Decision Tree Example: Placement in Computer Science Engineering

Dataset

Consider the following dataset:

ProgrammingScore	DataStructuresScore	Placement
80	75	1
60	65	0
90	80	1
70	50	0

Solution

Step 1: Calculate the entropy of the target variable (Placement)

Entropy formula:

$$\text{Entropy}(S) = -p_1 \log_2(p_1) - p_0 \log_2(p_0)$$

where p_1 is the proportion of class 1 (placement = 1) and p_0 is the proportion of class 0 (placement = 0).

From the dataset:

$$p_1 = \frac{2}{4} = 0.5$$
$$p_0 = \frac{2}{4} = 0.5$$

$$\text{Entropy}(S) = -0.5 \log_2(0.5) - 0.5 \log_2(0.5) = -0.5 \times (-1) - 0.5 \times (-1) = 1$$

Step 2: Calculate information gain for each feature

Information gain formula for a feature A:

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \times \text{Entropy}(S_v)$$

where S is the dataset, A is a feature, v is a value of A, $|S_v|$ is the number of elements in S for which $A = v$, and $|S|$ is the total number of elements in S.

For ProgrammingScore: - Split the dataset into two subsets based on ProgrammingScore (≤ 70 and > 70): - Subset 1: $(60, 65, 0), (70, 50, 0)$ - Subset 2: $(80, 75, 1), (90, 80, 1)$ - Calculate the entropy for each subset: - Subset 1: $p_1 = \frac{0}{2} = 0$, $p_0 = \frac{2}{2} = 1$, $\text{Entropy}(S_1) = 0$ - Subset 2: $p_1 = \frac{2}{2} = 1$, $p_0 = \frac{0}{2} = 0$, $\text{Entropy}(S_2) = 0$ - Calculate the information gain:

$$\text{Gain}(S, \text{ProgrammingScore}) = 1 - \left(\frac{2}{4} \times 0 + \frac{2}{4} \times 0\right) = 1$$

For DataStructuresScore: - Split the dataset into two subsets based on DataStructuresScore (≤ 70 and > 70): - Subset 1: $(60, 65, 0), (70, 50, 0)$ - Subset 2: $(80, 75, 1), (90, 80, 1)$ - Calculate the entropy for each subset: - Subset 1: $p_1 = \frac{0}{2} = 0$, $p_0 = \frac{2}{2} = 1$, Entropy$(S_1) = 0$ - Subset 2: $p_1 = \frac{2}{2} = 1$, $p_0 = \frac{0}{2} = 0$, Entropy$(S_2) = 0$ - Calculate the information gain:

$$\text{Gain}(S, \text{DataStructuresScore}) = 1 - \left(\frac{2}{4} \times 0 + \frac{2}{4} \times 0\right) = 1$$

Both ProgrammingScore and DataStructuresScore have the same information gain, so we can choose either one as the root of the tree. Let's choose ProgrammingScore for simplicity.

Step 3: Select the feature with the highest information gain as the root

The root of the tree will be ProgrammingScore.

Step 4: Split the dataset based on the selected feature

We split the dataset based on ProgrammingScore ≤ 70 and ProgrammingScore > 70.

Step 5: Recursively apply steps 1-4 to each subset

We continue this process recursively until all data points are classified.

Decision Tree Algorithm: Step-by-Step Explanation

Dataset

Consider a dataset with features X_1 and X_2 and a binary target variable Y:

X_1	X_2	Y
3	5	0
2	4	1
4	6	0
5	2	1

Solution

Step 1: Calculate the entropy of the target variable (Y)

Entropy formula:

$$\text{Entropy}(S) = -p_1 \log_2(p_1) - p_0 \log_2(p_0)$$

where p_1 is the proportion of class 1 (Y = 1) and p_0 is the proportion of class 0 (Y = 0).

From the dataset:

$$p_1 = \frac{2}{4} = 0.5$$
$$p_0 = \frac{2}{4} = 0.5$$

$$\text{Entropy}(S) = -0.5\log_2(0.5) - 0.5\log_2(0.5) = -0.5 \times (-1) - 0.5 \times (-1) = 1$$

Step 2: Calculate information gain for each feature

Information gain formula for a feature X:

$$\text{Gain}(S, X) = \text{Entropy}(S) - \sum_{v \in \text{Values}(X)} \frac{|S_v|}{|S|} \times \text{Entropy}(S_v)$$

where S is the dataset, X is a feature, v is a value of X, $|S_v|$ is the number of elements in S for which $X = v$, and $|S|$ is the total number of elements in S.

For X_1: - Split the dataset into two subsets based on X_1 (≤ 3 and > 3): - Subset 1: $(3, 5, 0)$ - Subset 2: $(4, 6, 0), (2, 4, 1), (5, 2, 1)$ - Calculate the entropy for each subset: - Subset 1: $p_1 = \frac{0}{1} = 0$, $p_0 = \frac{1}{1} = 1$, $\text{Entropy}(S_1) = 0$ - Subset 2: $p_1 = \frac{2}{3}$, $p_0 = \frac{1}{3}$, $\text{Entropy}(S_2) = -\left(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}\right)$ - Calculate the information gain:

$$\text{Gain}(S, X_1) = 1 - \left(\frac{1}{4} \times 0 + \frac{3}{4} \times \text{Entropy}(S_2)\right)$$

For X_2: - Split the dataset into two subsets based on X_2 (≤ 4 and > 4): - Subset 1: $(2, 4, 1), (5, 2, 1)$ - Subset 2: $(3, 5, 0), (4, 6, 0)$ - Calculate the entropy for each subset: - Subset 1: $p_1 = \frac{2}{2} = 1$, $p_0 = \frac{0}{2} = 0$, $\text{Entropy}(S_1) = 0$ - Subset 2: $p_1 = \frac{0}{2} = 0$, $p_0 = \frac{2}{2} = 1$, $\text{Entropy}(S_2) = 0$ - Calculate the information gain:

$$\text{Gain}(S, X_2) = 1 - \left(\frac{2}{4} \times 0 + \frac{2}{4} \times 0\right)$$

Both X_1 and X_2 have the same information gain, so we can choose either one as the root of the tree. Let's choose X_1 for simplicity.

Step 3: Select the feature with the highest information gain as the root

The root of the tree will be X_1.

Step 4: Split the dataset based on the selected feature

We split the dataset based on $X_1 \leq 3$ and $X_1 > 3$.

Step 5: Recursively apply steps 1-4 to each subset

We continue this process recursively until all data points are classified.

Solution: Loan Approval

Step 1: Calculate the entropy of the target variable (LoanApproval)

Entropy formula:

$$\text{Entropy}(S) = -p_1 \log_2(p_1) - p_0 \log_2(p_0)$$

where p_1 is the proportion of class 1 (LoanApproval = 1) and p_0 is the proportion of class 0 (LoanApproval = 0).

From the dataset:

$$p_1 = \frac{2}{4} = 0.5$$
$$p_0 = \frac{2}{4} = 0.5$$

$$\text{Entropy}(S) = -0.5 \log_2(0.5) - 0.5 \log_2(0.5) = -0.5 \times (-1) - 0.5 \times (-1) = 1$$

Step 2: Calculate information gain for each feature

Information gain formula for a feature X:

$$\text{Gain}(S, X) = \text{Entropy}(S) - \sum_{v \in \text{Values}(X)} \frac{|S_v|}{|S|} \times \text{Entropy}(S_v)$$

where S is the dataset, X is a feature, v is a value of X, $|S_v|$ is the number of elements in S for which $X = v$, and $|S|$ is the total number of elements in S.

For CreditScore: - Split the dataset into two subsets based on CreditScore ($\leq$ 700 and $>$ 700): - Subset 1: $(600, 30000, 0), (650, 40000, 0)$ - Subset 2: $(700, 50000, 1), (750, 80000, 1)$ - Calculate the entropy for each subset: - Subset 1: $p_1 = \frac{0}{2} = 0$, $p_0 = \frac{2}{2} = 1$, $\text{Entropy}(S_1) = 0$ - Subset 2: $p_1 = \frac{2}{2} = 1$, $p_0 = \frac{0}{2} = 0$, $\text{Entropy}(S_2) = 0$ - Calculate the information gain:

$$\text{Gain}(S, \text{CreditScore}) = 1 - \left(\frac{2}{4} \times 0 + \frac{2}{4} \times 0 \right) = 1$$

For Income: - Split the dataset into two subsets based on Income (≤ 50000 and > 50000): - Subset 1: $(600, 30000, 0), (650, 40000, 0)$ - Subset 2: $(700, 50000, 1), (750, 80000, 1)$ - Calculate the entropy for each subset: - Subset 1: $p_1 = \frac{0}{2} = 0$, $p_0 = \frac{2}{2} = 1$, $\text{Entropy}(S_1) = 0$ - Subset 2: $p_1 = \frac{2}{2} = 1$, $p_0 = \frac{0}{2} = 0$, $\text{Entropy}(S_2) = 0$ - Calculate the information gain:

$$\text{Gain}(S, \text{Income}) = 1 - \left(\frac{2}{4} \times 0 + \frac{2}{4} \times 0\right) = 1$$

Both CreditScore and Income have the same information gain, so we can choose either one as the root of the tree. Let's choose CreditScore for simplicity.

Step 3: Select the feature with the highest information gain as the root

The root of the tree will be CreditScore.

Step 4: Split the dataset based on the selected feature

We split the dataset based on CreditScore ≤ 700 and CreditScore > 700.

Step 5: Recursively apply steps 1-4 to each subset

We continue this process recursively until all data points are classified.