Decision Trees Theory

1 Introduction

Decision Trees are a popular machine learning algorithm used for both classification and regression tasks. They model decisions and their possible consequences as a tree-like structure. Each internal node of the tree represents a decision based on a feature, each branch represents the outcome of that decision, and each leaf node represents a class label or continuous value.

2 Tree Construction

A Decision Tree is constructed by recursively splitting the data into subsets based on feature values. The objective is to create subsets that are as homogeneous as possible with respect to the target variable.

2.1 Splitting Criterion

The quality of a split is typically measured using impurity measures. Two common impurity measures are:

- **Gini Impurity**:

  $$Gini = 1 - \sum_{i=1}^{C} p_i^2 \tag{1}$$

  where p_i is the probability of class i in the node, and C is the number of classes.

- **Entropy**:

  $$H = -\sum_{i=1}^{C} p_i \log_2(p_i) \tag{2}$$

  where p_i is the probability of class i in the node.

2.2 Recursive Splitting

The tree is built recursively by choosing the best feature and threshold that minimizes the impurity measure:

$$\text{Impurity}_{\text{split}} = \frac{n_{\text{left}}}{n_{\text{total}}}\text{Impurity}_{\text{left}} + \frac{n_{\text{right}}}{n_{\text{total}}}\text{Impurity}_{\text{right}} \tag{3}$$

where: - n_{left} and n_{right} are the number of samples in the left and right splits, respectively. - $\text{Impurity}_{\text{left}}$ and $\text{Impurity}_{\text{right}}$ are the impurity measures of the left and right splits.

3 Stopping Criteria

The recursion terminates when one of the following conditions is met:

- The node contains samples of only one class.
- The maximum tree depth is reached.
- The node contains fewer than a minimum number of samples.
- No further improvement in impurity can be achieved.

4 Pruning

Pruning is a technique used to reduce the complexity of the tree and improve generalization. It involves removing nodes or branches that have little importance. Two common pruning methods are:

- **Cost Complexity Pruning**: Also known as weakest link pruning, it removes nodes that have the least impact on the overall performance of the tree.
- **Reduced Error Pruning**: It removes nodes that do not improve the performance on a validation set.

5 Decision Tree Algorithms

Several algorithms are used to construct Decision Trees, including:

- **ID3** (Iterative Dichotomiser 3): Uses entropy as the impurity measure.
- **C4.5**: An extension of ID3 that handles continuous features and missing values.
- **CART** (Classification and Regression Trees): Uses Gini impurity for classification and mean squared error for regression.