Decision Trees: Implementing Classification and Regression from Scratch

Tejinder Singh Hunjan, Mohit Kamkhalia, Sashank Mishra

Abstract

Decision Trees are powerful and interpretable models used for both classification and regression tasks. This report explains how Decision Trees work internally, covering splitting criteria, impurity measures, and recursive construction. We demonstrate how both classification and regression trees (CART) are implemented from scratch without external libraries.

1. Introduction

A Decision Tree recursively partitions the input space based on feature values to make predictions. Internal nodes represent decisions, and leaf nodes represent outputs: class labels for classification or real values for regression. The tree is built top-down by selecting optimal feature thresholds.

2. Splitting Criteria for Classification

At each node, we evaluate all possible splits and choose the one that results in the largest reduction in impurity.

Gini Impurity

Gini measures the probability of misclassification:

$$Gini = 1 - \sum_{k=1}^{K} p_k^2$$

where p_k is the proportion of class k in the node.

Entropy and Information Gain

Entropy captures class unpredictability:

$$Entropy = -\sum_{k=1}^{K} p_k \log_2(p_k)$$

The best split maximizes the Information Gain:

$$IG = Entropy_{parent} - \left(\frac{n_{left}}{n} Entropy_{left} + \frac{n_{right}}{n} Entropy_{right}\right)$$

3. Splitting Criteria for Regression

Regression trees use variance reduction as the criterion, typically minimizing Mean Squared Error (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Each split aims to reduce the combined MSE of the resulting subsets.

4. Tree Building Algorithm

1. Begin at the root node with the full dataset.
2. For every feature and potential threshold:
    - Split the dataset into two parts.
    - Calculate impurity or MSE for the child nodes.
3. Choose the split that provides the best gain (lowest impurity or error).
4. Recursively apply the same process to left and right subsets.
5. Stop when a stopping criterion is met:
    - Maximum tree depth
    - Minimum number of samples in a node
    - No further improvement in impurity/error

5. Predictions

Classification

To classify a sample, traverse the tree using its feature values until a leaf is reached. The predicted class is the majority class in that leaf.

Regression

For regression, the predicted value is the mean of the target values in the reached leaf node.

6. Conclusion

Decision Trees provide an intuitive way to model both classification and regression problems. Understanding the splitting logic, impurity metrics, and recursive tree construction enables better insight into model behavior. Though they can overfit without pruning, trees are foundational to ensemble methods like Random Forests and Gradient Boosting.