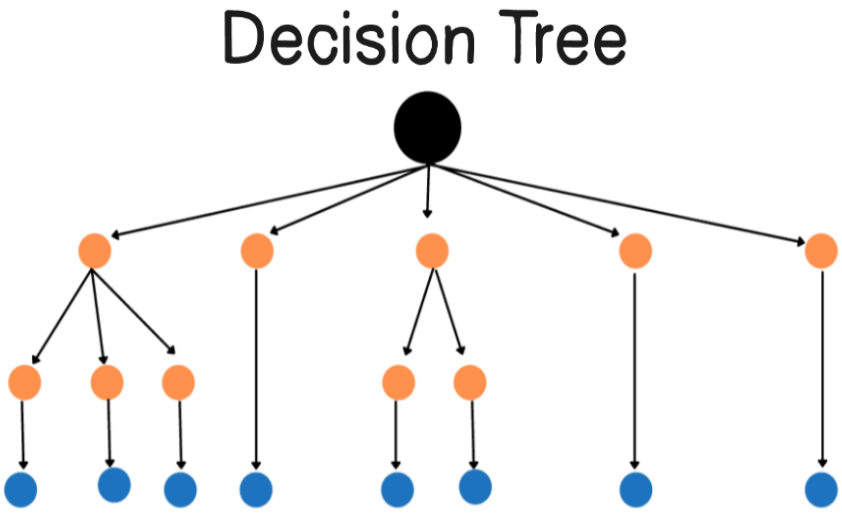
DECISION TREE

Introduction

A Decision Tree is a supervised learning algorithm used for classification and regression. It recursively partitions the feature space into regions by selecting features and split points that maximize class purity (classification) or reduce variance (regression). The model is interpretable and rule-based.



1. Problem Formulation

Input Space $\mathcal{X} \subseteq \mathbb{R}^p$

Output Space

- Classification: $\mathcal{Y} = \{1,2,\dots,K\}$
- Regression: $\mathcal{Y} \subseteq \mathbb{R}$

Data Distribution

$(x_i, y_i) \sim P_{X,Y}$ i.i.d samples.

Learning Objective

Minimize expected prediction error by partitioning feature space:

$$\min_{T} \mathbb{E}[\ell(y, f_T(x))]$$

where Trepresents the tree structure.

2. Model Specification

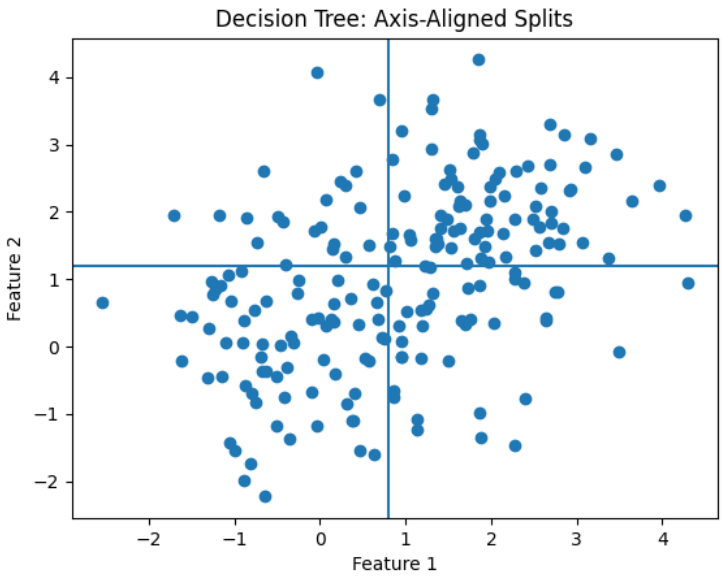
A decision tree splits data recursively.

At each node:

- Select feature j
- Select threshold t
- Split data into: $x_j \leq t$and$x_j > t$

Prediction:

- Classification → majority class
- Regression → mean value



3. Loss Function

Classification

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

Entropy $H = -\sum_{k=1}^{K} p_k \log\ p_k$

Choose split that maximizes impurity reduction.

Regression

Variance reduction: $\text{MSE} = \frac{1}{n}\sum(y_i - \bar{y})^2$

Choose split minimizing squared error.

4. Objective Function

At each split, maximize:

$$\Delta I = I(parent) - \left(\frac{n_L}{n} I(L) + \frac{n_R}{n} I(R)\right)$$

Where:

- I= impurity measure
- L, R= left and right nodes

5. Optimization Method

Greedy algorithm:

1. Try all features
2. Try all thresholds
3. Choose best impurity reduction
4. Repeat recursively

No global optimization. Purely greedy.

Complexity

Training: $O(pn\log\ n)$

Prediction: $O(\text{tree depth})$

6. Statistical Interpretation

Decision trees are nonparametric models.

They approximate: $E[y \mid x]$ by partitioning space into regions.

No distributional assumptions.

7. Regularization & Generalization

Trees overfit easily.

Control methods:

- Maximum depth
- Minimum samples per leaf
- Pruning
- Maximum number of leaves

Bias–Variance:

- Deep tree → low bias, high variance
- Shallow tree → high bias, low variance

8. Theoretical Properties

- Highly flexible
- Nonlinear decision boundaries
- Piecewise constant approximation
- High variance model

Not convex optimization.

9. Computational Complexity

Training: $O(pn\log\ n)$

Inference: $O(\text{depth})$

Memory: $O(\text{number of nodes})$

10. Limitations

1. Overfitting
2. High variance
3. Unstable to small data changes
4. Axis-aligned splits only
5. Poor extrapolation in regression