Decision Trees

Ishaan Jain

August 2023

1 Introduction:

A decision tree is a popular machine learning algorithm used for both classification and regression tasks. It is a simple yet powerful model that breaks down complex decision-making processes into a series of simple decisions. They can be thought as nested if-else conditions.

2 Algorithm Description:

A decision tree algorithm constructs a tree-like structure where each internal node represents a decision based on a specific attribute, and each leaf node represents a class label or a numerical value. The algorithm recursively splits the dataset based on the chosen attributes, aiming to maximize information gain (in classification) or minimize variance (in regression) at each step. This process continues until a stopping criterion is met, such as reaching a maximum depth or having a minimum number of samples in a node.

Entropy - It's the measure of randomness. It's defined as

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

Based on Entropy, the decision tree does classification.

3 Advantages:

Interpretability: Decision trees are easy to understand and visualize, making them suitable for explaining model decisions to non-technical stakeholders.

Handling Non-linearity: Decision trees can capture non-linear relationships in the data, making them versatile for a wide range of problem types.

Mixed Data Types: They can handle both categorical and numerical data, eliminating the need for extensive data preprocessing.

Feature Importance: Decision trees provide a feature importance score, indicating which features are most relevant in making predictions.

No Assumptions: Decision trees do not assume any specific distribution of data, making them suitable for various types of datasets.

4 Limitations:

Overfitting: Decision trees are prone to overfitting, especially when the tree becomes too deep and complex. Regularization techniques like pruning are used to mitigate this issue.

Instability: Small variations in the data can lead to significantly different tree structures, making the model somewhat unstable

Bias towards Dominant Classes: In classification tasks, decision trees tend to favor majority classes, leading to imbalanced predictions.

Inability to Capture Certain Relationships: Decision trees might struggle to capture complex relationships that require multiple sequential decisions.