Machine Learning

Dr.Hajialiasgari

Tehran University
Of
Medical Science

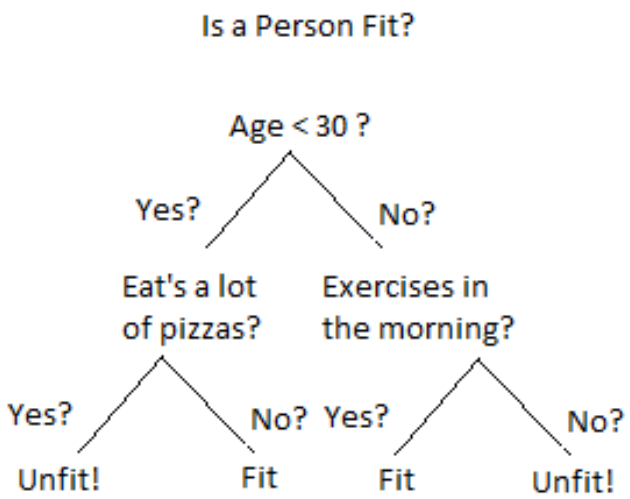
February 4, 2025

TEHRAN UNIVERSITY
OF
MEDICAL SCIENCES

Overview of Decision Trees

- Decision Trees are used for both classification and regression.
- They split data into branches based on feature values.
- The tree consists of **nodes** (decisions) and **leaves** (predictions).

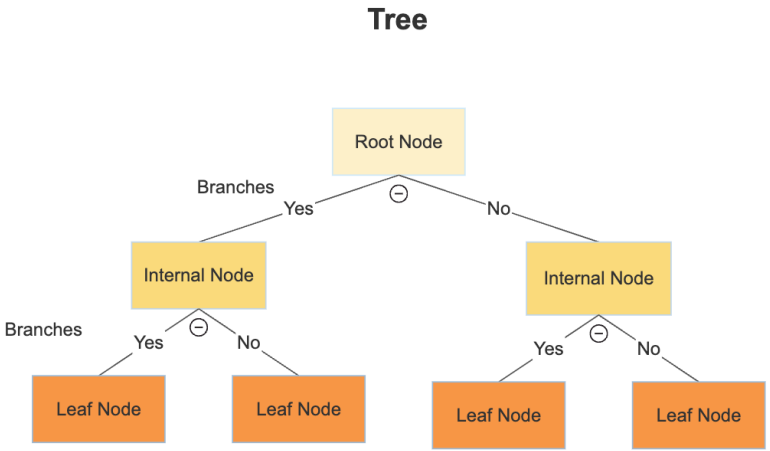
Decision Tree (Cont.)



How Decision Trees Work

- Starts with the entire dataset at the root.
- Splits the dataset based on a selected feature using a splitting criterion (e.g., Gini Impurity, Entropy, or Mean Squared Error).
- Repeats the process recursively until stopping criteria are met (e.g., max depth, minimum samples per leaf).
- Outputs a final prediction based on leaf nodes.

Decision Tree Structure



Splitting Criteria (1)

1. Gini Impurity:

$$Gini = 1 - \sum_{i=1}^{C} p_i^2 \tag{1}$$

2. Entropy:

$$Entropy = -\sum_{i=1}^{C} p_i \log_2 p_i \tag{2}$$

Splitting Criteria (Cont.)

3. Information Gain:

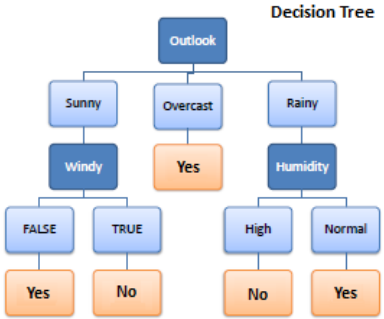
$$IG = Entropy(parent) - \sum_i \frac{|subset_i|}{|parent|} \times Entropy(subset_i) \tag{3}$$

4. Mean Squared Error (for Regression):

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{4}$$

Example

Predictors				Target
Outlook	**Temp**	**Humidity**	**Windy**	**Play Golf**
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No



Example: Medical Diagnosis

- Features: Age, Cholesterol, Blood Pressure, Smoking Status.
- Example Rule:
  - If Cholesterol > 240 and Blood Pressure > 140, classify as High Risk.
  - Else, classify as Low Risk.

Example: Insurance Risk Prediction

- Features: Age, Driving History, Number of Accidents, Type of Car.
- Example Rule:
  - If Age < 25 and Accidents > 2, classify as High Risk.
  - If Age >= 25 and No Accidents, classify as Low Risk.

Advantages of Decision Trees

- Easy to interpret and visualize.
- Handles both numerical and categorical data.
- Requires little data preprocessing.
- Works well with large datasets.

Disadvantages of Decision Trees

- Can overfit the data (prone to high variance).
- Sensitive to noisy data.
- Can create biased results if dataset is imbalanced.

Improving Decision Trees

- **Pruning:** Reducing tree size to avoid overfitting.
- **Ensemble Methods:** Combining multiple trees (Random Forest, Gradient Boosting).
- **Feature Selection:** Choosing relevant features improves accuracy.

Overview

Definition: Overfitting occurs when a decision tree learns patterns specific to the training data but fails to generalize to unseen data.

- The tree becomes too complex and captures noise instead of the true pattern.
- Leads to high accuracy on training data but poor performance on test data.

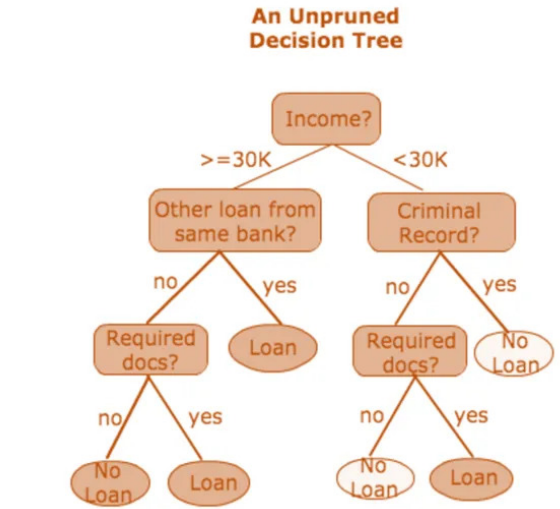
Causes of Overfitting:

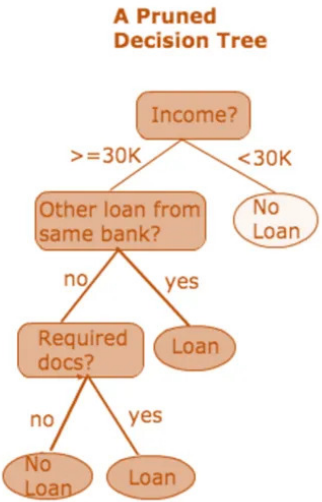
- Deep trees with too many splits.
- Small leaf nodes capturing noise.
- High variance in data leading to unstable decision boundaries.

Preventing Overfitting in Decision Trees

1. **Pruning**
   - **Pre-pruning (Early Stopping):** Stop tree growth based on depth or information gain threshold.
   - **Post-pruning:** Grow the full tree, then remove branches that do not improve validation accuracy.
2. **Restricting Tree Growth**
   - Limit maximum depth of the tree.
   - Require a minimum number of samples per leaf node.
   - Set a minimum number of samples needed to split a node.

A Pruned Tree





Advanced Techniques to Control Overfitting

3. **Ensemble Methods**
   - **Random Forests:** Combine multiple trees and average predictions to reduce variance.
   - **Boosting (e.g., AdaBoost, Gradient Boosting):** Train sequential trees that correct previous errors.
4. **Regularization Techniques**
   - **Cost Complexity Pruning (CCP):** Penalizes complex trees by adding a cost term for additional nodes.

For more information and code check the related notebook

End of Classification