Actividad 2 - Minería de Datos

Carmen Plata Fernández

28/03/2025

1. Realiza los cálculos para el resto de pasos del algoritmo hasta llegar a que todos los nodos sean de tipo hoja.

En el problema presentado en la tabla 2.1, se trataba de decidir si se debería jugar o no al tenis, basándose en los atributos relacionados con el clima, como la vista, la temperatura, la humedad y el viento.

Se ha determinado que el atributo que maximiza la ganancia de información es la *Vista*. Calculamos la ganancia de información para este atributo usando la siguiente fórmula:

$$\text{Ganancia de Información (Vista)} = -\left(\frac{5}{14}\log_2\left(\frac{5}{14}\right) + \frac{9}{14}\log_2\left(\frac{9}{14}\right)\right) = 0{,}9403$$

En este caso, tenemos que:

- $C = 2$ (jugadas posibles: "Sí" o "No")
- El número total de ejemplos es 14, con 9 ejemplos correspondientes a "Sí" y 5 a "No".

Al calcular la entropía para los valores de "Vista", se obtiene lo siguiente:

- Para "Soleado":

$$I(\text{Soleado}) = -\left(\frac{2}{5}\log_2\left(\frac{2}{5}\right) + \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right) = 0{,}9709$$

- Para "Nublado":

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

Por lo tanto, la ganancia de información para "Vista" es:

$$I(\text{Vista}) = \frac{5}{14} \times 0{,}9709 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0{,}9709 = 0{,}6935$$

La ganancia de información final para el atributo "Vista" es:

$$G(\text{Vista}) = 0{,}9403 - 0{,}6935 = 0{,}2468$$

Posteriormente, calculamos la ganancia de información para los atributos "Humedad", "Temperatura" y "Viento", resultando en los siguientes valores:

- "Humedad" = 0.151
- "Temperatura" = 0.029
- "Viento" = 0.048

Dado que la ganancia de información de "Vista" es la más alta, este se convierte en el primer nodo del árbol de decisión. Ahora, se dividen los datos en dos ramas:

1. **Rama 1**: Vista = Nublado En este caso, todos los ejemplos son "Sí", por lo que el nodo hoja es "Sí".
2. **Rama 2**: Vista = Soleado En este subconjunto, tenemos 5 ejemplos: 3 de ellos son "No" y 2 son "Sí". A continuación, calculamos la entropía para los atributos disponibles, comenzando con "Humedad", que tiene dos valores posibles:
   - *Humedad = Alta* (3 ejemplos → Todos "No") Entropía = 0 (puro).
   - *Humedad = Normal* (2 ejemplos → Todos "Sí") Entropía = 0 (puro).

   Por lo tanto, la ganancia de información para "Humedad" es 0.971. Continuamos con los demás atributos:
   - *Temperatura* tiene tres valores posibles:
     - Temperatura = Alta → (2 ejemplos → 1 "Sí", 1 "No") → Entropía = 1
     - Temperatura = Media → (2 ejemplos → 1 "Sí", 1 "No") → Entropía = 1
     - Temperatura = Baja → (1 ejemplo → Todos "Sí") → Entropía = 0

     La entropía total para "Temperatura" es 0.8, lo que nos da una ganancia de información de 0.171, que es menor que la de "Humedad". Por lo tanto, descartamos "Temperatura".
   - *Viento* tiene dos valores posibles:
     - Viento = No → (3 ejemplos → 2 "No", 1 "Sí") → Entropía = 0.918
     - Viento = Sí → (2 ejemplos → 1 "Sí", 1 "No") → Entropía = 1

     La ganancia de información para "Viento" es 0.021, que también es menor que la de "Humedad", por lo que descartamos "Viento".

   Finalmente, elegimos "Humedad" como el mejor atributo para dividir el conjunto "Soleado", con los siguientes resultados:
   - *Humedad Alta* → Nodo hoja "No"
   - *Humedad Normal* → Nodo hoja "Sí"

- **Rama 3**: Vista = Lluvioso Aquí, tenemos 5 ejemplos, de los cuales 3 son "Sí" y 2 son "No". Calculamos la entropía para el atributo "Viento" y obtenemos que es el mejor atributo para dividir este subconjunto, ya que la ganancia de información para "Viento" es la más alta. Al dividir con "Viento", obtenemos los siguientes nodos:
  - *Viento = No* → Nodo hoja "Sí"

- *Viento* = *Sí* → Nodo hoja "No"

De esta forma, el árbol de decisión queda completamente definido con todos los nodos convertidos en hojas, lo que asegura que la clasificación de los ejemplos sea precisa y eficiente.

2. Resuelve el problema anterior desarrollando los cálculos y explicando cada uno de los valores que aparecen en las expresiones para los diferentes cálculos.

Primero, organizamos los ejemplos según los valores del atributo Temperatura.

$$\text{Valores: } 64, 65, 68, 69, 70, 71, 72, 72, 75, 75, 80, 81, 83, 85$$

$$\text{Clases: } A, B, A, A, A, B, B, A, A, A, B, A, A, B$$

A continuación, identificamos los posibles puntos de división:

$$64{,}5, 66{,}5, 68{,}5, 69{,}5, 70{,}5, 71{,}5, 72, 73{,}5, 75, 77{,}5, 80{,}5, 82, 84$$

Eliminamos aquellos puntos que no son viables, considerando que cada subintervalo debe contener al menos dos ejemplos. En este caso, descartamos los puntos 64,5 y 84. Además, es importante no dividir el intervalo si el siguiente ejemplo pertenece a la misma clase.

$$66{,}5, 68{,}5, 69{,}5, 70{,}5, 71{,}5, 72, 73{,}5, 75, 77{,}5, 80{,}5, 82$$

Luego de eliminar los puntos marcados en rojo, se observa lo siguiente:

$$66{,}5, 70{,}5, 72, 77{,}5, 80{,}5$$

También debemos asegurarnos de no dividir el intervalo si el siguiente valor es igual al actual.

$$66{,}5, 70{,}5, 77{,}5, 80{,}5$$

Posteriormente, unimos subintervalos adyacentes que tengan la misma clase mayoritaria. En este proceso, eliminamos los puntos 70,5, ya que todo el intervalo es de clase A, y el punto 80,5 porque solo corresponde a un ejemplo.

Finalmente, los puntos candidatos a ser de corte son 66,5 y 77,5.

La fórmula para la información antes de aplicar el atributo es:

$$I = -\sum_{c=1}^{n_c} \frac{n_c}{n} \log_2 \left(\frac{n_c}{n} \right)$$

Donde:

- n_c es el número de ejemplos de la clase c, con dos clases: A y B, donde A = Sí juega y B = No juega.

- $n_A = 9$, $n_B = 5$
- $n = 14$ es el total de ejemplos.
- $n_c = 2$ es el número de clases.
- n_a es el número de valores del atributo A.
- $n_j(A)$ es el número de ejemplos con valor j en A.
- n_{ijk} es el número de ejemplos con valor j en A que pertenecen a la clase k.

El valor de la información inicial es:

$$I = -\left(\frac{5}{14}\log_2\left(\frac{5}{14}\right) + \frac{9}{14}\log_2\left(\frac{9}{14}\right)\right) = 0{,}9403$$

Tras realizar el análisis de los puntos posibles de división:
Los puntos de corte candidatos son 66,5, 70,5, 77,5 y 80,5.

Punto de corte en 66,5

Se dividen los datos en:

- $A \leq 66{,}5$ (2 ejemplos: 1 en A, 1 en B):

$$I(A \leq 66{,}5) = -\left(\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) = 1$$

- $A > 66{,}5$ (12 ejemplos: 8 en A, 4 en B):

$$I(A > 66{,}5) = -\left(\frac{8}{12}\log_2\left(\frac{8}{12}\right) + \frac{4}{12}\log_2\left(\frac{4}{12}\right)\right) = 0{,}9183$$

La información total después del corte en 66,5 es:

$$I(A \leq 66{,}5) = \frac{2}{14} \times 1 + \frac{12}{14} \times 0{,}9183 = 0{,}93$$

La ganancia de información para el punto 66,5 es:

$$G(A \leq 66{,}5) = 0{,}9403 - 0{,}93 = 0{,}0103$$

Punto de corte en 77,5

Se dividen los datos en:

- $A \leq 77{,}5$ (10 ejemplos: 7 en A, 3 en B):

$$I(A \leq 77{,}5) = -\left(\frac{7}{10}\log_2\left(\frac{7}{10}\right) + \frac{3}{10}\log_2\left(\frac{3}{10}\right)\right) = 0{,}8813$$

- $A > 77{,}5$ (4 ejemplos: 3 en A, 1 en B):

$$I(A > 77{,}5) = -\left(\frac{3}{4}\log_2\left(\frac{3}{4}\right) + \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right) = 0{,}8113$$

La información total después del corte en 77,5 es:

$$I(A \leq 77{,}5) = \frac{10}{14} \times 0{,}8813 + \frac{4}{14} \times 0{,}8113 = 0{,}8582$$

La ganancia de información para el punto 77,5 es:

$$G(A \leq 77{,}5) = 0{,}9403 - 0{,}8582 = 0{,}0821$$

2.1. Comparación de las ganancias de información:

$$G(A \leq 66{,}5) = 0{,}0103$$
$$G(A \leq 77{,}5) = 0{,}0251$$

El punto de corte seleccionado es 77,5 debido a que tiene una mayor ganancia de información.

3. Para el ejemplo anterior, realiza la clasificación para el siguiente ejemplo de test:

Datos:

Vista	Temperatura	Humedad	Viento	Jugar
Lluvioso	Media	Alta	Sí	?

De acuerdo con el clasificador Naive Bayesiano, la probabilidad de que una instancia pertenezca a la clase v_j dada una serie de atributos $a_1, a_2, \ldots, a_n$ se calcula como:

$$P(a_1, a_2, \ldots, a_n \mid v_j)P(v_j) = \prod_i P(a_i \mid v_j)P(v_j)$$

Los clasificadores Naive Bayesianos suponen que el impacto de cada atributo sobre la clase es independiente de los demás atributos.

Se tiene que:

$$P(\text{Jugar} = \text{Sí}) = \frac{9}{14}, \quad P(\text{Jugar} = \text{No}) = \frac{5}{14}$$

Para cada atributo, dado que Jugar = Sí, debemos estimar las probabilidades condicionales de que cada valor de atributo se dé bajo esa condición.

$$P(\text{Vista} = \text{Lluvioso} \mid \text{Jugar} = \text{Sí}) = \frac{3}{9}, \quad P(\text{Temperatura} = \text{Media} \mid \text{Jugar} = \text{Sí}) = \frac{4}{9}$$

$$P(\text{Humedad} = \text{Alta} \mid \text{Jugar} = \text{Sí}) = \frac{3}{9}, \quad P(\text{Viento} = \text{Sí} \mid \text{Jugar} = \text{Sí}) = \frac{3}{9}$$

Multiplicando las probabilidades, obtenemos:

$$P(\text{Jugar} = \text{Sí} \mid X) = P(\text{Sí}) \times \prod_i P(a_i \mid \text{Sí}) = \frac{9}{14} \times \frac{3}{9} \times \frac{4}{9} \times \frac{3}{9} \times \frac{3}{9} = 0{,}01058$$

Ahora, para cada atributo dado que Jugar = No, se tiene:

$$P(\text{Vista} = \text{Lluvioso} \mid \text{Jugar} = \text{No}) = \frac{2}{5}, \quad P(\text{Temperatura} = \text{Media} \mid \text{Jugar} = \text{No}) = \frac{2}{5}$$

$$P(\text{Humedad} = \text{Alta} \mid \text{Jugar} = \text{No}) = \frac{4}{5}, \quad P(\text{Viento} = \text{Sí} \mid \text{Jugar} = \text{No}) = \frac{3}{5}$$

Multiplicamos estas probabilidades:

$$P(\text{Jugar} = \text{No} \mid X) = \frac{5}{14} \times \frac{2}{5} \times \frac{2}{5} \times \frac{4}{5} \times \frac{3}{5} = 0{,}0274$$

Finalmente, calculamos las probabilidades posteriores:

$$P(\text{Jugar} = \text{Sí} \mid X) = \frac{0{,}01058}{0{,}01058 + 0{,}0274} = 0{,}27856 \quad (27.86\,\%)$$

$$P(\text{Jugar} = \text{No} \mid X) = \frac{0{,}0274}{0{,}01058 + 0{,}0274} = 0{,}72143 \quad (72.14\,\%)$$

Por lo tanto, en el 72.14 % de los casos en los que se presenta un clima lluvioso, temperatura media, humedad alta y viento, la respuesta será Jugar = No, y en el 27.86 % de los casos, la respuesta será Jugar = Sí.

4. Repetir el ejemplo anterior considerando en este caso el atributo temperatura con el valor igual a Media.

- **Número total de ejemplos:** $n = 14$
- **Número de valores que puede tomar Temperatura:** $n_t = 3$ (Alta, Media, Baja)
- **Número de ejemplos con Temperatura = Alta:** $n_{\text{alta}} = 4$
- **Número de ejemplos con Temperatura = Media:** $n_{\text{media}} = 6$
- **Número de ejemplos con Temperatura = Baja:** $n_{\text{baja}} = 4$

Además, consideramos $b = 20$ y $\epsilon = 0{,}01$.

Objetivo:

$$f = \frac{b}{100}(n - n_{\text{media}}) + n_{\text{media}} = 0{,}2 \times (14 - 6) + 6 = 0{,}2 \times 8 + 6 = 1{,}6 + 6 = 7{,}6$$

Iteración 0 (Inicio):

$$f_{\text{inferior }0} = 0 + \frac{\epsilon}{2} = 0 + 0{,}01/2 = 0{,}005$$

$$P(\text{alta} \mid \text{media}) = 1 - s\frac{n_t}{n} = 1 - 0{,}005\frac{3}{14} = 0{,}02369$$

$$P(\text{baja} \mid \text{media}) = P(\text{alta} \mid \text{media}) = 0{,}02369$$

$$P(\text{media} \mid \text{media}) = \frac{(s + 1 - s\frac{n_t}{n})}{n} = \frac{(0{,}005 + 1 - 0{,}005\frac{3}{14})}{14} = 0{,}02405$$

$$\begin{aligned}\text{sf}_{\text{rainferior }0} &= \frac{\sum P(x \mid \text{media}) n_x}{2} \Big/ \sum P(x \mid \text{media})^2 n_x \\ &= \frac{(0{,}02369 \times 4 + 0{,}02369 \times 4 + 0{,}02405 \times 6)}{2} \\ &\Big/ (0{,}02369^2 \times 4 + 0{,}02369^2 \times 4 + 0{,}02405^2 \times 6) = 13{,}9992\end{aligned}$$

$$f_{\text{superior }0} = 1 - \frac{\epsilon}{2} = 1 - \frac{0{,}01}{2} = 0{,}995$$

$$P(\text{alta} \mid \text{media}) = 1 - s\frac{n_t}{n} = 1 - 0{,}995\frac{3}{14} = 0{,}000119$$

$$P(\text{baja} \mid \text{media}) = P(\text{alta} \mid \text{media}) = 0{,}000119$$

$$P(\text{media} \mid \text{media}) = \frac{(s + 1 - s\frac{n_t}{n})}{n} = 0{,}07119$$

$$\begin{aligned}\text{sf}_{\text{superior }0} &= \frac{\sum P(x \mid \text{media}) n_x}{2} \Big/ \sum P(x \mid \text{media})^2 n_x \\ &= \frac{(0{,}000119 \times 4 + 0{,}000119 \times 4 + 0{,}07119 \times 6)}{2} \\ &\Big/ (0{,}000119^2 \times 4 + 0{,}000119^2 \times 4 + 0{,}07119^2 \times 6) = 6{,}0268\end{aligned}$$

$$f_{\text{valor }0} = 1 - \frac{b}{100} = 0{,}8$$

$$P(\text{alta} \mid \text{media}) = 1 - s\frac{n_t}{n} = 1 - 0{,}8\frac{3}{14} = 0{,}00476$$

$$P(\text{baja} \mid \text{media}) = P(\text{alta} \mid \text{media}) = 0{,}00476$$

$$P(\text{media} \mid \text{media}) = \frac{(s + 1 - s\frac{n_t}{n})}{n} = \frac{(0{,}8 + 1 - 0{,}8\frac{3}{14})}{14} = 0{,}06190$$

$$\begin{aligned}\text{sf}_{v0} &= \frac{\sum P(x \mid \text{media}) n_x}{2} \Big/ \sum P(x \mid \text{media})^2 n_x \\ &= \frac{(0{,}00476 \times 4 + 0{,}00476 \times 4 + 0{,}0619 \times 6)}{2} \\ &\Big/ (0{,}00476^2 \times 4 + 0{,}00476^2 \times 4 + 0{,}0619^2 \times 6) = 7{,}2364\end{aligned}$$

$$|\text{sf} - k| \not\leq \epsilon \quad \text{(Iteración 1)}$$

$$\text{sf}_{sup0} = 6{,}0268 < \text{sf}_{v0} = 7{,}2364 < k = 7{,}6 < \text{sf}_{inferior0} = 13{,}9992$$

$$\text{inferior } 1 = \text{inferior } 0 = 0{,}005; \quad \text{superior } 1 = \text{valor } 0 = 0{,}8$$

$$f_1 = \frac{\text{inferior } 1 + \text{superior } 1}{2} = \frac{0{,}005 + 0{,}8}{2} = 0{,}4025$$

$$P(\text{alta} \mid \text{media}) = 1 - 0{,}4025\frac{3}{14} = 0{,}0142$$

$$P(\text{baja} \mid \text{media}) = P(\text{alta} \mid \text{media}) = 0{,}0142$$

$$P(\text{media} \mid \text{media}) = \frac{(0{,}4025 + 1 - 0{,}4025\frac{3}{14})}{14} = 0{,}0430$$

$$\text{sf} = \frac{(0{,}0142 \times 4 + 0{,}0142 \times 4 + 0{,}0430 \times 6)}{2} \Big/ (0{,}0142^2 \times 4 + 0{,}0142^2 \times 4 + 0{,}0430^2 \times 6) = 10{,}8669$$

$$|\text{sf} - k| = |10{,}8669 - 7{,}6| = 3{,}2669 \not\leq \epsilon \quad \text{(Aún no converge)}$$

$$\text{sf}_{sup1} = 7{,}2364 < k = 7{,}6 < \text{sf}_{v1} = 10{,}8669 < \text{sf}_{inferior1} = 13{,}9992$$

$$\text{inferior } 1 = 0{,}005 < \text{valor } 1 = 0{,}4025 < \text{superior } 1 = 0{,}8$$

$$f_2 = \frac{\text{inferior } 2 + \text{superior } 2}{2} = \frac{0{,}4025 + 0{,}8}{2} = 0{,}60125$$

$$P(\text{alta} \mid \text{media}) = 1 - 0{,}60125\frac{3}{14} = 0{,}00949$$

$$P(\text{baja} \mid \text{media}) = P(\text{alta} \mid \text{media}) = 0{,}00949$$

$$P(\text{media} \mid \text{media}) = \frac{(0{,}60125 + 1 - 0{,}60125\frac{3}{14})}{14} = 0{,}05244$$

$$\text{sf} = \frac{(0{,}00949 \times 4 + 0{,}00949 \times 4 + 0{,}05244 \times 6)}{2} \Big/ (0{,}00949^2 \times 4 + 0{,}00949^2 \times 4 + 0{,}05244^2 \times 6) = 8{,}8580$$

$$|\text{sf} - k| = |8{,}8580 - 7{,}6| = 1{,}258 \not\leq \epsilon \quad \text{(Aún no converge)}$$

Iteración 3

$$\text{sf}_{sup2} = 7{,}2364 < k = 7{,}6 < \text{sf}_{v2} = 8{,}8580 < \text{sf}_{inf2} = 10{,}8669$$

$$\text{inferior } 2 = 0{,}4025 < \text{valor } 2 = 0{,}60125 < \text{superior } 2 = 0{,}8$$

$$\text{inferior } 3 = \text{valor } 2 = 0{,}60125; \quad \text{superior } 2 = \text{superior } 3 = 0{,}8$$

$$f_3 = \frac{\text{inferior } 3 + \text{superior } 3}{2} = \frac{0{,}60125 + 0{,}8}{2} = 0{,}70062$$

$$P(\text{alta} \mid \text{media}) = 1 - 0{,}70062\frac{3}{14} = 0{,}0071$$

$$P(\text{baja} \mid \text{media}) = P(\text{alta} \mid \text{media}) = 0{,}0071$$

$$P(\text{media} \mid \text{media}) = \frac{(0{,}70062 + 1 - 0{,}70062\frac{3}{14})}{14} = 0{,}0572$$

$$\text{sf} = \frac{(0{,}0071 \times 4 + 0{,}0071 \times 4 + 0{,}0572 \times 6)}{2} \Big/ (0{,}0071^2 \times 4 + 0{,}0071^2 \times 4 + 0{,}0572^2 \times 6) = 7{,}9863$$

$$|\text{sf} - k| = |7{,}9863 - 7{,}6| = 0{,}3863 \not\leq \epsilon \quad \text{(Aún no converge)}$$

Iteración 4

$$\text{sf}_{sup3} = 7{,}2364 < k = 7{,}6 < \text{sf}_{v3} = 7{,}9863 < \text{sf}_{inf3} = 8{,}8580$$

$$\text{inferior } 2 = 0{,}60125 < \text{valor } 3 = 0{,}70062 < \text{superior } 3 = 0{,}8$$

$$\text{inferior } 4 = \text{valor } 3 = 0{,}70062; \quad \text{superior } 3 = \text{superior } 4 = 0{,}8$$

$$f_4 = \frac{\text{inferior } 4 + \text{superior } 4}{2} = \frac{0{,}70062 + 0{,}8}{2} = 0{,}7503$$

$$P(\text{alta} \mid \text{media}) = 1 - 0{,}7503\frac{3}{14} = 0{,}0059$$

$$P(\text{baja} \mid \text{media}) = P(\text{alta} \mid \text{media}) = 0{,}0059$$

$$P(\text{media} \mid \text{media}) = \frac{(0{,}7503 + 1 - 0{,}7503\frac{3}{14})}{14} = 0{,}0595$$

$$\text{sf} = \frac{(0{,}0059 \times 4 + 0{,}0059 \times 4 + 0{,}0595 \times 6)}{2} \Big/ (0{,}0059^2 \times 4 + 0{,}0059^2 \times 4 + 0{,}0595^2 \times 6) = 7{,}5919$$

$$|\text{sf} - k| = |7{,}5919 - 7{,}6| = 0{,}0081 < \epsilon = 0{,}01 \quad \text{(Converge)}$$

Son necesarias 4 iteraciones, siendo el resultado del proceso el valor $s = 0{,}7503$.