Decision Tree in Machine Learning

March 11, 2025

Introduction to Decision Trees

- **What is a Decision Tree?**
  - A supervised machine learning algorithm.
  - Used for both **classification** and **regression** tasks.
  - Represents decisions and their possible consequences in a tree-like structure.
- **Key Components:**
  - **Nodes:** Represent decisions or tests on attributes.
  - **Edges/Branches:** Represent outcomes of decisions.
  - **Leaf Nodes:** Represent final decisions or predictions.

Why Use Decision Trees?

- **Advantages:**
  - Easy to understand and interpret (visual representation).
  - Can handle both numerical and categorical data.
  - Requires minimal data preprocessing.
  - Robust to outliers and missing values.
- **Disadvantages:**
  - Prone to overfitting, especially with complex trees.
  - Can be unstable (small changes in data can lead to different trees).
  - May create biased trees if some classes dominate.

Key Terminology

- **Root Node:** The topmost node representing the entire dataset.
- **Splitting:** Dividing a node into sub-nodes based on a condition.
- **Decision Node:** A node that splits into further sub-nodes.
- **Leaf/Terminal Node:** A node with no further splits (final output).
- **Pruning:** Removing sub-nodes to reduce overfitting.
- **Branch/Sub-Tree:** A subsection of the decision tree.

How Decision Trees Work

- **Step 1:** Start with the entire dataset at the root node.
- **Step 2:** Select the best attribute to split the data (using metrics like **Gini Index**, **Entropy**, or **Information Gain**).
- **Step 3:** Split the dataset into subsets based on the selected attribute.
- **Step 4:** Repeat the process recursively for each subset.
- **Step 5:** Stop when a stopping criterion is met (e.g., maximum depth, pure nodes).

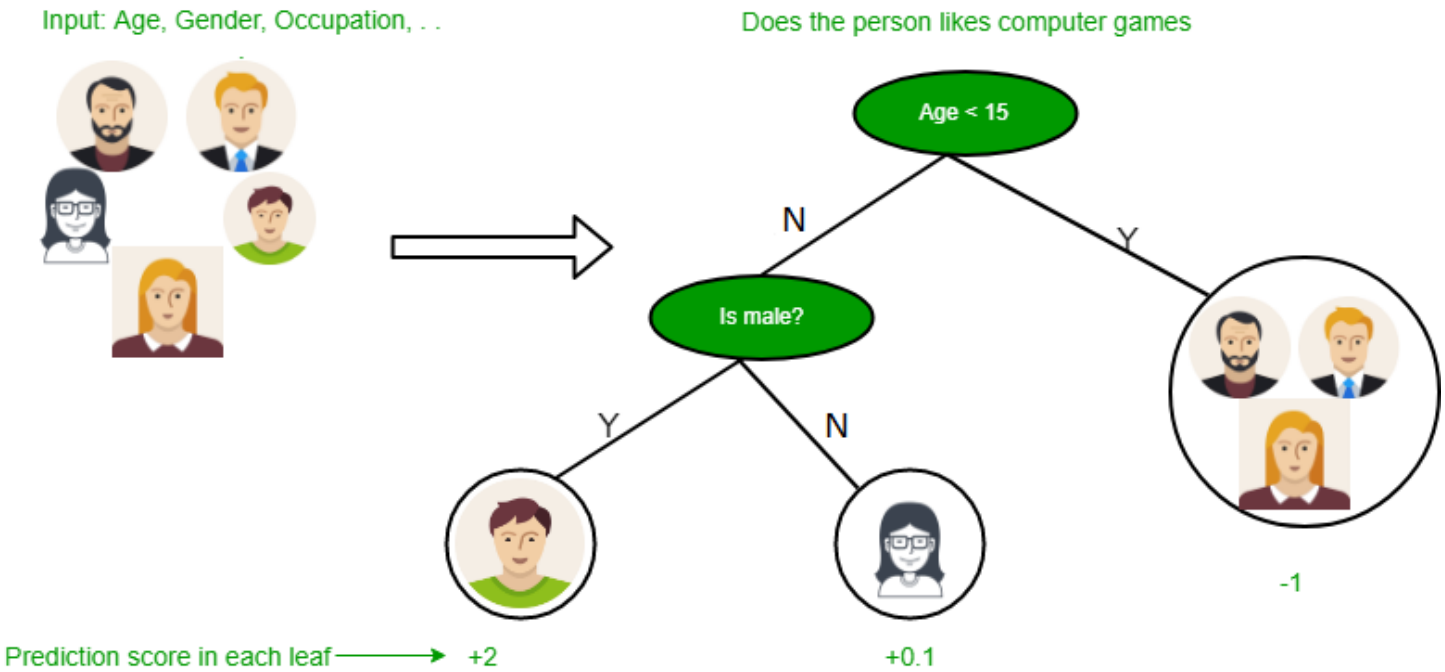
Splitting Criteria

- **Gini Index:**
  - Measures impurity in a node.
  - A Gini Index of 0 means the node is pure.
- **Entropy:**
  - Measures randomness or uncertainty.
  - Higher entropy means more uncertainty.
- **Information Gain:**
  - Measures the reduction in entropy after a split.
  - Higher information gain means a better split.

Example of a Decision Tree

- **Problem:** Classify whether a person will play tennis based on weather conditions.
- **Attributes:** Outlook, Temperature, Humidity, Wind.
- **Target Variable:** Play Tennis (Yes/No).
- **Tree Structure:**
  - Root Node: Outlook (Sunny, Overcast, Rain).
  - Sub-nodes: Temperature, Humidity, Wind.
  - Leaf Nodes: Play Tennis (Yes/No).

Visual Representation



Overfitting and Pruning

- **Overfitting:**
  - When the tree becomes too complex and captures noise in the data.
- **Pruning:**
  - Technique to reduce the size of the tree by removing unnecessary branches.
  - Improves generalization on unseen data.

Applications of Decision Trees

- **Classification:**
  - Spam detection, disease diagnosis, customer segmentation.
- **Regression:**
  - Predicting house prices, stock prices, or sales.
- **Other Uses:**
  - Feature selection, data exploration.

Conclusion

- Decision trees are a powerful and interpretable tool for machine learning.
- They are easy to implement and visualize but require careful handling to avoid overfitting.
- Pruning and parameter tuning can improve their performance.

Questions and Answers