Algoritmo ID3: Introducción y Ejemplo Didáctico

Profesor Cosijopii García

January 15, 2025

¿Qué es el Algoritmo ID3?

- ID3 (Iterative Dichotomiser 3) es un algoritmo para construir árboles de decisión.
- Utiliza un enfoque voraz para seleccionar el atributo que mejor clasifica los datos en cada paso.
- Basado en los conceptos de **entropía** y **ganancia de información**.

Concepto Clave: Entropía

- La entropía mide la incertidumbre de una variable aleatoria.
- Fórmula general:

$$H(S) = -\sum_{i=1}^{n} p_i \log_2(p_i),$$

  donde p_i es la probabilidad de cada clase en el conjunto S.
- Valores de entropía:
  - $H = 0$: Conjunto homogéneo (todas las muestras son de la misma clase).
  - $H > 0$: Conjunto heterogéneo (mezcla de clases).

Concepto Clave: Ganancia de Información

- La ganancia de información mide la reducción en la entropía al dividir un conjunto de datos según un atributo.
- Fórmula:

$$\text{Ganancia}(S, A) = H(S) - \sum_{v \in \text{Valores}(A)} \frac{|S_v|}{|S|} H(S_v),$$

donde:

- S: Conjunto original de datos.
- A: Atributo usado para dividir S.
- S_v: Subconjunto donde el atributo A toma el valor v.

Ejemplo: Construcción del Árbol de Decisión

Conjunto de datos:

Clima	Temperatura	Humedad	Viento	Jugar
Soleado	Calor	Alta	Débil	No
Soleado	Calor	Alta	Fuerte	No
Nublado	Calor	Alta	Débil	Sí
Lluvia	Templado	Alta	Débil	Sí
Lluvia	Frío	Normal	Débil	Sí
Lluvia	Frío	Normal	Fuerte	No
Nublado	Frío	Normal	Fuerte	Sí
Soleado	Templado	Alta	Débil	No
Soleado	Frío	Normal	Débil	Sí
Lluvia	Templado	Normal	Débil	Sí
Soleado	Templado	Normal	Fuerte	Sí
Nublado	Templado	Alta	Fuerte	Sí
Nublado	Calor	Normal	Débil	Sí
Lluvia	Templado	Alta	Fuerte	No

Cálculo de la Entropía Inicial

Frecuencia de la clase Jugar:

- **Sí**: 9 ejemplos.
- **No**: 5 ejemplos.

Entropía del conjunto S:

$$H(S) = -\left(\frac{9}{14}\log_2\frac{9}{14} + \frac{5}{14}\log_2\frac{5}{14}\right)$$

$$H(S) \approx 0.94$$

Cálculo de la Ganancia de Información (Atributo: Clima) I

División por valores del atributo Clima:

- **Soleado**: $|S_{\text{Soleado}}| = 5$ (No: 3, $Sí$: 2).
- **Nublado**: $|S_{\text{Nublado}}| = 4$ ($Sí$: 4).
- **Lluvia**: $|S_{\text{Lluvia}}| = 5$ (No: 2, $Sí$: 3).

Entropías por subconjuntos:

$$H(S_{\text{Soleado}}) = -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) \approx 0.97$$

$$H(S_{\text{Nublado}}) = -\left(\frac{4}{4}\log_2\frac{4}{4}\right) = 0$$

$$H(S_{\text{Lluvia}}) = -\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right) \approx 0.97$$

Ganancia de información:

Cálculo de la Ganancia de Información (Atributo: Clima) II

$$
\begin{aligned}
\text{Ganancia}(S, \text{Clima}) &= H(S) - \left(\frac{5}{14} H(S_{\text{Soleado}}) + \frac{4}{14} H(S_{\text{Nublado}}) + \frac{5}{14} H(S_{\text{Lluvia}}) \right) \\
&= 0.94 - \left(\frac{5}{14}(0.97) + \frac{4}{14}(0) + \frac{5}{14}(0.97) \right) \\
&= 0.94 - 0.693 \approx 0.247
\end{aligned}
$$

Iteración del Algoritmo

- Seleccionamos el atributo con mayor ganancia de información como la raíz del árbol. En este caso, *Clima*.
- Repetimos el cálculo para cada nodo hijo hasta que:
  - Todos los ejemplos en un nodo tienen la misma clase, o
  - No quedan atributos por dividir.