# Tutorial: Decision Trees

*Cheat Sheet and Practice Problems*

ES335 - Machine Learning
IIT Gandhinagar

July 23, 2025

## 1 Summary from Slides

### 1.1 Key Concepts

**What are Decision Trees?**

- Tree-like model for making decisions through series of questions
- Internal nodes: feature tests, Leaves: class labels (classification) or values (regression)
- Interpretable - can trace prediction path

**Types of Decision Trees:**

- **Discrete Input, Discrete Output**: Classification with categorical features
- **Discrete Input, Real Output**: Regression with categorical features
- **Real Input, Discrete Output**: Classification with continuous features
- **Real Input, Real Output**: Regression with continuous features

### 1.2 Information Theory

**Entropy**: Measure of impurity/uncertainty

$$H(S) = -\sum_{i=1}^{c} p_i \log_2(p_i)$$

**Information Gain**: Reduction in entropy after split

$$\text{Gain}(S, A) = H(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} H(S_v)$$

**Weighted Entropy**: For weighted samples

$$H_w(S) = -\sum_{i=1}^{c} p_{w,i} \log_2(p_{w,i})$$

where $p_{w,i} = \frac{\sum_{j:y_j=i} w_j}{\sum_j w_j}$

## 1.3 Tree Construction Algorithm

1. Calculate entropy of current node
2. For each feature, calculate information gain
3. Choose feature with maximum information gain
4. Split on that feature
5. Recursively apply to child nodes
6. Stop when: pure node, max depth reached, min samples reached

**For Regression Trees**: Use variance reduction instead of information gain

$$\text{VarReduction}(S, A) = \text{Var}(S) - \sum_v \frac{|S_v|}{|S|}\text{Var}(S_v)$$

## 1.4 Key Properties

**Advantages**:

- Highly interpretable
- No assumptions about data distribution
- Handles both numerical and categorical features
- Automatic feature selection
- Non-parametric

**Disadvantages**:

- Prone to overfitting
- Unstable (small data changes can drastically change tree)
- Biased toward features with more levels
- Greedy algorithm (may not find globally optimal tree)

## 2 Practice Problems

**Problem : Information Gain Calculation**

Given the tennis dataset from slides:

| Outlook | Temperature | Humidity | Wind | Play? |
|---|---|---|---|---|
| Sunny | Hot | High | Weak | No |
| Sunny | Hot | High | Strong | No |
| Overcast | Hot | High | Weak | Yes |
| Rain | Mild | High | Weak | Yes |
| Rain | Cool | Normal | Weak | Yes |
| Rain | Cool | Normal | Strong | No |
| Overcast | Cool | Normal | Strong | Yes |
| Sunny | Mild | High | Weak | No |
| Sunny | Cool | Normal | Weak | Yes |
| Rain | Mild | Normal | Weak | Yes |
| Sunny | Mild | Normal | Strong | Yes |
| Overcast | Mild | High | Strong | Yes |
| Overcast | Hot | Normal | Weak | Yes |
| Rain | Mild | High | Strong | No |

Calculate the information gain for splitting on the "Wind" attribute.

**Problem : Tree Construction**

Build a complete decision tree for this simple dataset:

| X1 | X2 | Y |
|---|---|---|
| 0 | 0 | A |
| 0 | 1 | B |
| 1 | 0 | B |
| 1 | 1 | A |

Show all entropy calculations and explain your splitting decisions.

**Problem : Regression Tree**

For this regression dataset:

| Size | Price |
|---|---|
| 1000 | 100 |
| 1200 | 120 |
| 1500 | 180 |
| 1800 | 200 |
| 2000 | 250 |

Find the best split point using variance reduction. Calculate the variance before and after the split.

**Problem : Weighted Entropy**

Given a weighted dataset where samples have weights [2, 1, 3, 1] and classes [A, B, A, B]:
a) Calculate the weighted entropy b) Compare with unweighted entropy c) Explain when weighted entropy is useful

### Problem : Overfitting Analysis

You have a decision tree with the following training and validation accuracies at different depths:

| Depth | Training Acc | Validation Acc |
|---|---|---|
| 1 | 0.70 | 0.68 |
| 2 | 0.80 | 0.76 |
| 3 | 0.90 | 0.82 |
| 4 | 0.95 | 0.79 |
| 5 | 0.98 | 0.75 |
| 6 | 1.00 | 0.70 |

a) At what depth does overfitting begin? b) What depth would you choose for deployment? c) Suggest three techniques to reduce overfitting.

### Problem : Feature Selection Bias

Explain why decision trees are biased toward features with more distinct values. Give a concrete example with: - Feature A: 2 possible values - Feature B: 10 possible values - Same information content
How would you address this bias?

### Problem : Continuous Feature Splits

For a continuous feature "Age" with values [25, 30, 35, 40, 45, 50] and binary target [0, 0, 1, 1, 0, 1]: a) List all possible split points b) Calculate information gain for split at Age $\leq 37.5$ c) Find the optimal split point

### Problem : Pruning Strategies

Compare pre-pruning vs post-pruning:
a) Define minimum samples per leaf = 3. How does this affect tree growth? b) For post-pruning, explain how you would use a validation set c) What is the trade-off between model complexity and generalization?

### Problem : Missing Values

How would you handle missing values in decision trees during: a) Training phase b) Prediction phase Propose two different strategies for each phase and discuss their pros/cons.

**Problem : Implementation Challenge**

Implement the decision tree algorithm pseudocode:

```
function BuildTree(data, features, target):
    if stopping_condition(data):
        return create_leaf(data)

    best_feature = find_best_split(data, features)
    tree = create_node(best_feature)

    for each value v in best_feature:
        subset = filter_data(data, best_feature == v)
        subtree = BuildTree(subset, features, target)
        add_branch(tree, v, subtree)

    return tree
```

Define each helper function and specify the stopping conditions.

**Problem : Ensemble Motivation**

Given the instability of decision trees:
a) Explain why small changes in data can lead to very different trees b) How does this motivate ensemble methods like Random Forests? c) Describe the bias-variance trade-off for decision trees d) How do ensembles help with this trade-off?

**Problem : Real-World Application**

Design a decision tree for a loan approval system with features: - Income (continuous) - Credit Score (continuous) - Employment Type (categorical: Full-time, Part-time, Self-employed, Unemployed) - Previous Defaults (binary)
a) What would be appropriate stopping criteria? b) How would you handle class imbalance (more approvals than rejections)? c) What interpretability constraints might be required for regulatory compliance? d) How would you validate the model's fairness across different demographic groups?