Tech Bytes

ML Refresher - Decision Trees

by Pranali Bose

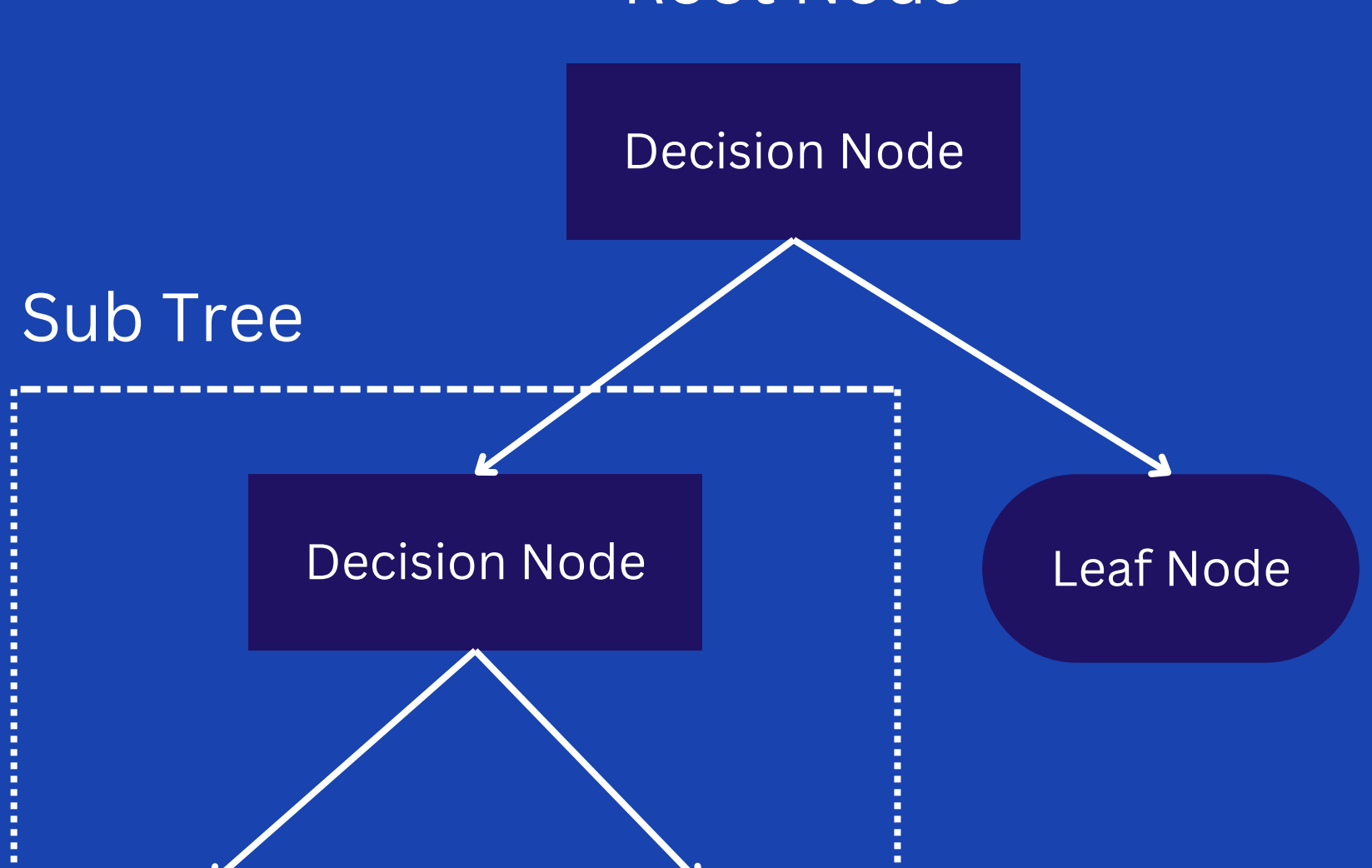
What is it?

Versatile ML algorithm that can perform both classification and regression tasks.

It is a powerful algorithm capable of fitting complex datasets.

It is also a fundamental component of Random Forest, one of the most powerful ML algorithms available today.

Structure


Root Node
Decision Node
Sub Tree
Decision Node
Leaf Node
Leaf Node
Leaf Node

Terminologies

- **Split**: Dividing a node into two or more sub-nodes based on a feature.
- **Impurity**: Measure of how mixed the classes are at a node.
- **Pruning**: Removing branches from the tree to prevent overfitting and improve generalization.
- **Depth**: Length of the longest path from the root node to a leaf node

Intuition Behind

- **Feature Splitting**: The algorithm selects the best feature to separate classes using metrics like Gini impurity.

- **Recursive Partitioning**: Splitting continues until a stopping criterion, such as maximum depth, is reached.

Feature Splitting?

One of the ways...

- **Step 1**: Calculate Gini Impurity for the nodes
- **Step 2**: Evaluate possible splits
- **Step 3**: Compare the Gini values and pick the least

Formulae

Gini = $\Sigma(p_i^2)$

Gini Impurity = 1 - Gini

Alternatively,

Entropy=- $\Sigma(p_i \log_2(p_i))$

Information Gain = 1 - Entropy

where p_i is the probability of class i

Example

Type	Color	Pet
Red	Cat	Yes
Red	Dog	No
Blue	Cat	Yes
Blue	Dog	Yes

Step 1: Calculate Gini Impurity for the Parent Node

- Total Instances: 4 (2 Yes, 2 No)

- $\textbf{Gini}_{\textbf{Parent}} = 1 - (p_{Yes}^2 + p_{No}^2) = 0.5$

Step 2: Evaluate Possible Splits

Split by Color

Red - 2 (1 Yes, 1 No)

$Gini_{Red} = 0.5$

Blue - 2 (1 Yes, 1 No)

$Gini_{Blue} = 0.5$

$$\mathbf{Gini_{Color}} = 2/4 * Gini_{Red} + 2/4 * Gini_{Blue} = \mathbf{0.5}$$

Split by Type

Cat - 2 (2 Yes)

$Gini_{Cat} = 0$

Dog - 2 (1 Yes, 1 No)

$Gini_{Dog} = 0.5$

$$\mathbf{Gini_{Type}} = 2/4 * Gini_{Cat} + 2/4 * Gini_{Dog} = \mathbf{0.25}$$

Step 3: Compare Splits and Choose

Split	Gini Impurity
Color	0.5
Type	0.25
No split	0.5

Conclusion:
The best split is by Type because it results in the lowest Gini impurity.

Hyperparameters

- **max_depth**: Maximum depth of the tree. Limiting the depth can help prevent overfitting
- **min_samples_split**: minimum number of samples required to split an internal node. Increasing this value can also reduce overfitting
- **criterion**: The function to measure the quality of a split
- **class_weight**: Weights associated with classes in the classification problem. It can help with imbalanced datasets

Pros and Cons

- No need for feature scaling
- Can handle both numerical and categorical data
- Flexible in modeling complex relationships

- Prone to overfitting if not properly managed
- Sensitive to small changes in the data

Ponder Upon

- What criteria can be used to determine the best split in a decision tree? Why does entropy have a negative sign?

- Describe the process of pruning in decision trees. Why is it important?

- What are the limitations of decision trees compared to ensemble methods like Random Forests?

- How do decision trees handle missing values during training?

Tech Bytes

Find this useful?

Let me know in the comments which topic would you like to see next :)

Follow for more...