**Random Forests**

**Practical Notes for Exams**

**1. Comparison: Random Forest vs. Decision Trees**

|  |  |  |
| --- | --- | --- |
| **Aspect** | **Decision Tree** | **Random Forest** |
| **Overfitting** | Prone to overfitting, especially on noisy data. | Reduces overfitting by averaging multiple trees. |
| **Robustness** | Sensitive to small changes in data, leading to different trees. | Robust to data variability due to ensemble averaging. |
| **Stability** | Produces a single tree, which can be unstable. | Combines multiple trees, making predictions more stable. |
| **Accuracy** | Lower accuracy for complex datasets. | Higher accuracy due to ensemble voting or averaging. |

**2. Bagging vs. Boosting**

|  |  |  |
| --- | --- | --- |
| **Aspect** | **Bagging (Random Forest)** | **Boosting (Gradient Boosting)** |
| **Approach** | Trains trees independently on random subsets (bootstrapping). | Trains trees sequentially, each focusing on correcting previous errors. |
| **Objective** | Reduces variance. | Reduces bias by improving weak learners. |
| **Overfitting** | Less prone due to averaging. | Can overfit if not regularized properly. |
| **Computation** | Parallel training possible. | Sequential training, harder to parallelize. |
| **Strengths** | Simpler, robust, reduces overfitting. | Better for complex patterns, more accurate but sensitive to hyperparameters. |

**3. Feature Importance in Random Forest**

Random Forest evaluates feature importance using:

1. **Gini Importance (Mean Decrease in Impurity)**:
   * Measures how much each feature reduces impurity (e.g., Gini Index) across all trees.
   * Higher reductions = Higher importance.
2. **Permutation Importance**:
   * Randomly permutes the values of a feature and measures the drop in model accuracy.
   * Larger drops indicate more important features.

Key Points for Exams:

* Feature importance is derived from the overall contribution of a feature in splitting nodes across the forest.
* Random Forest helps identify dominant features in datasets for dimensionality reduction or further analysis.

**Key Concepts of Random Forests:**

1. **Ensemble Method**: Random Forest combines multiple decision trees to form a "forest." It uses majority voting (classification) or averaging (regression) to improve predictions.
2. **Bootstrapping**: Each tree in the forest is trained on a different subset of the training data (with replacement). This technique is called bagging.
3. **Feature Randomness**: At each split, a random subset of features is considered to reduce correlation between trees.
4. **Out-of-Bag (OOB) Error**: Since each tree uses only a subset of data, the unused samples (OOB samples) can evaluate the model's performance without separate validation data.

**Steps in Random Forest Construction:**

1. Draw multiple bootstrap samples from the original dataset.
2. Build a decision tree for each sample, splitting nodes based on the best feature from a random subset of features.
3. Aggregate predictions from all trees for final output (majority vote or average).

**Advantages:**

1. **Reduces Overfitting**: By averaging multiple trees, it reduces the variance compared to a single decision tree.
2. **Robustness**: Handles missing data, noisy data, and imbalanced datasets well.
3. **Feature Importance**: Provides a straightforward way to calculate feature importance.