Study Guide: Decision Trees

Student Name

October 9, 2024

1 Introduction

Decision trees are a widely used method for both classification and regression tasks. The structure of a decision tree is a flowchart-like model where each internal node represents a decision based on a feature, each branch corresponds to an outcome of that decision, and each leaf node represents a prediction (class or regression value). Decision trees are interpretable, easy to implement, and form the foundation of more complex models like random forests.

2 Sample Complexity of Decision Trees (Section 18.1)

The sample complexity of decision trees refers to the number of training samples required to achieve good generalization. A key factor in determining the complexity is the depth of the tree.

2.1 VC Dimension of Decision Trees

The VC dimension of a decision tree with k internal nodes is $O(k)$. This suggests that the sample complexity grows linearly with the number of nodes. Shallow trees, which have fewer internal nodes, have lower complexity but might underfit, while deep trees might overfit if not enough data is provided.

2.2 Tree Pruning

Pruning is a technique used to reduce the complexity of a decision tree by removing branches that have little impact on the prediction accuracy. This helps control overfitting by reducing the tree's depth, thus lowering the VC dimension.

3 Decision Tree Algorithms (Section 18.2)

Several algorithms are available for building decision trees. Most decision tree algorithms work by recursively splitting the data based on features that maximize a splitting criterion.

3.1 Splitting Criteria

Common splitting criteria include:

- **Gini Index:** Measures the impurity of a node. A split is chosen to minimize the Gini impurity in the resulting child nodes.
- **Entropy (Information Gain):** Based on the concept of entropy from information theory, information gain measures how much information is gained by making a split.
- **Mean Squared Error:** Used in regression trees, this criterion minimizes the variance of the target values within each split.

3.2 CART Algorithm

The **CART (Classification and Regression Trees)** algorithm is one of the most commonly used algorithms for building decision trees. It constructs binary trees by splitting the data based on the criterion that maximizes either Gini index or mean squared error.

3.3 ID3 Algorithm

The **ID3 (Iterative Dichotomiser 3)** algorithm uses the information gain criterion to select the best feature at each step. ID3 is mainly used for classification tasks.

4 Random Forests (Section 18.3)

A **random forest** is an ensemble method that builds multiple decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees. Each tree is trained on a random subset of the data and the features, which reduces the variance of the model and improves generalization.

4.1 Bagging

Bagging (Bootstrap Aggregating) is the technique used to generate the random subsets of the training data for each tree in the random forest. By averaging over multiple trees, random forests reduce the risk of overfitting compared to individual decision trees.

5 Summary (Section 18.4)

- Decision trees are powerful and interpretable models used for both classification and regression.
- The sample complexity of a decision tree grows with its depth, and techniques like pruning help to manage this complexity.
- Common algorithms for building decision trees include CART and ID3, which use criteria like Gini index and information gain to select the best splits.
- Random forests are ensembles of decision trees that improve accuracy by reducing variance and mitigating overfitting.