Tema 7: Aprendizaje de árboles de decisión

José A. Alonso Jiménez
Miguel A. Gutiérrez Naranjo
Francisco J. Martín Mateos
José L. Ruiz Reina

Dpto. de Ciencias de la Computación e Inteligencia Artificial

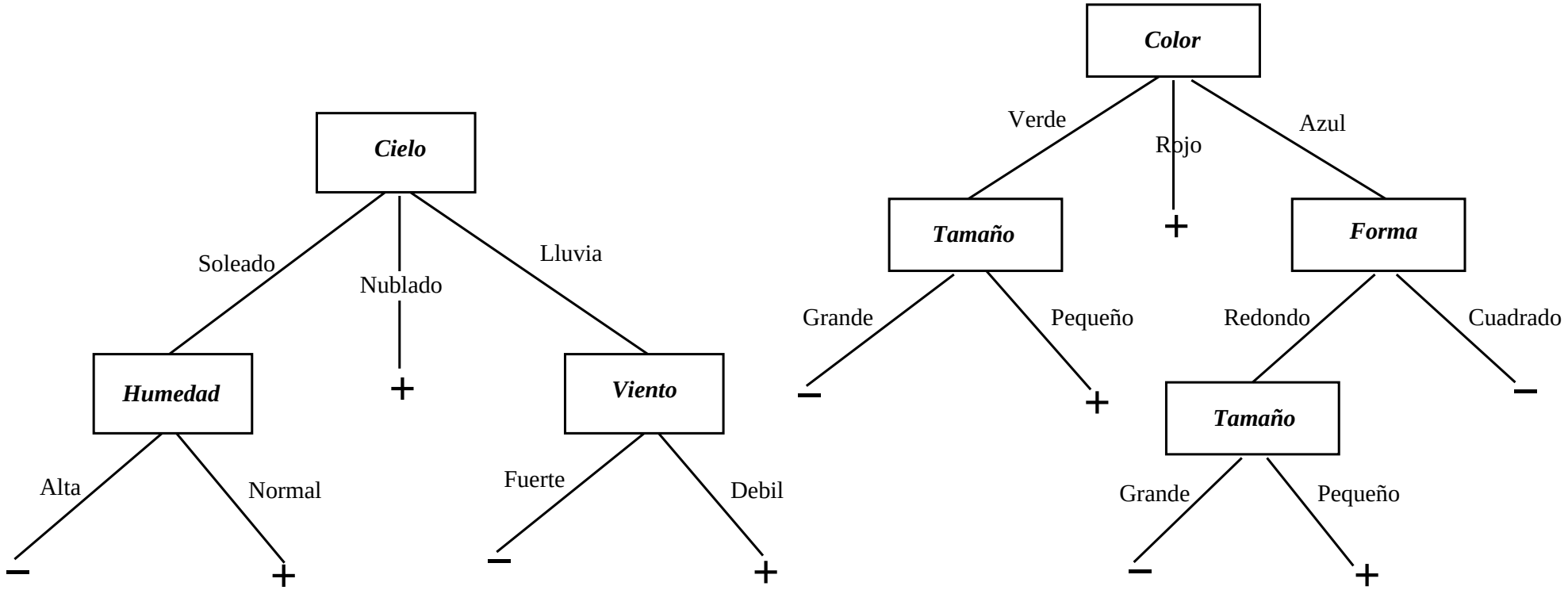
Universidad de Sevilla

Contenido

- Árboles de decisión
- El algoritmo ID3
- Entropía e información
- Ejemplos
- Búsqueda y sesgo inductivo
- Sobreajuste y ruido
- Poda
- Otras cuestiones

Árboles de decisión

- **Ejemplos de árboles de decisión**



Árboles de decisión

- **Árboles de decisión**
  - **Nodos interiores: atributos**
  - **Arcos: posibles valores del nodo origen**
  - **Hojas: valor de clasificación (usualmente + ó −, aunque podría ser cualquier conjunto de valores, no necesariamente binario)**
  - **Representación de una función objetivo**
- **Disyunción de reglas proposicionales:**

$$
\begin{aligned}
& (Cielo = Soleado \wedge Humedad = Alta \rightarrow Jugar\_Tenis = -) \\
\vee\ & (Cielo = Soleado \wedge Humedad = Normal \rightarrow Jugar\_Tenis = +) \\
\vee\ & (Cielo = Nublado \rightarrow Jugar\_Tenis = +) \\
\vee\ & (Cielo = Lluvioso \wedge Viento = Fuerte \rightarrow Jugar\_Tenis = -) \\
\vee\ & (Cielo = Lluvioso \wedge Viento = Debil \rightarrow Jugar\_Tenis = +)
\end{aligned}
$$

- **Capaz de representar cualquier subconjunto de instancias**

Aprendizaje de árboles de decisión

- **Objetivo: aprender un árbol de decisión consistente con los ejemplos**
  - **Para posteriormente clasificar ejemplos nuevos**
- **Ejemplos de conjuntos de entrenamiento:**

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

Ej.	Color	Forma	Tamaño	Clase
O_1	Rojo	Cuadrado	Grande	+
O_2	Azul	Cuadrado	Grande	+
O_3	Rojo	Redondo	Pequeño	-
O_4	Verde	Cuadrado	Pequeño	-
O_5	Rojo	Redondo	Grande	+
O_6	Verde	Cuadrado	Grande	-

Algoritmo ID3

- **ID3(Ejemplos, Atributo-objetivo, Atributos)**

```
1. Si todos los Ejemplos son positivos, devolver un nodo etiquetado con +
2. Si todos los Ejemplos son negativos, devolver un nodo etiquetado con -
3. Si Atributos está vacío, devolver un nodo etiquetado con el valor más
   frecuente de Atributo-objetivo en Ejemplos.
4. En otro caso:
   4.1. Sea A el atributo de Atributos que MEJOR clasifica Ejemplos
   4.2. Crear Árbol, con un nodo etiquetado con A.
   4.3. Para cada posible valor v de A, hacer:
     * Añadir un arco a Árbol, etiquetado con v.
     * Sea Ejemplos(v) el subconjunto de Ejemplos con valor
       del atributo A igual a v.
     * Si Ejemplos(v) es vacío:
       - Entonces colocar debajo del arco anterior un nodo etiquetado con
         el valor más frecuente de Atributo-objetivo en Ejemplos.
       - Si no,  colocar debajo del arco anterior el subárbol
         ID3(Ejemplos(v), Atributo-objetivo, Atributos-{A}).
   4.4  Devolver Árbol
```

¿Cómo saber qué atributo clasifica mejor?

- **Entropía de un conjunto de ejemplos D (resp. de una clasificación):**

$$Ent(D) = -\frac{|P|}{|D|} \cdot log_2 \frac{|P|}{|D|} - \frac{|N|}{|D|} \cdot log_2 \frac{|N|}{|D|}$$

  donde P y N son, resp., los subconjuntos de ejemplos positivos y negativos de D

  - **Notación:** $Ent([p+, n-])$**, donde** $p = |P|$ **y** $n = |N|$

- **Intuición:**
  - **Mide la ausencia de "homegeneidad" de la clasificación**
  - **Teoría de la Información: cantidad media de información (en bits) necesaria para codificar la clasificación de un ejemplo de** D

- **Ejemplos:**
  - $Ent([9+, 5-]) = -\frac{9}{14} \cdot log_2 \frac{9}{14} - \frac{5}{14} \cdot log_2 \frac{5}{14} = 0.94$
  - $Ent([k+, k-]) = 1$ **(ausencia total de homogeneidad)**
  - $Ent([p+, 0]) = Ent([0, n-]) = 0$ **(homogeneidad total)**

Ganancia de información

- **Preferimos nodos con menos entropía (árboles pequeños)**
- **Entropía *esperada* después de usar un atributo A en el árbol:**

$$\sum_{v \in Valores(A)} \frac{|D_v|}{|D|} \cdot Ent(D_v)$$

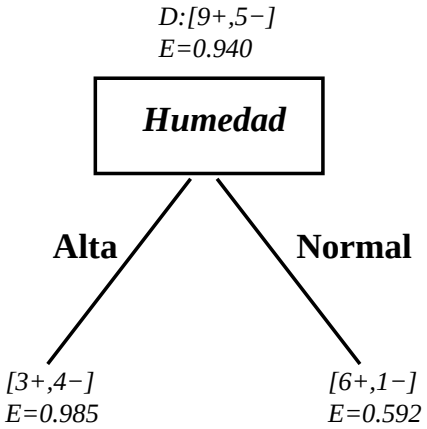
donde D_v es el subconjunto de ejemplos de D con valor del atributo A igual a v

- **Ganancia de información *esperada* después de usar un atributo A:**

$$Ganancia(D, A) = Ent(D) - \sum_{v \in Valores(A)} \frac{|D_v|}{|D|} \cdot Ent(D_v)$$

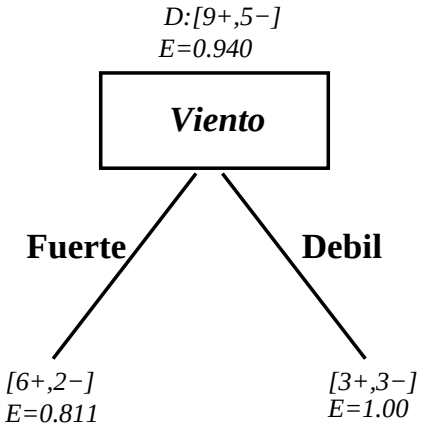
- **En el algoritmo ID3, en cada nodo usamos el atributo con mayor ganancia de información (considerando los ejemplos correspondientes al nodo)**

Algoritmo ID3 (ejemplo 1)



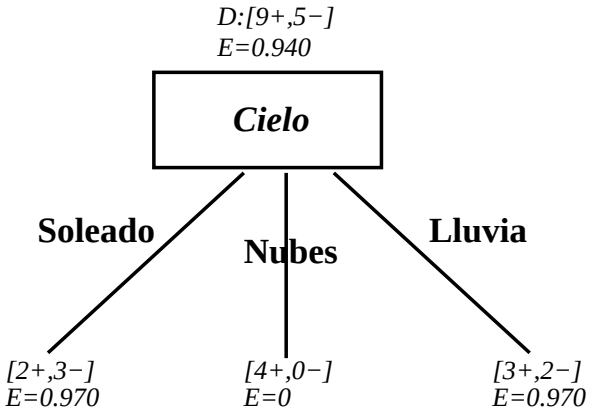
Ganancia(D,Humedad)=

0.94−(7/14)·0.985−(7/14)·0.592=0.151



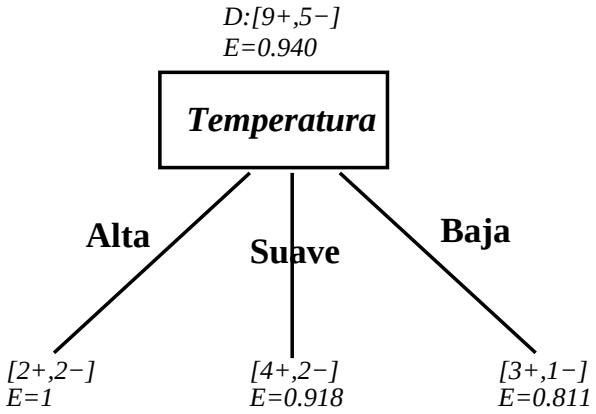
Ganancia(D,Viento)=

0.94−(8/14)·0.811−(6/14)·1=0.048



Ganancia(D,Cielo)=

0.94−(5/14)·0.97−(4/14)·0−(5/14)·0.97=0.246



Ganancia(D,Temperatura)=

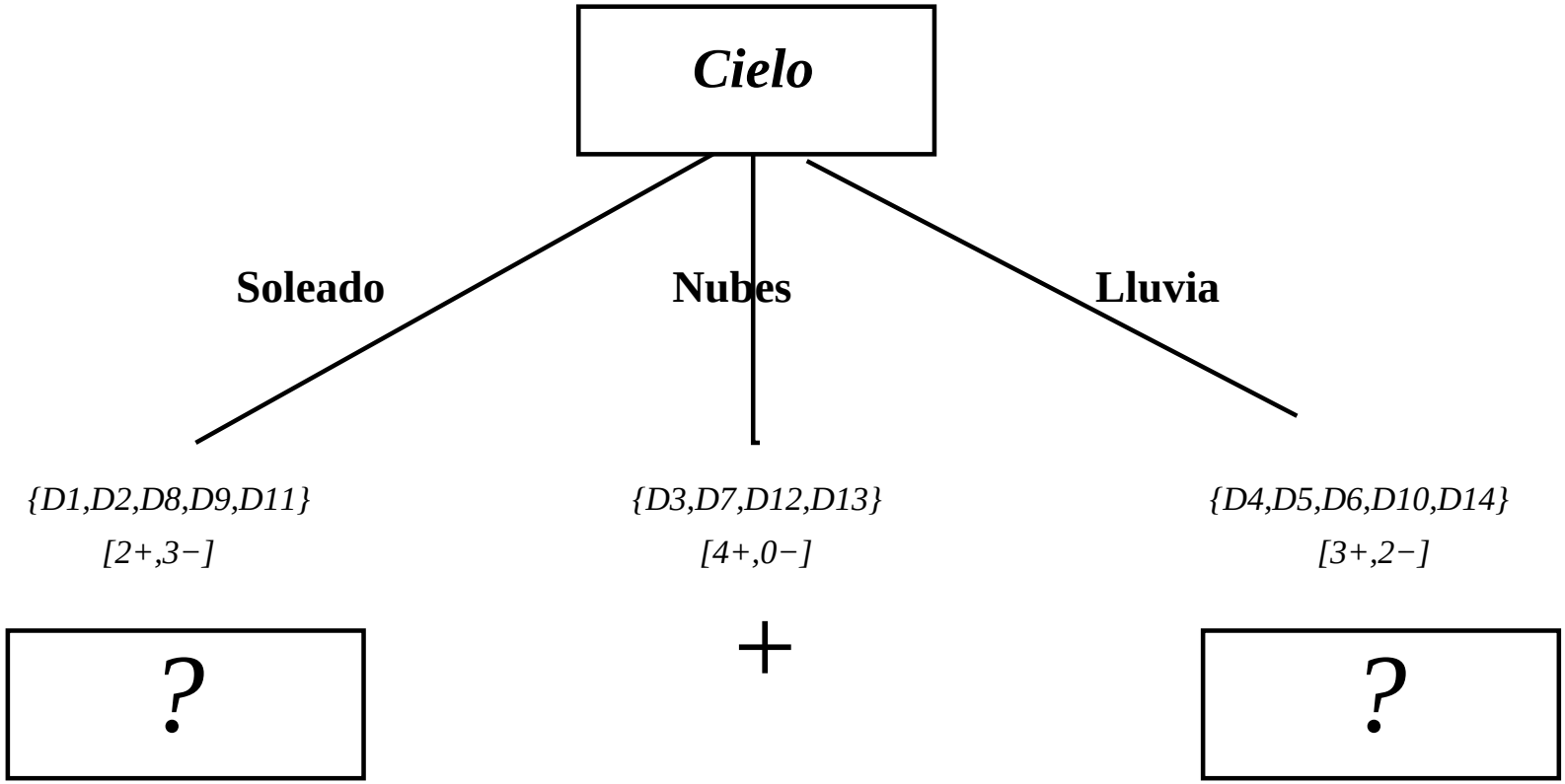
0.94−(4/14)·1−(6/14)·0.918−(4/14)·0.811=0.02

Algoritmo ID3 (ejemplo 1)

- **Entropía inicial en el ejemplo** $Jugar\_tenis$**,** $Ent([9+, 5-]) = 0.94$
- **Selección del atributo para el nodo raiz:**
  - $Ganancia(D, Humedad) = 0.94 - \frac{7}{14} \cdot 0.985 - \frac{7}{14} \cdot 0.592 = 0.151$
  - $Ganancia(D, Viento) = 0.94 - \frac{8}{14} \cdot 0.811 - \frac{6}{14} \cdot 1 = 0.048$
  - $Ganancia(D, Cielo) = 0.94 - \frac{5}{14} \cdot 0.970 - \frac{4}{14} \cdot 0 - \frac{5}{14} \cdot 0.970 = 0.246$ **(mejor atributo)**
  - $Ganancia(D, Temperatura) = 0.94 - \frac{4}{14} \cdot 1 - \frac{6}{14} \cdot 0.918 - \frac{4}{14} \cdot 0.811 = 0.02$
  - **Se selecciona el atributo** $Cielo$**, que es el que produce mayor ganacia de información**

Algoritmo ID3 (ejemplo 1)

- **Árbol parcialmente construido:**

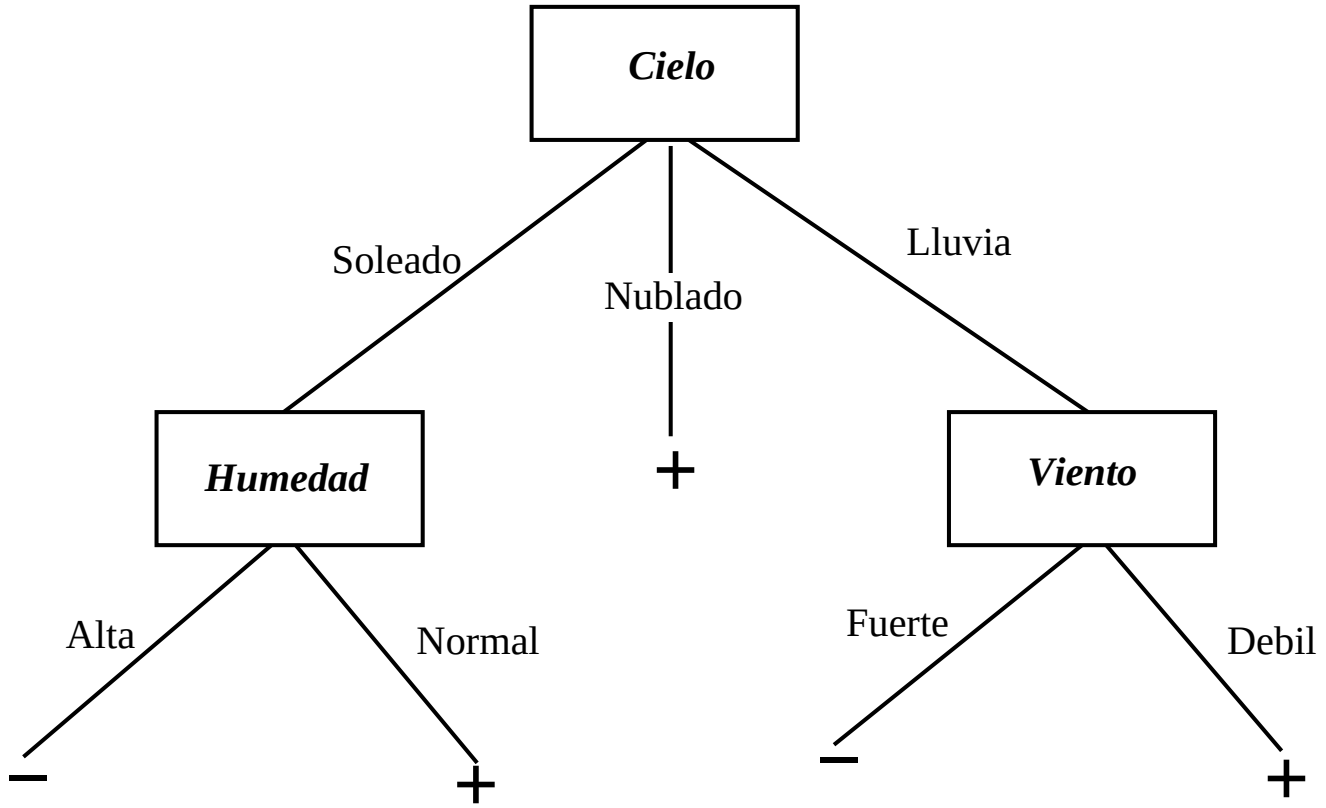


Algoritmo ID3 (ejemplo 1)

- **Selección del atributo para el nodo $Cielo = Sol$:**
  - $D_{Sol} = \{D_1, D_2, D_8, D_9, D_{11}\}$ **con entropía** $Ent([2+, 3-]) = 0.971$
  - $Ganancia(D_{Sol}, Humedad) = 0.971 - \frac{3}{5} \cdot 0 - \frac{2}{5} \cdot 0 = 0.971$ **(mejor atributo)**
  - $Ganancia(D_{Sol}, Temperatura) = 0.971 - \frac{2}{5} \cdot 0 - \frac{2}{5} \cdot 1 - \frac{1}{5} \cdot 0 = 0.570$
  - $Ganancia(D_{Sol}, Viento) = 0.971 - \frac{2}{5} \cdot 1 - \frac{3}{5} \cdot 0.918 = 0.019$
- **Selección del atributo para el nodo $Cielo = Lluvia$:**
  - $D_{Lluvia} = \{D_4, D_5, D_6, D_{10}, D_{14}\}$ **con entropía** $Ent([3+, 2-]) = 0.971$
  - $Ganancia(D_{Lluvia}, Humedad) = 0.971 - \frac{2}{5} \cdot 1 - \frac{3}{5} \cdot 0.918 = 0.820$
  - $Ganancia(D_{Lluvia}, Temperatura) = 0.971 - \frac{3}{5} \cdot 0.918 - \frac{2}{5} \cdot 1 = 0.820$
  - $Ganancia(D_{Lluvia}, Viento) = 0.971 - \frac{3}{5} \cdot 0 - \frac{2}{5} \cdot 0 = 0.971$ **(mejor atributo)**

Algoritmo ID3 (ejemplo 1)

- **Árbol finalmente aprendido:**

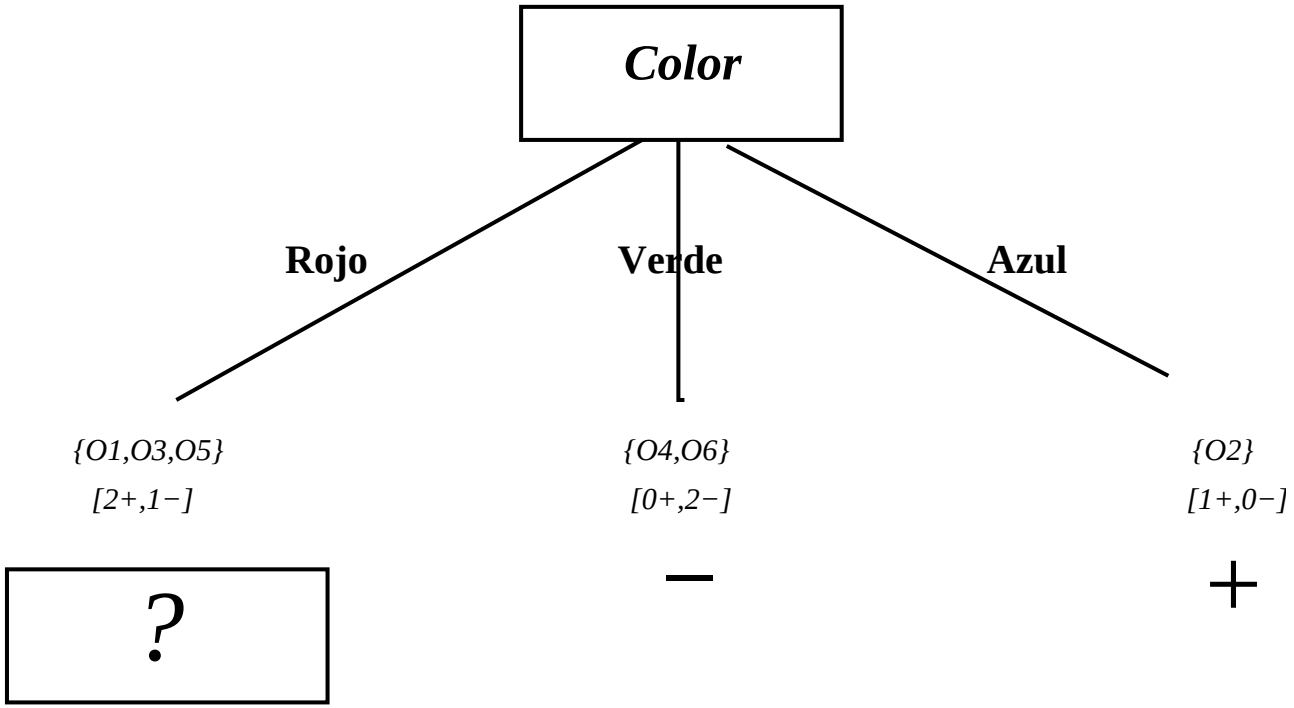


Algoritmo ID3 (ejemplo 2)

- **Entropía inicial en el ejemplo de los objetos,** $Ent([3+, 3-]) = 1$
- **Selección del atributo para el nodo raiz:**
  - $Ganancia(D, Color) = 1 - \frac{3}{6} \cdot Ent([2+, 1-]) - \frac{1}{6} \cdot Ent([1+, 0-]) - \frac{2}{6} \cdot Ent([0+, 2-]) = 0.543$
  - $Ganancia(D, Forma) = 1 - \frac{4}{6} \cdot Ent([2+, 2-]) - \frac{2}{6} \cdot Ent([1+, 1-]) = 0$
  - $Ganancia(D, Tamano) = 1 - \frac{4}{6} \cdot Ent([3+, 1-]) - \frac{2}{6} \cdot Ent([0+, 2-]) = 0.459$
  - **El atributo seleccionado es** $Color$

Algoritmo ID3 (ejemplo 2)

- **Árbol parcialmente construido:**

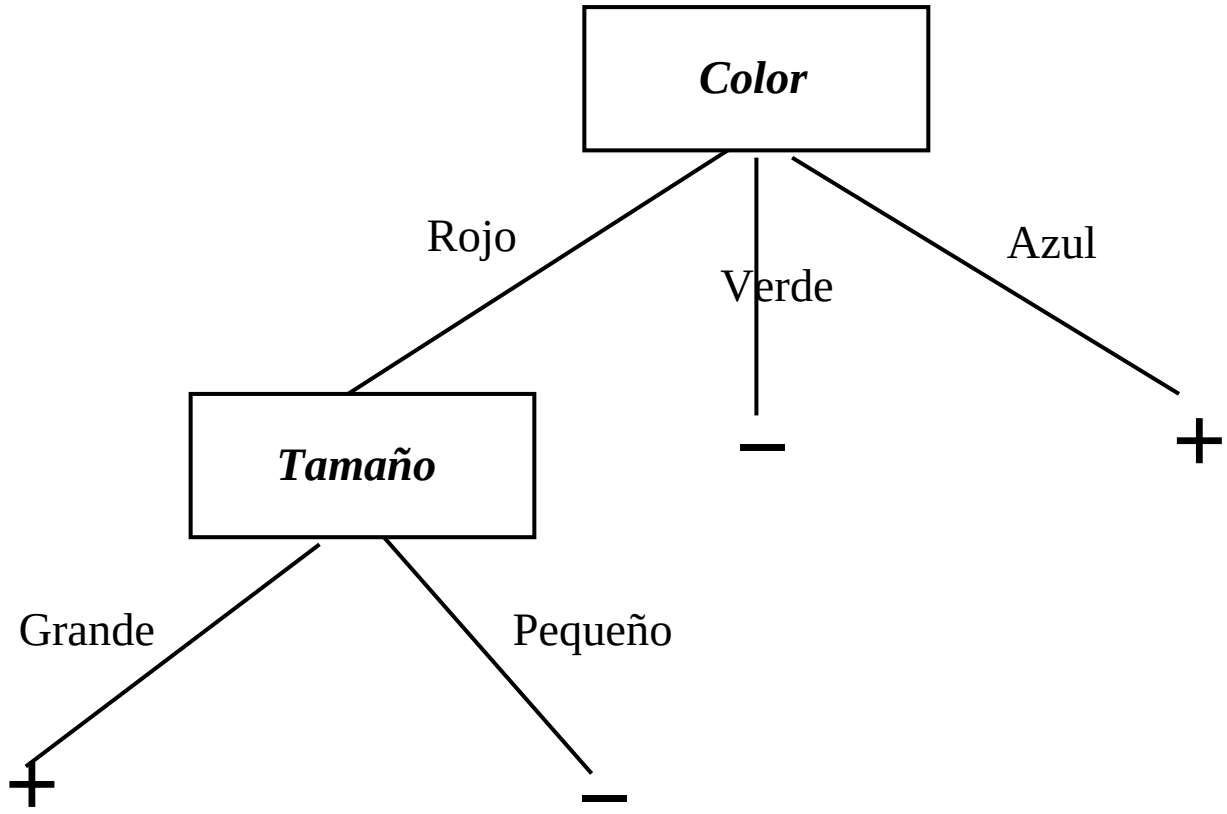


Algoritmo ID3 (ejemplo 2)

- **Selección del atributo para el nodo $Color = Rojo$:**
  - $D_{Rojo} = \{O_1, O_3, O_5\}$ **con entropía** $Ent([2+, 1-]) = 0.914$
  - $Ganancia(D_{Rojo}, Forma) = 0.914 - \frac{1}{3} \cdot Ent([1+, 0-]) - \frac{2}{3} \cdot Ent([1+, 1-]) = 0.247$
  - $Ganancia(D_{Rojo}, Tamano) = 0.914 - \frac{2}{3} \cdot Ent([2+, 0-]) - \frac{1}{3} \cdot Ent([0+, 1-]) = 0.914$
  - **El atributo seleccionado es *Tamaño***

Algoritmo ID3 (ejemplo 2)

- **Árbol finalmente aprendido:**



Búsqueda y sesgo inductivo

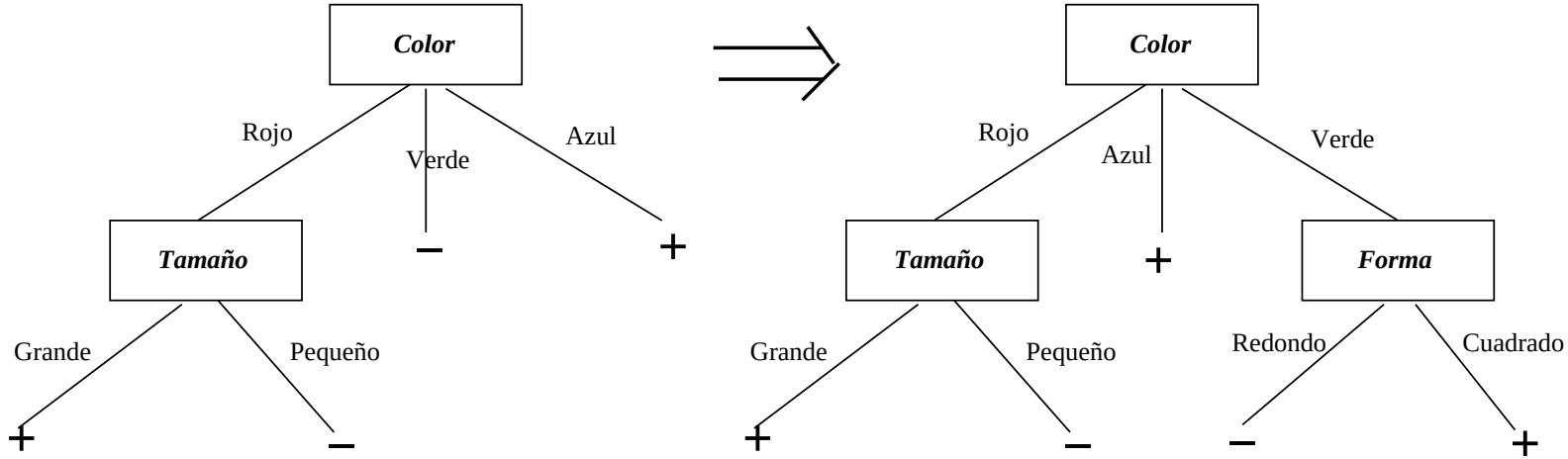
- **Búsqueda en un espacio de hipótesis**
  - **Espacio de hipótesis completo**
  - **Un único árbol candidato en cada paso**
  - **Sin retroceso (peligro de óptimos locales), búsqueda en escalada**
  - **Decisiones tomadas a partir de conjuntos de ejemplos**
- **Sesgo inductivo**
  - **Se prefieren árboles más cortos sobre los más largos**
  - **Sesgo preferencial, implícito en la búsqueda**
  - **Principio de la navaja de Occam**

Medida del rendimiento del aprendizaje

- **Conjunto de entrenamiento y conjunto de prueba**
  - **Aprender con el conjunto de entrenamiento**
  - **Medida del rendimiento: proporción de ejemplos bien clasificados en el conjunto de prueba**
- **Repetición de este proceso**
  - **Curva de aprendizaje**
  - **Estratificación: cada clase correctamente representada en el entrenamiento y en la prueba**
- **Validación cruzada**
  - **Dividir en k partes, y hace k aprendizajes, cada uno de ellos tomando como prueba una de las partes y entrenamiento el resto. Finalmente hacer la media de los rendimientos.**
  - **En la práctica: validación cruzada, con $k = 10$ y estratificación**

Sobreajuste y ruido

- **Una hipótesis $h \in H$ *sobreajusta* los ejemplos de entrenamiento si existe $h' \in H$ que se ajusta peor que h a los ejemplos pero actúa mejor sobre la distribución completa de instancias.**
- ***Ruido:* ejemplos incorrectamente clasificados. Causa sobreajuste**
- **Ejemplo: supongamos que por error, se incluye el ejemplo $< Verde, Redondo, Pequeno >$ como ejemplo positivo**
- **El árbol aprendido en este caso sería (sobrejustado a los datos):**



Sobreajuste y ruido

- Otras causas de sobreajuste:
  - Atributos que en los ejemplos presentan una aparente regularidad pero que no son relevantes en realidad
  - Conjuntos de entrenamiento pequeños
- Maneras de evitar el sobreajuste:
  - Parar el desarrollo del árbol antes de que se ajuste perfectamente a todos los datos
  - Podar el árbol *a posteriori*
- Poda *a posteriori*, dos aproximaciones:
  - Transformación a reglas, podado de las condiciones de las reglas
  - Realizar podas directamente en el árbol
  - Las podas se producen siempre que reduzcan el error sobre un conjunto de prueba

Podado de árboles

- **Un algoritmo de poda para reducir el error**

```
1. Dividir el conjunto de ejemplos en Entrenamiento y Prueba
2. Árbol=arbol obtenido por ID3 usando Entrenamiento
3. Medida = proporción de ejemplos en Prueba correctamente clasificados por Árbol
   Continuar=True
4. Mientras Continuar:
      * Por cada nodo interior N de Árbol:
         - Podar temporalmente Árbol en el nodo N y sustituirlo por una
           hoja etiquetada con la clasificación mayoritaria en ese nodo
         - Medir la proporción de ejemplos correctamente clasificados
           en el conjunto de prueba.
      * Sea K el nodo cuya poda produce mejor rendimiento
      *  Si este rendimiento es mejor que Medida, entonces
         Árbol = resultado de podar permanentemente Árbol en K
      * Si no, Continuar=Falso
5. Devolver Árbol
```

Otra cuestiones prácticas del algoritmo ID3

- **Extensiones del algoritmo:**
  - **Atributos con valores contínuos**
  - **Otras medidas para seleccionar atributos**
  - **Otras estimaciones de error**
  - **Atributos sin valores**
  - **Atributos con coste**
- **Algoritmos C4.5 y C5.0 (Quinlan)**

Bibliografía

- **Mitchell, T.M. *Machine Learning* (McGraw-Hill, 1997)**
  - **Cap. 3: "Decision tree learning"**
- **Russell, S. y Norvig, P. *Inteligencia artificial (Un enfoque moderno)* (Prentice–Hall Hispanoamericana, 1996)**
  - **Cap. 18: "Aprendiendo de observaciones"**
- **Witten, I.H. y Frank, E. *Data mining* (Morgan Kaufmann Publishers, 2000)**
  - **Cap. 3: "Output: Knowledge representation"**
  - **Cap. 4: "Algorithms: The basic methods"**
  - **Cap. 5: "Credibility: Evaluating what's has been learned"**
  - **Cap. 6: "Implementations: Real machine learning schemes"**