Árboles de decisión

Verónica E. Arriola-Rios

Inteligencia Artificial

11 de mayo de 2020

Temas

Árboles de decisión

- Dado un conjunto de atributos y sus valores, determinar si un ejemplar pertenece o no a una *clase*.
- Generar una jerarquía: preguntar primero por los atributos que dan más información.

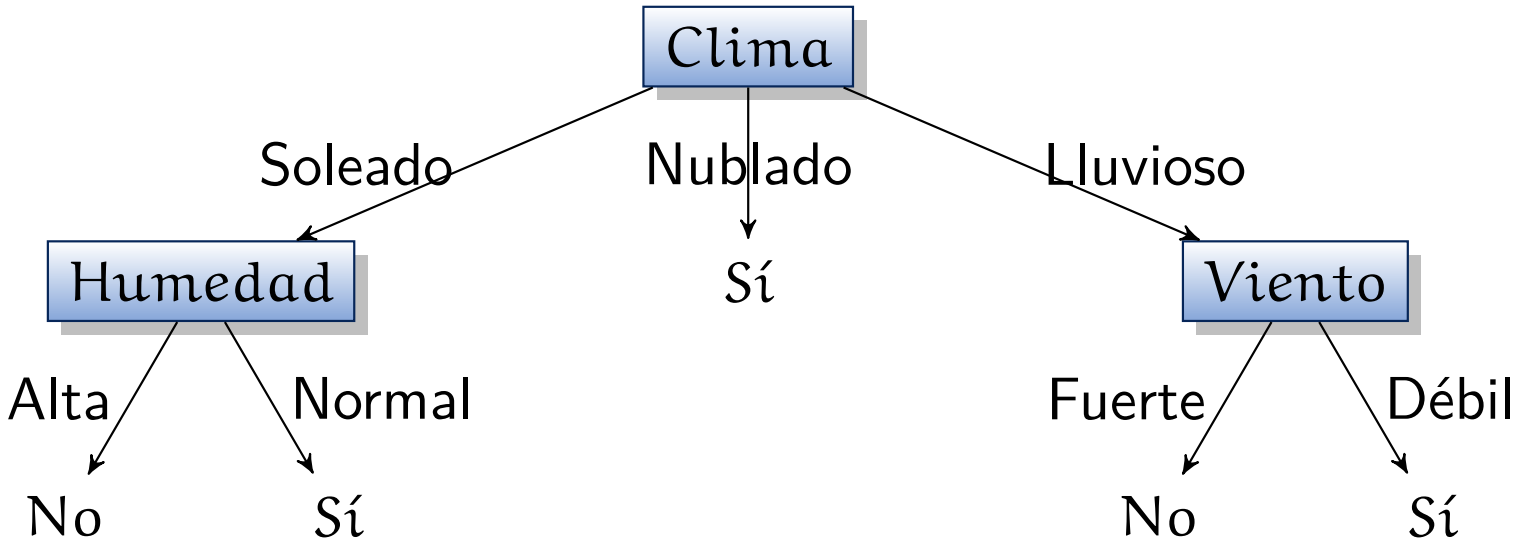


Figura: ¿Es un buen día para jugar tenis?

Ejemplo: ¿Adivina quién?

Figura: Tratemos de hacer la menor cantidad de preguntas. Mayor riesgo: características particulares; menor riesgo: dividir mitad y mitad.

¿Qué apreden?

Un árbol de decisión aprende conceptos de la forma:

- Buen día para jugar tenis :=
  - (Clima = Soleado $\wedge$ Humedad = Normal) $\vee$
  - (Clima = Nublado) $\vee$
  - (Clima = Lluvioso $\wedge$ Viento = Débil)

Ejemplo

Ejemplo tomado de Mitchell 1997.

Día	Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
D1	Soleado	Cálido	Alta	Débil	No
D2	Soleado	Cálido	Alta	Fuerte	No
D3	Nublado	Cálido	Alta	Débil	Sí
D4	Lluvioso	Templado	Alta	Débil	Sí
D5	Lluvioso	Frío	Normal	Débil	Sí
D6	Lluvioso	Frío	Normal	Fuerte	No
D7	Nublado	Frío	Normal	Fuerte	Sí
D8	Soleado	Templado	Alta	Débil	No
D9	Soleado	Frío	Normal	Débil	Sí
D10	Lluvioso	Templado	Normal	Débil	Sí
D11	Soleado	Templado	Normal	Fuerte	Sí
D12	Nublado	Templado	Alta	Fuerte	Sí
D13	Nublado	Cálido	Normal	Débil	Sí
D14	Lluvioso	Templado	Alta	Fuerte	No

Temas

Temas

Objetivo

- Se ve como un problema de búsqueda en el espacio de hipótesis:

  El espacio de hipótesis es el conjunto de todos los árboles de decisión posibles.

- El objetivo óptimo sería encontrar el árbol que permita encontrar la clasificación correcta para cada ejemplar con el menor número de preguntas.
- Los requerimientos se pueden relajar de las formas siguientes:
  - Encontrar un árbol que realize pocas preguntas (aunque no sea el más corto).
  - Encontrar el que cometa la menor cantidad de errores posibles, en caso de que los atributos utilizados no permitan separar a los ejemplares unívocamente.

Temas

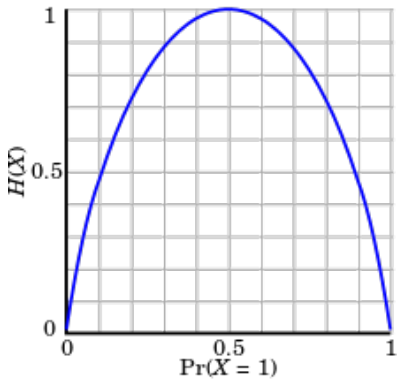
Entropía

Sea D una colección de muestras.

- La *entropía* de D especifica el número mínimo de bits necesario para codificar la clasificación de cualquier miembro de D. Para un atributo clase con c diferentes valores posibles:

$$\text{Entropía}(D) \equiv \sum_{i=1}^{c} -p_i \log_2(p_i) \tag{1}$$

donde p_i es la proporción de muestras en D que pertenecen a la clase i.



Ejemplo

Día	Clima	Temperatura	Humedad	Viento	¿Jugar tenis?	
D1	Soleado	Cálido	Alta	Débil	No	
D2	Soleado	Cálido	Alta	Fuerte	No	
D3	Nublado	Cálido	Alta	Débil	Sí	1
D4	Lluvioso	Templado	Alta	Débil	Sí	2
D5	Lluvioso	Frío	Normal	Débil	Sí	3
D6	Lluvioso	Frío	Normal	Fuerte	No	
D7	Nublado	Frío	Normal	Fuerte	Sí	4
D8	Soleado	Templado	Alta	Débil	No	
D9	Soleado	Frío	Normal	Débil	Sí	5
D10	Lluvioso	Templado	Normal	Débil	Sí	6
D11	Soleado	Templado	Normal	Fuerte	Sí	7
D12	Nublado	Templado	Alta	Fuerte	Sí	8
D13	Nublado	Cálido	Normal	Débil	Sí	9
D14	Lluvioso	Templado	Alta	Fuerte	No	

$$
\begin{aligned}
\text{Entropía}(D) &\equiv \sum_{i=1}^{c} -p_i \log_2(p_i) \\
&= -p_+ \log_2 p_+ - p_- \log_2 p_- \\
&= -\frac{9}{14} \log_2 \left(\frac{9}{14}\right) - \frac{5}{14} \log_2 \left(\frac{5}{14}\right) \\
&= -\frac{9}{14} \frac{\ln\left(\frac{9}{14}\right)}{\ln(2)} - \frac{5}{14} \frac{\ln\left(\frac{5}{14}\right)}{\ln(2)} \\
&= 0.940
\end{aligned}
$$

¿Cómo se genera el árbol?

- ¿Qué tanta información se gana al preguntar por el valor de un atributo A?

$$\text{Ganancia}(D, A) \equiv \text{Entropía}(D) - \sum_{v \in Valores(A)} \frac{|S_V|}{|S|} \text{Entropía}(S_V) \tag{2}$$

Donde el segundo término es la suma de las entropías de cada subconjunto S_V, pesado por la fracción de muestras $\frac{|S_V|}{|S|}$ que pertenecen a S_V.

Ej. ¿Cuánto se gana si se conoce la temperatura?

Día	Clima	Temperatura	Humedad	Viento	¿Jugar tenis?
D1	Soleado	Cálido	Alta	Débil	No
D2	Soleado	Cálido	Alta	Fuerte	No
D3	Nublado	Cálido	Alta	Débil	Sí
D4	Lluvioso	Templado	Alta	Débil	Sí
D5	Lluvioso	Frío	Normal	Débil	Sí
D6	Lluvioso	Frío	Normal	Fuerte	No
D7	Nublado	Frío	Normal	Fuerte	Sí
D8	Soleado	Templado	Alta	Débil	No
D9	Soleado	Frío	Normal	Débil	Sí
D10	Lluvioso	Templado	Normal	Débil	Sí
D11	Soleado	Templado	Normal	Fuerte	Sí
D12	Nublado	Templado	Alta	Fuerte	Sí
D13	Nublado	Cálido	Normal	Débil	Sí
D14	Lluvioso	Templado	Alta	Fuerte	No

$$\begin{aligned}
&\text{Ganancia}(D, \text{Temperatura}) \\
&\quad = \text{Entropía}(D) - \sum_{v \in \{\text{Cálido}, \text{Templado}, \text{Frío}\}} \frac{|S_v|}{|S|} \text{Entropía}(S_v) \\
&\quad = 0.940 - \frac{4}{14} \text{Entropía}(D_{\text{Cálido}}) - \frac{6}{14} \text{Entropía}(D_{\text{Templado}}) \\
&\qquad - \frac{4}{14} \text{Entropía}(D_{\text{Frío}}) \qquad (3)
\end{aligned}$$

$$\text{Entropía}(D_{\text{Cálido}}) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1.0 \quad (4)$$

$$\text{Entropía}(D_{\text{Templado}}) = -\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6} = 0.9183 \quad (5)$$

$$\text{Entropía}(D_{\text{Frío}}) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = 0.8113 \quad (6)$$

Sustituyendo en (3):

$$\begin{aligned}
&\text{Ganancia}(D, \text{Temperatura}) \\
&= 0.940 - \frac{4}{14}1.0 - \frac{6}{14}0.9183 - \frac{4}{14}0.8113 \\
&= 0.940 - 0.9111 \\
&= 0.0289
\end{aligned} \quad (7)$$

ID3

ID3

Va seleccionando los atributos que producen la mayor ganancia...

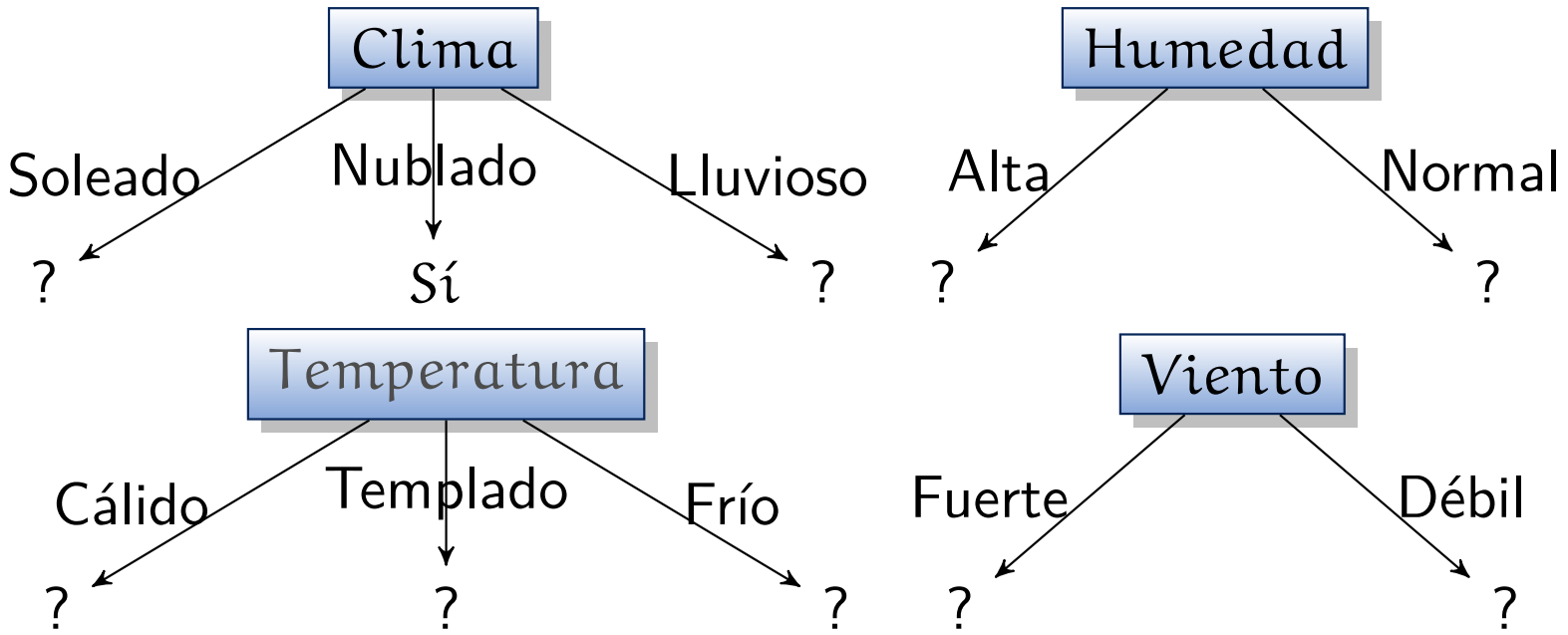


Figura: ¿Es un buen día para jugar tenis?

ID3

ID3

Va seleccionando los atributos que producen la mayor ganancia...

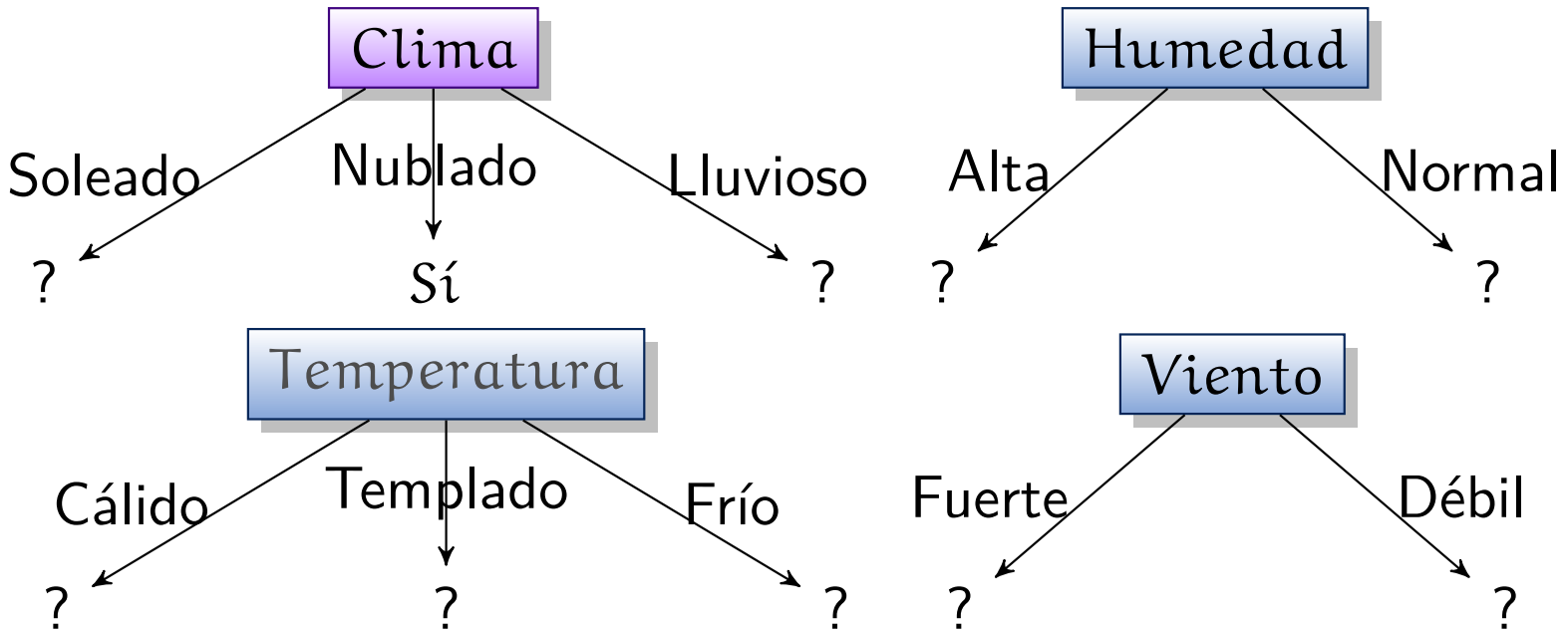


Figura: ¿Es un buen día para jugar tenis?

Clima=Soleado

Día	Temperatura	Humedad	Viento	¿Jugar tenis?
D1	Cálido	Alta	Débil	No
D2	Cálido	Alta	Fuerte	No
D8	Templado	Alta	Débil	No
D9	Frío	Normal	Débil	Sí
D11	Templado	Normal	Fuerte	Sí

Búsqueda de una hipótesis

- El algoritmo ID3 realiza una escalada de colinas, a partir del árbol más sencillo (el árbol vacío), hacia árboles más complejos hasta encontrar uno que clasifique correctamente a todas las muestras.
- La función de evaluación que dirige la escalada es $\mathrm{Ganancia}(\mathrm{D}, \mathrm{A})$.

Referencias I

 Mitchell, Tom M. (1997). *Machine Learning*. McGrawHill.