DOCUMENTATION

Imports and Node Class

- We import numpy for numerical operations and Counter from collections for counting elements.
- The Node class represents a node in the decision tree:
    - feature: The index of the feature used for splitting at this node.
    - threshold: The threshold value for the split.
    - left and right: References to the left and right child nodes.
    - current: For leaf nodes, this stores the predicted class.

DecisionTree Class

- The DecisionTree class initializes with two hyperparameters:
    - maxDepth: Maximum depth of the tree.
    - minSplit: Minimum number of samples required to split an internal node.
- self.root will store the root node of the tree.

Fit Method

- fit method starts the tree-building process by calling growTree.

Grow Tree Method

- This method recursively builds the tree:
    - It first checks stopping criteria (max depth, pure node, or min samples).
    - If stopping criteria are met, it creates a leaf node with the most common label.
    - Otherwise, it selects the best feature and threshold for splitting.
    - It then splits the data and recursively builds left and right subtrees.

Best Split Method

- This method finds the best feature and threshold for splitting:
    - It iterates through features and potential thresholds.
    - For each combination, it calculates the information gain.
    - It keeps track of the split with the highest information gain.

Information Gain Method

- This method calculates the information gain for a potential split:
    - It computes the entropy of the parent node.
    - It splits the data based on the threshold.

    - It calculates the weighted average entropy of the child nodes.
    - The information gain is the difference between parent and child entropies.

Helper Methods

- split: Splits data based on a threshold.
- entropy: Calculates the entropy of a set of labels.
- printTree: Prints the decision tree.

Prediction Methods

- predict: Predicts the class for a set of samples.
- traverseTree: Recursively traverses the tree for a single sample to make a prediction.

Link: Video Demonstration