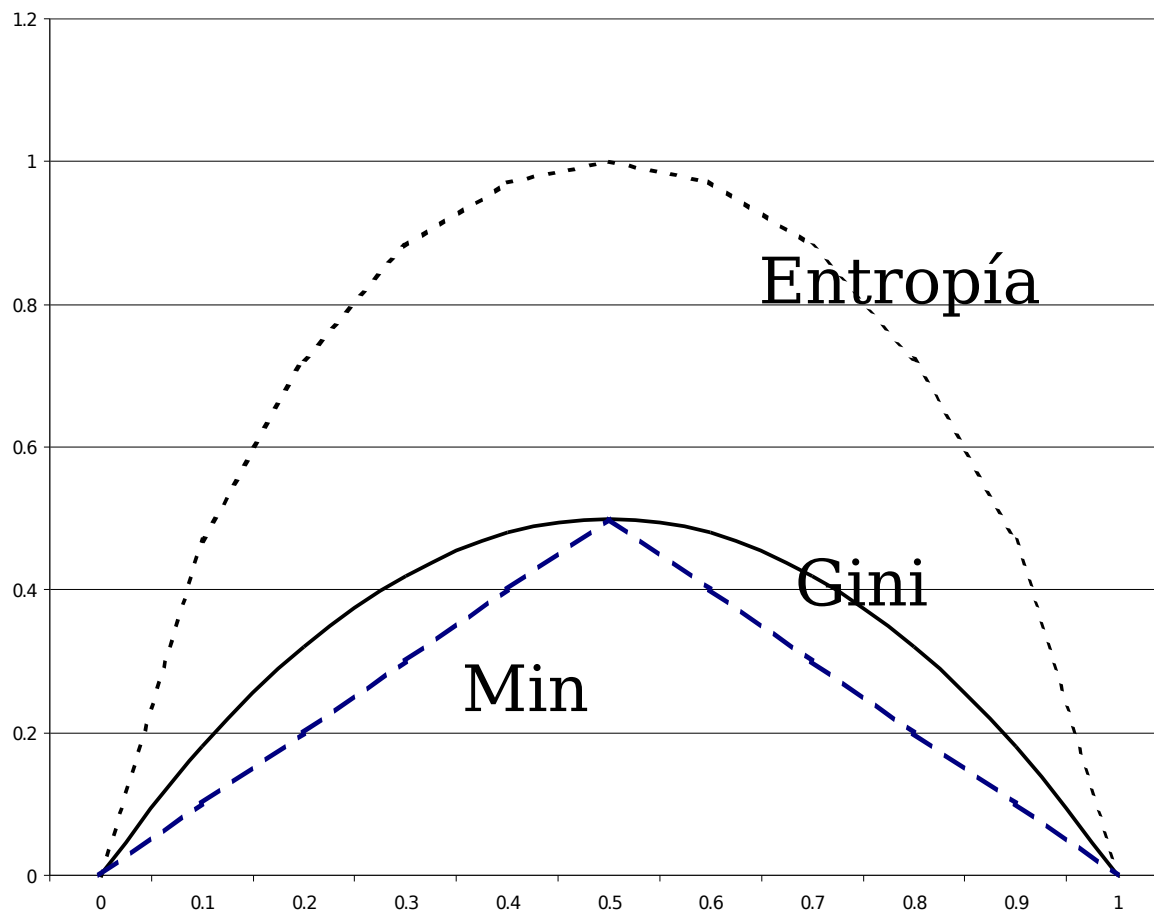
M	cod.	long.	prob	
A	000	3	0,125	0,375
B	001	3	0,125	0,375
C	01	2	0,250	0,500
D	1	1	0,500	0,500
average message length				1,750

Si usamos este código para enviar mensajes (A,B,C, o D) con la distribución de prob. dada, en promedio cada mensaje require 1.75 bits.

Entropía

- Entropía: mínimo teórico de bits promedio necesarios para transmitir un conjunto de mensajes sobre $\{x_1, x_2, \dots, x_n\}$ con distribución de prob. $P(X=x_i)=p_i$
- $\text{Entropy}(P) = -(p_1 \cdot \log(p_1) + p_2 \cdot \log(p_2) + \dots + p_n \cdot \log(p_n))$
- Información asociada a la distribución de probabilidades P .
- Ejemplos:
 - Si P es $(0.5, 0.5)$, $\text{Entropy}(P) = 1$
 - Si P es $(0.67, 0.33)$, $\text{Entropy}(P) = 0.92$
 - Si P is $(1, 0)$, $\text{Entropy}(P)=0$
- Mientras más uniforme es P , mayor es su entropía

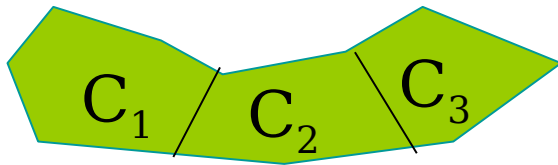
Entropía



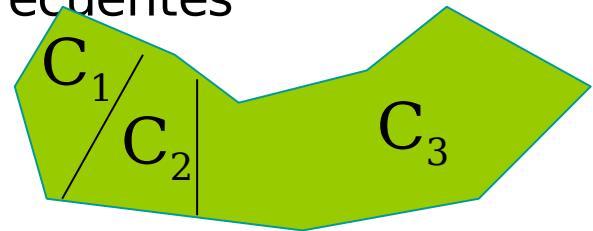
Selección de splits usando Entropía

$$Entropy(t) = \sum_{c \in C} -p_{t,c} \log_2 p_{t,c}$$

- La entropía mide la impureza de los datos S
- Mide la información (num de bits) promedio necesaria para codificar las clases de los datos en el nodo t
- Casos extremos
 - Todos los datos pertenecen a la misma clase (Nota: $0 \cdot \log 0 = 0$)
 - Todas las clases son igualmente frecuentes



Entropía alta



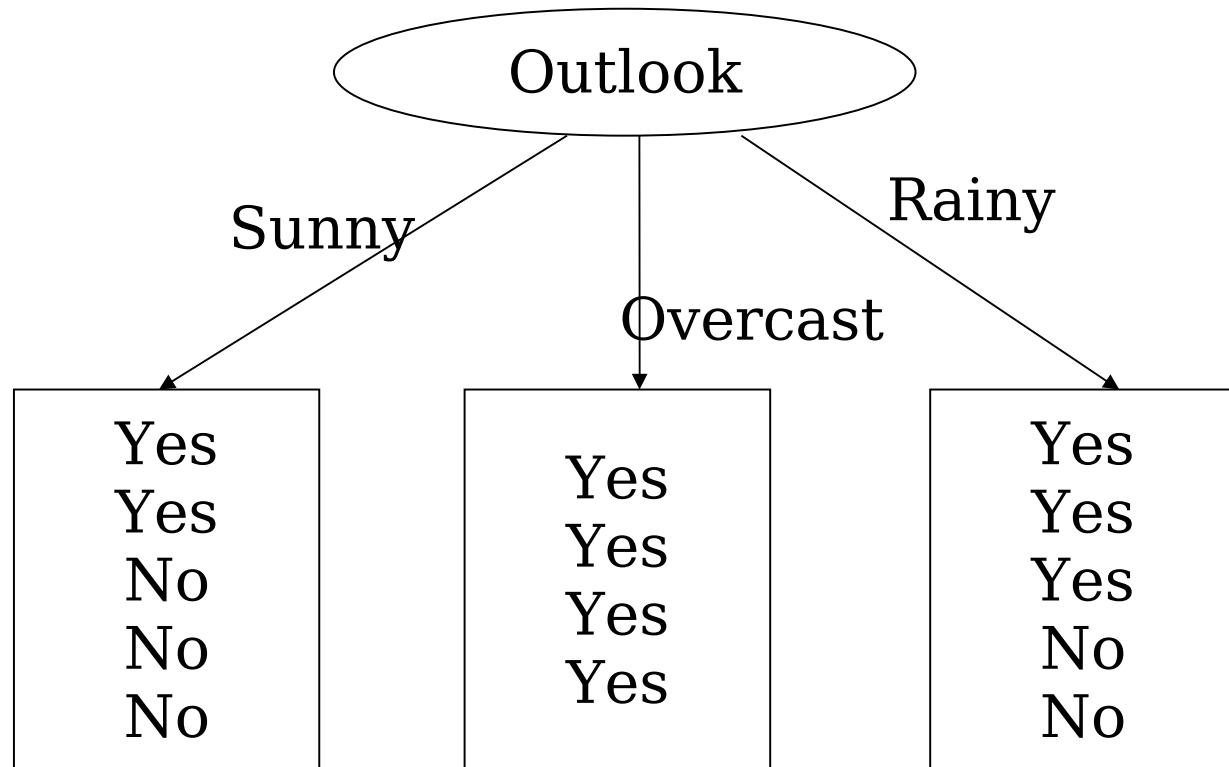
Entropía baja

Selección de Splits: Ganancia de Información

- Criterio para elegir un split: seleccionar el split con la mayor ganancia de información (Gain)
- Equivalente a seleccionar el split con la menor entropía ponderada
- Dado un split $S = \{s_1, \dots, s_n\}$ de t

$$Gain(t, S) = Entropy(t) - \sum_{s \in S} \frac{|s|}{|t|} Entropy(s)$$

Ejemplo: consideremos el split de Outlook



Ejemplo: Split simple de Outlook

$$Entropy(S) = \sum_{i=1}^2 -p_i \log_2 p_i$$

$$i - \frac{9}{14} \cdot \log_2 \frac{9}{14} - \frac{5}{14} \cdot \log_2 \frac{5}{14} = 0.92$$

$$Entropy(S_{sunny}) = -\frac{2}{5} \cdot \log_2 \frac{2}{5} - \frac{3}{5} \cdot \log_2 \frac{3}{5} = 0.97 \quad i$$

$$Entropy(S_{overcast}) = -\frac{4}{4} \cdot \log_2 \frac{4}{4} - \frac{0}{4} \cdot \log_2 \frac{0}{4} = 0.00$$

$$Entropy(S_{rainy}) = -\frac{3}{5} \cdot \log_2 \frac{3}{5} - \frac{2}{5} \cdot \log_2 \frac{2}{5} = 0.97$$

Ejemplo: Cálculo de la Ganancia de información

$$Gain(S, outlook) = Entropy(S) - \sum_{v \in Values(a)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$i = 0.92 - \frac{5}{14} \cdot 0.97 - \frac{4}{14} \cdot 0 - \frac{5}{14} \cdot 0.97$$

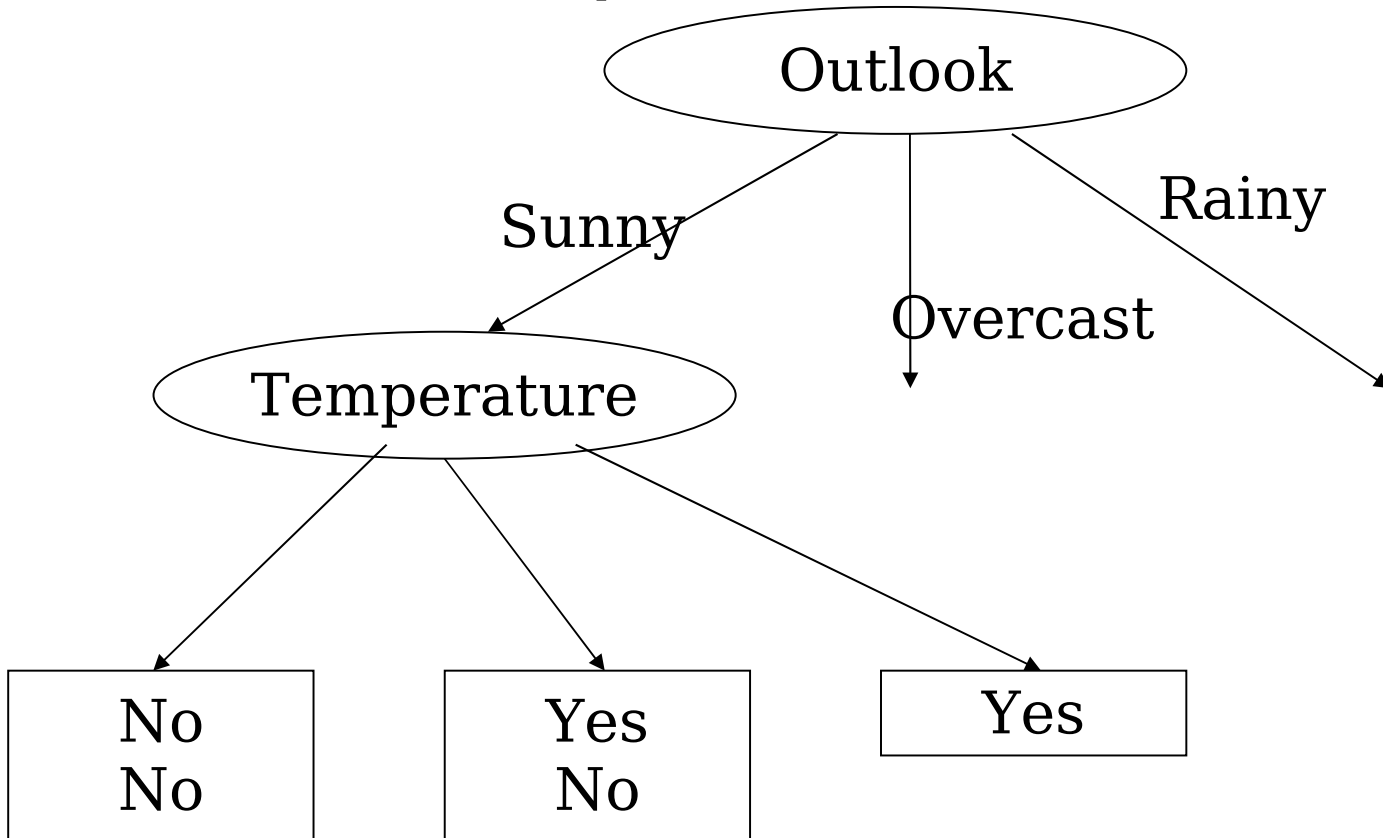
$$i = 0.92 - 0.69 = 0.23$$

$$Gain(S, temp) = 0.03$$

$$Gain(S, humidity) = 0.15$$

$$Gain(S, windy) = 0.05$$

Ejercicio: Expansión de nodo “Sunny” (probamos el split “Temperature”)



Cálculo de la Entropía para los nodos del split “Temperature”

$$Entropy(S) = 0.97$$

$$Entropy(S_{hot}) = -\frac{0}{2} \cdot \log_2 \frac{0}{2} - \frac{2}{2} \cdot \log_2 \frac{2}{2} = 0$$

$$Entropy(S_{mild}) = -\frac{1}{2} \cdot \log_2 \frac{1}{2} - \frac{1}{2} \cdot \log_2 \frac{1}{2} = 1$$

$$Entropy(S_{cool}) = -\frac{1}{1} \cdot \log_2 \frac{1}{1} - \frac{0}{1} \cdot \log_2 \frac{0}{1} = 0$$

Cálculo de la Ganancia para “Temperature”

$$Gain(S, temp) = Entropy(S) - \sum_{v \in Values(a)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$\text{¿ } 0.97 - \frac{2}{5} \cdot 0 - \frac{2}{5} \cdot 1 - \frac{1}{5} \cdot 0$$

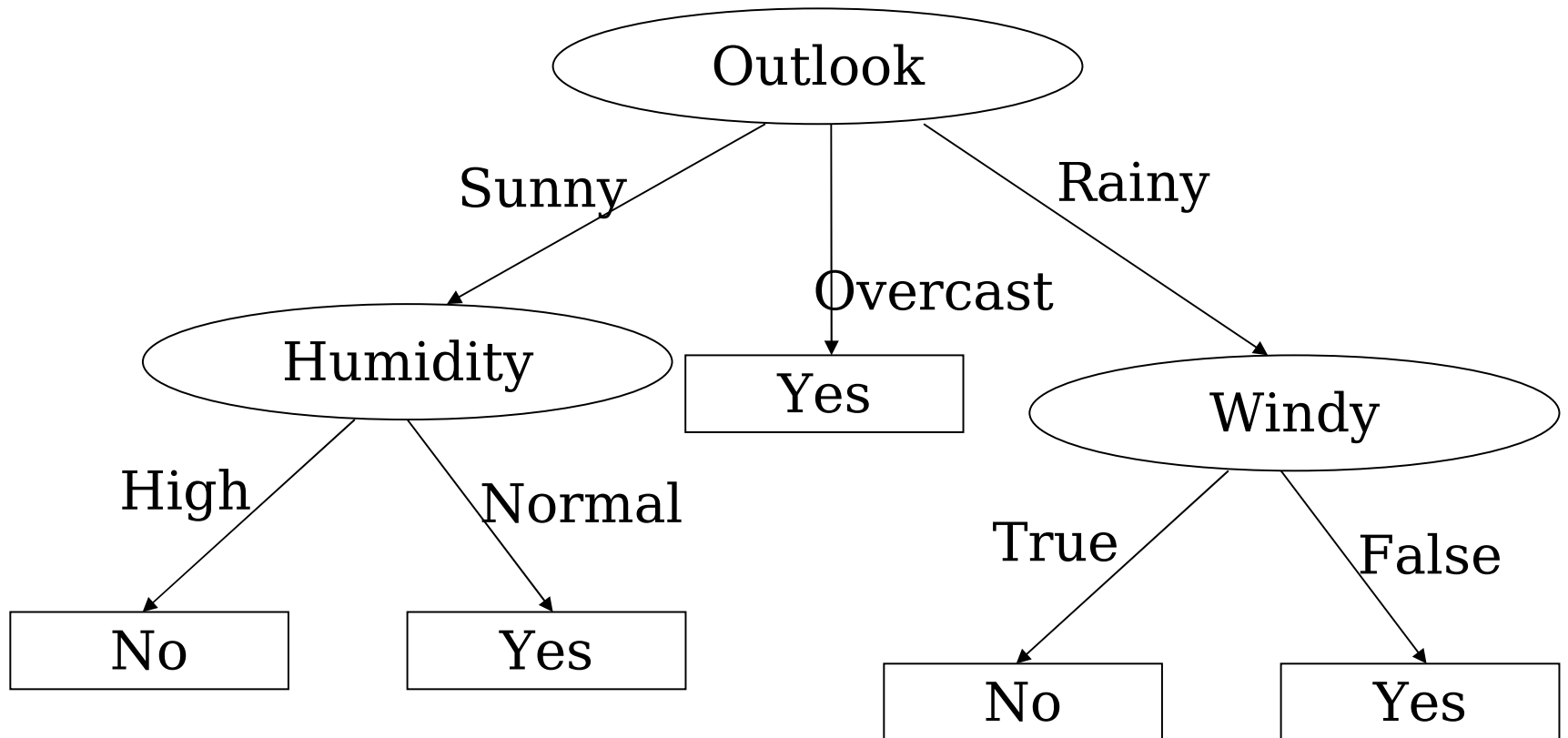
$$\text{¿ } 0.97 - 0.40 = 0.57$$

$$Gain(S, humidity) = 0.97$$

$$Gain(S, windy) = 0.02$$

La ganancia de “humidity” es mayor
Por lo que seleccionamos este split

Ejemplo: Arbol resultante

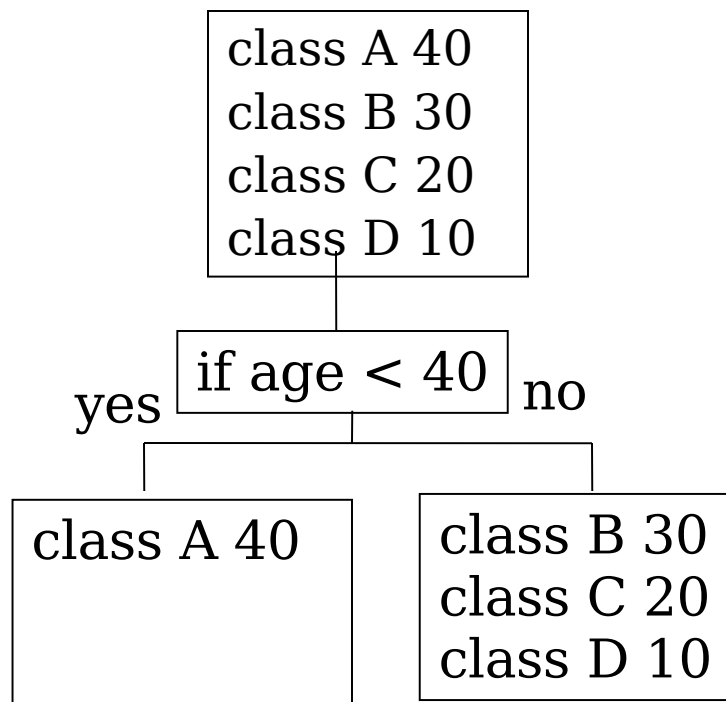


Entropía vs. Índice Gini

- Breiman, L., (1996). Technical note: Some properties of splitting criteria, *Machine Learning* 24, 41-47. Conclusiones de estudio empírico:
- Gini:
 - Tiende a seleccionar splits que aíslan una clase mayoritaria en un nodo y el resto en otros nodos.
 - Tiende a crear splits desbalanceados.
- Entropía:
 - Favorece splits balanceados en número de datos.

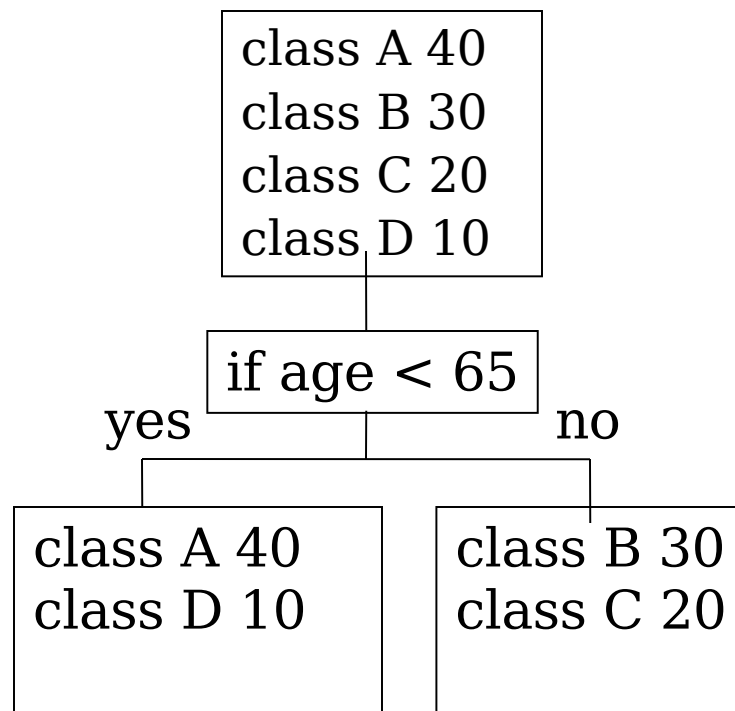
Entropía vs. Índice Gini

- Índice Gini tiende a aislar clases numerosas de otras clases



GiniSplit = 0,264 EntPond=0,259

- Entropía tiende a encontrar grupos de clases que suman más del 50% de los datos



GiniSplit=0,4 EntPond=0,254

Selección de split usando test Chi-cuadrado

- El test de chi-cuadrado se usa para testear la hipótesis nula de que dos variables son independientes.
- Es decir la ocurrencia de un valor de una variable no induce la ocurrencia del valor de otra variable.
- Usando el test de Chi-cuadrado mediremos la dependencia entre la variable del split y la variable de la clase.

Noción de dependencia de dos variable

- Consideramos que los datos son instancias de un vector aleatorio

$$(X_1, \dots, X_n, C)$$

- Es decir, cada instancia $(x_1, \dots, x_n, c)$ de este vector tiene asociada una probabilidad:

$$\Pr(X_1 = x_1, \dots, X_n = x_n, C = c)$$

Ejercicio: probabilidades para weather.nominal

- El vector de variables es (Outlook,Temp,Humidity,Windy,Play)
- Asumiendo que los datos corresponden a las probabilidades reales (esto no es siempre así ¿por qué?)
- La variable de la clase es Play.
- Estime de los datos las siguientes probabilidades:
 - $P(\text{Play}=\text{yes})$ $P(\text{Play}=\text{no})$
 - $P(\text{Play} = \text{yes}, \text{Outlook} = \text{sunny})$, $P(\text{play}=\text{no}, \text{Outlook}=\text{overcast})$
 - $P(\text{Outlook},\text{Temp},\text{Humidity})$ para todas las combinaciones.
 - $P(\text{Play}=\text{yes} \mid \text{Windy}=\text{false})$

weather.nominal

Outlook	Temp.	Humidity	Windy	Play
Sunny	Hot	High	FALSE	No
Sunny	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Rainy	Mild	High	FALSE	Yes
Rainy	Cool	Normal	FALSE	Yes
Rainy	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Sunny	Mild	High	FALSE	No
Sunny	Cool	Normal	FALSE	Yes
Rainy	Mild	Normal	FALSE	Yes
Sunny	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes
Rainy	Mild	High	TRUE	No

Dependencia

- Para dos variables Y, Z la probabilidad condicional se define como:

$$\Pr(Y = y|Z = z) = \frac{\Pr(Y = y, Z = z)}{\Pr(Z = z)}$$

- Y, Z son independientes si para todo par de valores y, z se tiene

$$\Pr(Y = y|Z = z) = \Pr(Y = y) \Pr(Z = z)$$

Dependencia

weather.nominal

- ¿Son Outlook y Play independientes?
- Para responder esto podríamos
 - Estimar la probabilidad
 - Verificar si se cumple la condición de independencia
- Ejercicio: verificar esto.

Test de Chi-cuadrado

- En el caso general, los datos de entrenamiento sólo son una muestra de los datos en que usaremos el modelo (datos objetivo).
- El test de Chi-cuadrado es un método estadístico para decidir si hay dependencia en los datos objetivo en base a lo que observamos en los datos de entrenamiento.

Test de Chi-cuadrado

- Ejemplo: venta de té y café en un supermercado.
 - Tenemos dos variables Café (C) y Té (T)
 - Valores son compró o no-compró

Tabla de contingencia observada

	c	$\bar{c}$	Total
t	20	5	25
$\bar{t}$	70	5	75
Total	90	10	100

Tabla de Contingencia Observada vs. Esperada

- Cada celda $r=(r_1,\dots,r_n)$, contiene el número de veces $O(r)$ que ocurren juntos en los datos los valores $r_1,\dots,r_n$
- Para una celda $r=(r_1,\dots,r_n)$, definimos el valor esperado de la celda si las variables son dependientes $E(r)$:

$$E[r] = n \times \frac{E[r_1]}{n} \times \dots \times \frac{E[r_k]}{n}$$

donde $E(r_i)=O(r_i)$ es el número de ocurrencias de r_i en la tabla

Tabla de Contingencia Observada vs. Esperada

	c	$\bar{c}$	Total
t	20	5	25
$\bar{t}$	70	5	75
Total	90	10	100

$$E[t\bar{c}] = 100 \times \frac{25}{100} \times \frac{10}{100} = 2.5$$

$$E[tc] = 100 \times \frac{25}{100} \times \frac{90}{100} = 22.5$$

Estimador Chi-cuadrado

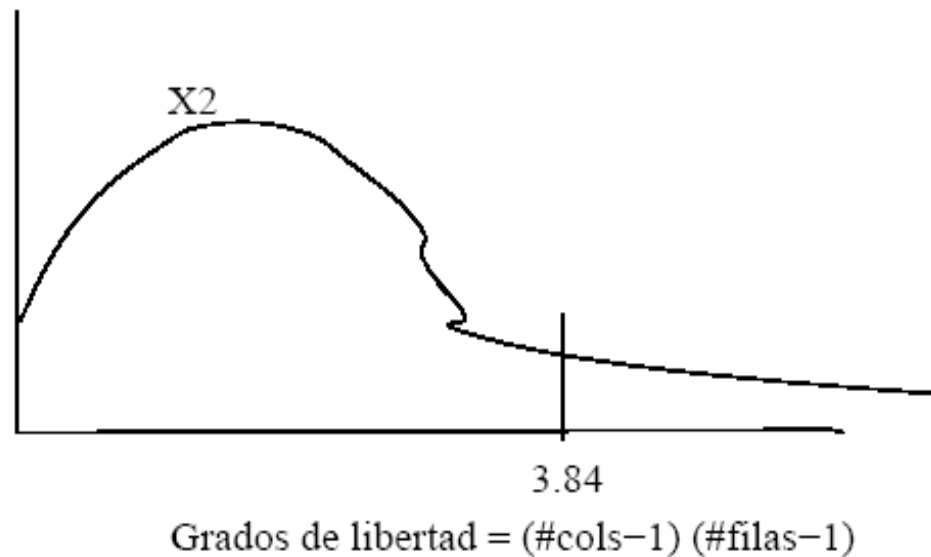
Estimador Chi-Cuadrado:

$$\chi^2 = \sum_{r \in R} \frac{(O(r) - E[r])^2}{E[r]}$$

Desviación normalizada entre la tabla de contingencia real
y la esperada

Estimador Chi-cuadrado (cont.)

Si χ^2 es mayor que 3.84 rechazamos la suposición de independencia con un 95% de nivel de confianza.

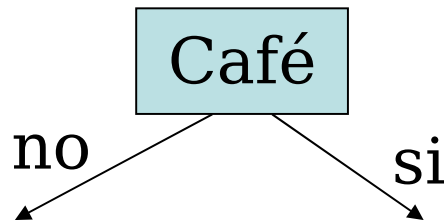


Uso de Chi-cuadrado para seleccionar splits

- Consideremos un split binario sobre una variable A .
- Las variables para la tabla de contingencia son C (variable de la clase) y A
 - Cada nodo hijo representa un valor distinto para A
- Se calcula tabla de contingencia y estimador chi-cuadrado
- Se elige el split con el mejor estimador chi-cuadrado

Ejemplo

- Supongamos que el atributo A es Café (C), y la variable de la clase es Té (T)
- El split quedaría



Ejemplo: consideremos el split simple de Outlook

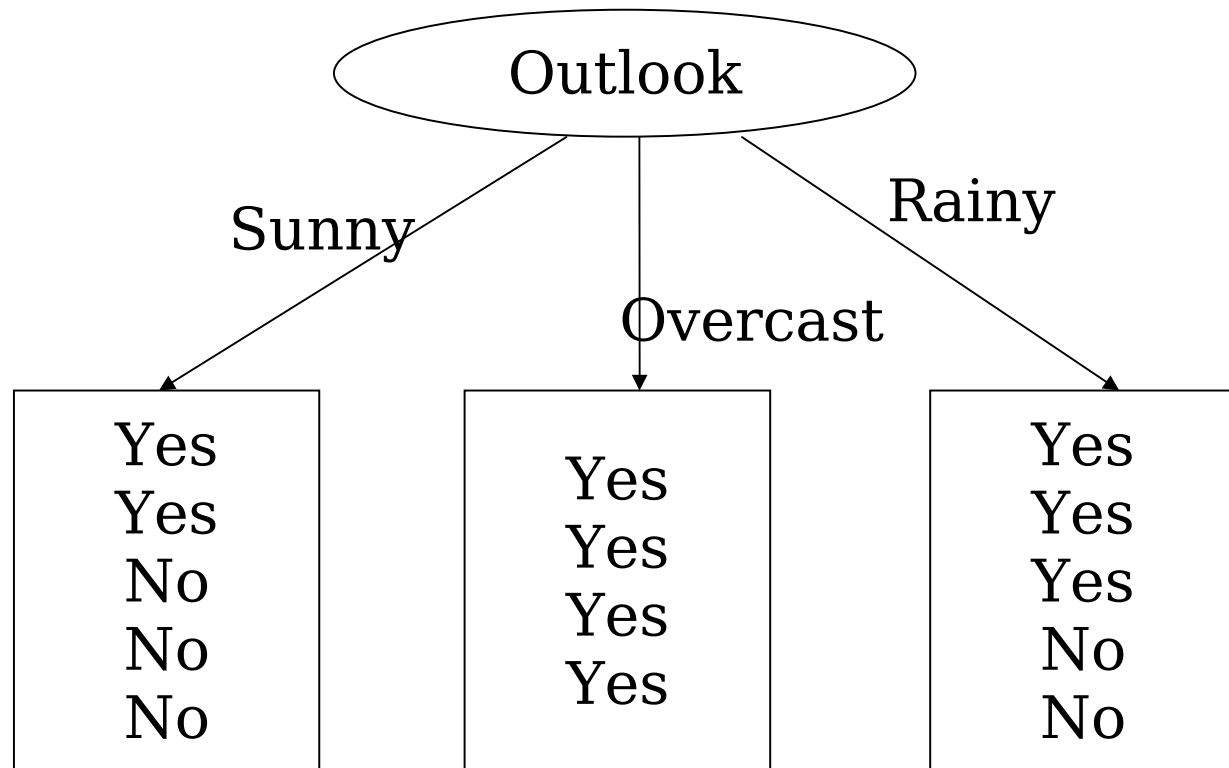


Tabla de Contingencia Observada vs. Esperada

Observada		Yes	No	Total
	Sunny	2	3	5
	Overcast	4	0	4
	Rainy	3	2	5
	Total	9	5	14

Esperada		Yes	No	Total
	Sunny	3.21	1.79	5
	Overcast	2.57	1.43	4
	Rainy	3.21	1.79	5
	Total	9	5	14

Tabla de Contingencia Esperada

	Yes	No	Total
Sunny	3.21	1.79	5
Overcast	2.57	1.43	4
Rainy	3.21	1.79	5
Total	9	5	14

Estimador Chi-Cuadrado

	Yes	No	Total
Sunny	0.46	0.83	
Overcast	0.79	1.43	
Rainy	0.01	0.03	
Total			3.55

Grados de libertad =2, valor crítico a 95% de conf. = 5.99

Luego no rechazamos suposición de independencia

Nota: el test de chi-cuadrado sólo se puede usar si no más del 10% de las celdas tienen menos de 5 de frecuencia observada..

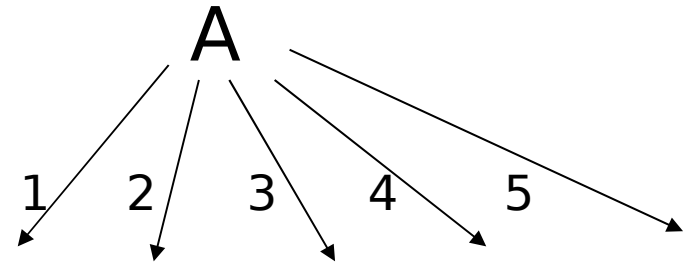
Ejercicio: selección de splits usando chi-cuadrado en weather.nominal

1. Calcule el estimador chi-cuadrado para los splits simples de Windy y Humidity
2. Diga en base al test chi-cuadrado si estos splits son independientes de la variable de la clase.

Problemas de entropía e Índice Gini

- Favorecen splits con muchos nodos hijos. Favorece atributos con muchos valores.

A	Clas
1	Si
2	No
3	No
4	Si
5	si



Problemas de Entropía, Gini y Chi-cuadrado

- Favorecen splits con muchos nodos hijos. Favorece atributos con muchos valores.
- Solución 1.: usar splits binarios
- Solución 2: usar GainRatio

$$GainRatio(t, S) = Gain(t, S) / SplitInfo(t, S)$$

$$SplitInfo(t, S) = Entropy\left(\frac{|S_1|}{|t|}, \dots, \frac{|S_n|}{|t|}\right)$$

Efecto de la elección del criterio de selección de splits

- A pesar de lo reportado por Breiman, un estudio de Mingers (1989) concluye:
 - Los árboles obtenidos con distintos criterios de selección de splits no difieren significativamente en su poder predictivo
 - El criterio sí afecta el tamaño del árbol obtenido
 - La elección aleatoria de atributos produce árboles de alrededor del doble de tamaño que si utilizamos una medida de impureza
 - Entre las medidas vistas, Gain ratio produce árboles pequeños y Chi-cuadrado produce árboles grandes

Referencias

(no son obligatorias para el curso)

L. Breiman.

Technical note: some properties of splitting criteria.

Machine Learning, 24:41–47, 1996.

J. Mingers.

An empirical comparison of selection measures for decision-tree induction.

Machine Learning, 3:319–342, 1989.

S. Rogic.

An overview of methods for decision tree induction.

Term project for CS 532B.

K.V. Sreerama.

On growing better decision trees from data.

PhD thesis, The Johns Hopkins University, 1995.