Algoritmo de un árbol de decisión para la clasificación

UNMSM – FISI – EP DE INGENIERÍA DE SOFTWARE

Minería de Datos

PROFESOR: JUAN GAMARRA MORENO

Algoritmo (1/3)

El algoritmo de un árbol de decisión para clasificación sigue un proceso relativamente simple pero poderoso para dividir un conjunto de datos en clases:

1. **Inicio**: Comienza con todos los datos en un nodo raíz.
2. **Selección de la característica**: Selecciona una característica que se utilizará para dividir los datos en subconjuntos más pequeños. El objetivo es elegir la característica que mejor separe las clases.
3. **División de los datos**: Divide los datos en subconjuntos basados en los valores de la característica seleccionada en el paso anterior. Cada subconjunto se convierte en un nodo hijo del nodo actual.

Algoritmo (2/3)

4. **Repetición recursiva**: Repite los pasos 2 y 3 para cada nodo hijo creado en el paso anterior. Este proceso se repite recursivamente hasta que se cumple un criterio de parada, como una profundidad máxima predefinida, un número mínimo de muestras en un nodo o una pureza mínima en un nodo.
5. **Asignación de etiquetas**: Cuando se alcanza un nodo hoja (terminal), se asigna una etiqueta de clase basada en la mayoría de las muestras en ese nodo.
6. **Construcción completa del árbol**: Una vez que se ha completado la división y asignación de etiquetas para todos los nodos, se ha construido el árbol de decisión.

Algoritmo (3/3)

- Durante la construcción del árbol, el algoritmo busca la mejor manera de dividir los datos en cada nodo para minimizar la impureza en los nodos hijos. La medida de impureza más comúnmente utilizada es la entropía o la ganancia de información (criterio de división de información). Otras medidas de impureza incluyen el índice de Gini y la reducción de error de clasificación.
- Una vez que se ha construido el árbol, se puede utilizar para hacer predicciones sobre nuevas muestras asignándolas a una clase específica siguiendo las decisiones tomadas en cada nodo a lo largo del camino desde el nodo raíz hasta un nodo hoja.

Fórmulas cuando se usa Entropía como medida de impureza (1/3)

Entropía: La entropía se utiliza para medir la impureza de un conjunto de datos. Se calcula utilizando la fórmula:

$$Entropía = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

donde p_i es la probabilidad de la clase i en el conjunto de datos y n es el número total de clases. Esta fórmula se basa en la teoría de la información y cuánto menos ordenada esté una colección de datos (mayor incertidumbre o desorden), mayor será la entropía.

Fórmulas cuando se usa Entropía como medida de impureza (2/3)

Ganancia de Información (GI): La ganancia de información se utiliza para seleccionar la mejor característica para dividir los datos en cada nodo del árbol. Se calcula restando la entropía de un nodo de la entropía ponderada de los nodos hijos resultantes de dividir los datos en función de una característica específica.

$$GI = Entropía(NodoPadre) - \sum_{i=1}^{m} \left(\frac{N_i}{N} \times Entropía(NodoHijo_i) \right)$$

donde N es el número total de muestras en el nodo padre, N_i es el número de muestras en el nodo hijo i y m es el número de nodos hijos resultantes de dividir los datos en función de la característica.

Fórmulas cuando se usa Entropía como medida de impureza (3/3)

Estas fórmulas se utilizan durante la construcción del árbol para determinar qué características son las más informativas para dividir los datos y cómo se deben dividir para maximizar la homogeneidad de los nodos hijos.

Ejemplo con entropía

Se muestra un ejemplo simple de cómo se aplica el algoritmo de árbol de decisión utilizando la entropía como medida de impureza. Supongamos que tenemos un conjunto de datos con dos características (A y B) y queremos clasificar las muestras en dos clases (0 y 1).

Muestra	Característica A	Característica B	Clase
1	1	0	0
2	1	1	0
3	0	1	1
4	0	1	1
5	1	0	1
6	0	0	0

Vamos a calcular la entropía para el nodo raíz y luego ver cómo se divide el conjunto de datos en subconjuntos más puros.

Paso 1: Calcular la Entropía para el Nodo Raíz

Para calcular la entropía, utilizamos la fórmula:

$$Entropía = -p_1 \log_2(p_1) - p_2 \log_2(p_2)$$

donde p_1 y p_2 son las probabilidades de las clases 0 y 1, en el nodo raíz.

En nuestro conjunto de datos, hay 3 muestras de la clase 0 y 3 muestras de la clase 1, por lo que las probabilidades son $p_1 = \frac{3}{6}$ y $p_2 = \frac{3}{6}$

Entonces, la entropía para el nodo raíz es:

$$Entropía = -\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right) = 1$$

Paso 2: Dividir el conjunto de datos

Para cada característica, calcularemos la ganancia de información y seleccionaremos la característica que maximice la ganancia de información.

Para la Característica A:

- Si A es 0: tenemos 2 muestras de la clase 0 y 1 muestra de la clase 1.
- Si A es 1: tenemos 1 muestra de la clase 0 y 2 muestras de la clase 1.

Entropía si A es 0:

$$Entropía(A = 0) = -\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right) \approx 0.92$$

Entropía si A es 1:

$$Entropía(A = 1) = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right) \approx 0.92$$

Ganancia de Información (GI) de A:

$$GI(A) = Entropía - \left(p_1 \times Entropía(A = 0) + p_2 \times Entropía(A = 1)\right)$$

$$GI(A) = 1 - \left(\frac{3}{6} \times 0.92 + \frac{3}{6} \times 0.92\right) = 0.08$$

Para la Característica B:

- Si B es 0: tenemos 2 muestras de la clase 0 y 1 muestra de la clase 1.
- Si B es 1: tenemos 1 muestra de la clase 0 y 2 muestras de la clase 1.

Entropía si B es 0:

$$Entropía(B = 0) = -\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right) \approx 0.92$$

Entropía si A es 1:

$$Entropía(B = 1) = -\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right) \approx 0.92$$

Ganancia de Información (GI) de B:

$$GI(B) = Entropía - \left(p_1 \times Entropía(B = 0) + p_2 \times Entropía(B = 1)\right)$$

$$GI(A) = 1 - \left(\frac{3}{6} \times 0.92 + \frac{3}{6} \times 0.92\right) = 0.08$$

- Como la ganancia de información es igual para ambas características, podemos seleccionar cualquiera de ellas.
- Este proceso continúa recursivamente hasta alcanzar un criterio de parada (por ejemplo, una profundidad máxima del árbol o una pureza mínima en los nodos).