Detailed Notes on Decision Trees

1. Introduction to Decision Tree

- **Definition**: A Decision Tree is a supervised machine learning algorithm used for both classification and regression tasks. It splits the dataset into subsets based on feature values, creating a tree-like structure of decisions.
- **Key Concepts**:
    - **Nodes**: Represent decisions or tests on features.
    - **Branches**: Represent the outcome of a decision (e.g., Yes/No).
    - **Leaf Nodes**: Final output (class label or continuous value).
- **Types of Decision Trees**:
    - **ID3**: Allows multi-way splits (more than two branches).
    - **CART (Classification and Regression Trees)**: Only allows binary splits (two branches per node).
- **Usage**: Decision Trees are foundational for ensemble techniques like Random Forest and Boosting.

---

2. Entropy and Gini Impurity

- **Entropy**:
    - **Definition**: A measure of impurity or disorder in a dataset. It quantifies the uncertainty in the data.
    - **Formula**:

$$H(S) = -p_+ \log_2(p_+) - p_- \log_2(p_-)$$

where p_+ and p_- are the probabilities of positive and negative classes.

-
    - **Range**: 0 (pure) to 1 (impure).
    - **Interpretation**: Higher entropy indicates more disorder.
- **Gini Impurity**:
    - **Definition**: Another measure of impurity, often used in CART. It calculates the probability of misclassifying a randomly chosen element.

- **Formula**:

$$Gini(S)=1-\sum_{i=1}^{n} p_i^2$$

where p_i is the probability of class i.

- **Range**: 0 (pure) to 0.5 (impure).
- **Interpretation**: Lower Gini Impurity indicates a better split.

---

3. Information Gain

- **Definition**: Information Gain measures the reduction in entropy or Gini Impurity after a dataset is split on a feature. It helps in selecting the best feature for splitting.
- **Formula**:

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

where:

- $H(S)$: Entropy of the parent node.
- S_v: Subset of data after splitting on feature A.
- $\frac{|S_v|}{|S|}$: Weight of the subset.

- **Usage**: The feature with the highest Information Gain is chosen for splitting.

---

4. Entropy vs Gini Impurity

- **Entropy**:
  - **Pros**: More sensitive to changes in class probabilities.
  - **Cons**: Computationally expensive due to logarithmic calculations.
  - **Best for**: Smaller datasets.
- **Gini Impurity**:
  - **Pros**: Faster to compute as it doesn't involve logarithms.
  - **Cons**: Less sensitive to class probability changes.
  - **Best for**: Larger datasets.

- **Default Choice**: Most libraries (e.g., scikit-learn) use Gini Impurity by default.

---

5. Decision Tree Split for Numerical Features

- **Process**:
    1. **Sort the Feature Values**: Arrange numerical values in ascending order.
    2. **Create Thresholds**: Consider each value as a potential threshold for splitting.
    3. **Calculate Information Gain**: For each threshold, calculate the Information Gain or Gini Impurity.
    4. **Select Best Threshold**: Choose the threshold that maximizes Information Gain or minimizes Gini Impurity.
- **Disadvantage**: High time complexity for large datasets due to the need to evaluate multiple thresholds.

---

6. Post Pruning & Pre Pruning

- **Post Pruning**:
    - **Definition**: After constructing the full decision tree, remove branches that do not contribute significantly to the model's performance.
    - **Process**:
        1. Build the complete tree.
        2. Prune branches that lead to overfitting.
    - **Best for**: Smaller datasets.
- **Pre Pruning**:
    - **Definition**: Stop the tree construction process early by setting constraints (hyperparameters) to prevent overfitting.
    - **Hyperparameters**:
        - **Max Depth**: Limits the depth of the tree.
        - **Min Samples Split**: Minimum number of samples required to split a node.

        - **Min Samples Leaf**: Minimum number of samples required in a leaf node.
    - **Best for**: Larger datasets.

---

7. Decision Tree Regression

- **Definition**: A Decision Tree used for regression tasks where the output is a continuous value.
- **Key Differences from Classification**:
    - **Splitting Criterion**: Uses **Variance Reduction** instead of Entropy or Gini Impurity.
    - **Output**: The average value of the target variable in the leaf node.
- **Variance Reduction**:
    - **Formula**:

$$\text{Variance Reduction} = \text{Var}(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \text{Var}(S_i)$$

where:

        - $\text{Var}(S)$: Variance of the parent node.
        - $\text{Var}(S_i)$: Variance of the child node.
    - **Process**:
        1. Calculate the variance of the parent node.
        2. Calculate the variance of child nodes after splitting.
        3. Select the split that maximizes variance reduction.
- **Example**: Predict salary based on experience and career gap. The tree splits based on thresholds and calculates the average salary in each leaf node.

Summary

- **Decision Trees** are versatile algorithms for both classification and regression.
- **Entropy** and **Gini Impurity** are used to measure impurity in classification tasks.

- **Information Gain** helps in selecting the best feature for splitting.
- **Post Pruning** and **Pre Pruning** are techniques to prevent overfitting.
- **Decision Tree Regression** uses **Variance Reduction** to split continuous data and predict average values in leaf nodes.

Possible interview questions related to Decision Trees

---

1. What is a Decision Tree?

- **Answer**: A Decision Tree is a supervised machine learning algorithm used for classification and regression. It splits data into subsets based on feature values, creating a tree-like structure of decisions, where nodes represent features, branches represent decisions, and leaves represent outcomes.

---

2. What are the types of Decision Trees?

- **Answer**:
  - **ID3**: Allows multi-way splits (more than two branches per node).
  - **CART (Classification and Regression Trees)**: Uses binary splits (only two branches per node) and is commonly used in libraries like scikit-learn.

---

3. What is Entropy in Decision Trees?

- **Answer**: Entropy measures the impurity or disorder in a dataset. It ranges from 0 (pure) to 1 (impure). The formula is:

$H(S)=-p_+\log_2(p_+)-p_-\log_2(p_-)$$H(S)=-p_+\log_2(p_+)-p_-\log_2(p_-)$

where p_+ and p_- are probabilities of positive and negative classes.

---

4. What is Gini Impurity?

- **Answer**: Gini Impurity measures the probability of misclassifying a randomly chosen element. It ranges from 0 (pure) to 0.5 (impure). The formula is:

$$Gini(S)=1-\sum_{i=1}^{n} p_i^2$$

where p_i is the probability of class i.

---

5. What is Information Gain?

- **Answer**: Information Gain measures the reduction in entropy or Gini Impurity after splitting a dataset on a feature. It helps select the best feature for splitting. The formula is:

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

where $H(S)$ is the entropy of the parent node, and S_v is the subset after splitting.

---

6. What is the difference between Entropy and Gini Impurity?

- **Answer**:
  - **Entropy**: More sensitive to changes in class probabilities, but computationally expensive due to logarithms. Best for smaller datasets.
  - **Gini Impurity**: Faster to compute, less sensitive to class probabilities. Best for larger datasets. Most libraries use Gini by default.

---

7. How does a Decision Tree handle numerical features?

- **Answer**:
  1. Sort the numerical feature values.
  2. Consider each value as a threshold for splitting.
  3. Calculate Information Gain or Gini Impurity for each threshold.

4. Select the threshold that maximizes Information Gain or minimizes Gini Impurity.

---

8. What is Post Pruning and Pre Pruning?

- **Answer**:
    - **Post Pruning**: Build the full tree first, then remove branches that do not contribute to model performance to prevent overfitting.
    - **Pre Pruning**: Stop tree growth early by setting constraints like max depth, min samples split, or min samples leaf to prevent overfitting.

---

9. What is Variance Reduction in Decision Tree Regression?

- **Answer**: Variance Reduction is used in Decision Tree Regression to split continuous data. It measures the reduction in variance after splitting. The formula is:

Variance Reduction=Var(S)−∑i=1n|Si||S|Var(Si)Variance Reduction=Var(*S*)−*i*=1∑*n*|*S*||*Si*|Var(*Si*)

where Var(S)Var(*S*) is the variance of the parent node, and Var(Si)Var(*Si*) is the variance of child nodes.

---

10. What are the advantages of Decision Trees?

- **Answer**:
    - Easy to understand and interpret (visualizable).
    - Can handle both numerical and categorical data.
    - Requires little data preprocessing (e.g., no need for scaling).

---

11. What are the disadvantages of Decision Trees?

- **Answer**:
    - Prone to overfitting, especially with deep trees.
    - Sensitive to small changes in data.
    - Can create biased trees if some classes dominate.

12. How do you prevent overfitting in Decision Trees?

- **Answer**:
  - Use **Post Pruning** to remove unnecessary branches.
  - Use **Pre Pruning** by setting constraints like max depth, min samples split, or min samples leaf.
  - Use ensemble methods like Random Forest or Gradient Boosting.

---

13. What is the time complexity of building a Decision Tree?

- **Answer**: The time complexity is O(n·m·log(m))$O(n \cdot m \cdot \log(m))$, where:
  - nn: Number of samples.
  - mm: Number of features.
  - log(m)$\log(m)$: Depth of the tree.

---

14. Can Decision Trees handle missing values?

- **Answer**: Yes, Decision Trees can handle missing values by:
  - Using surrogate splits (alternative splits when data is missing).
  - Assigning the most common value or using imputation techniques.

---

15. What is the difference between Decision Tree Classifier and Regressor?

- **Answer**:
  - **Classifier**: Predicts discrete class labels (e.g., Yes/No). Uses Entropy or Gini Impurity for splitting.
  - **Regressor**: Predicts continuous values (e.g., salary). Uses Variance Reduction for splitting.

---

16. What is the role of the root node in a Decision Tree?

- **Answer**: The root node is the topmost node in the tree, representing the feature that provides the best split (highest Information Gain or lowest Gini Impurity) to start the tree.

---

17. What is a leaf node in a Decision Tree?

- **Answer**: A leaf node is the final node in a Decision Tree, representing the predicted class (in classification) or the average value (in regression).

---

18. What is the difference between Random Forest and Decision Tree?

- **Answer**:
  - **Decision Tree**: A single tree prone to overfitting.
  - **Random Forest**: An ensemble of multiple Decision Trees, reducing overfitting and improving accuracy by averaging predictions.

---

19. How do you choose the best feature for splitting in a Decision Tree?

- **Answer**: Use **Information Gain** (for Entropy) or **Gini Impurity** to evaluate features. The feature with the highest Information Gain or lowest Gini Impurity is chosen for splitting.

---

20. What is the output of a Decision Tree Regressor?

- **Answer**: The output is the **average value** of the target variable in the leaf node. For example, if the leaf node contains salaries [40k, 42k], the predicted output is 41k.