Answers

1)R-squared is a better measure of goodness of fit.

* R-squared shows the proportion of variance explained by the model, making it easier to interpret.
* RSS only measures the total error, without considering the overall variability in the data.

2) TSS, ESS, and RSS in Brief

* TSS (Total Sum of Squares): Total variation in the data.

* ESS (Explained Sum of Squares): Variation explained by the regression model.

* RSS (Residual Sum of Squares): Variation not explained by the model (error).

Relationship: TSS = ESS + RSS

Essentially, the total variation in the data can be divided into the part explained by the model and the part that remains unexplained.

3) Regularization prevents models from becoming too complex and memorizing training data (overfitting). It improves a model's ability to make accurate predictions on new data.

4) Gini Impurity is a measure of how impure or mixed a dataset is with respect to its class distribution. It's used in decision trees to determine the best split for data.

5) Yes, unregularized decision trees are prone to overfitting.

This happens because:

* Deep trees: They can become extremely deep, capturing noise and outliers in the data.

* Memorization: The tree might memorize the training data perfectly but fail to generalize to new data.

* Lack of stopping criteria: Without limits, the tree can continue splitting until every data point is in its own leaf, leading to overfitting.

To combat this, techniques like pruning and setting maximum depth are used.

6) Ensemble techniques combine multiple models to create a better model. Think of it as a team of experts making a decision together. This often leads to more accurate and reliable predictions.

7) Bagging creates multiple models independently and combines their results, reducing variance. Boosting creates models sequentially, learning from previous mistakes, reducing bias.

8) Out-of-bag error is a way to estimate a random forest's accuracy without using a separate test set. It uses data points not used in building each tree to evaluate the model's performance.

9)K-fold cross-validation divides data into K parts, trains on K-1 parts, tests on the remaining part, and repeats to get a reliable estimate of model performance.

10) Hyperparameter tuning is finding the best settings for a machine learning model to improve its performance.

11) A large learning rate in Gradient Descent can lead to:

* Overshooting the minimum: The algorithm might jump over the optimal solution.

* Divergence: The algorithm might fail to converge and keep oscillating.

* Instability: The model's parameters can fluctuate wildly.

* Difficulty in finding the global minimum: It's more likely to get stuck in local minima.

12) No, Logistic Regression is not ideal for non-linear data. It's designed for linearly separable data. For complex patterns, use other methods.

13) Adaboost vs. Gradient Boosting

Adaboost and Gradient Boosting are both ensemble techniques that combine multiple weak learners to create a strong predictive model. However, they differ in their approach:

Adaboost

* Focuses on misclassified instances: Assigns higher weights to misclassified data points in subsequent iterations.

* Weights weak learners: Each weak learner is assigned a weight based on its performance.

* Less flexible: Primarily designed for classification problems.

Gradient Boosting

* Focuses on minimizing a loss function: Treats the model as an additive function and fits new models to reduce the loss.

* Uses gradient descent: Optimizes the loss function by iteratively moving in the direction of the steepest descent.

* More flexible: Can be applied to both classification and regression problems.

In essence: Adaboost focuses on correcting mistakes, while Gradient Boosting aims to minimize the overall error. Gradient Boosting is generally more flexible and often outperforms Adaboost in terms of accuracy.

14) Bias-variance trade-off is the challenge of finding the right balance between simplicity and complexity in a machine learning model.

* Bias: Error due to overly simplistic assumptions. High bias leads to underfitting.

* Variance: Sensitivity to small changes in the training data. High variance leads to overfitting.

The goal is to find a model that minimizes both bias and variance for optimal performance on new data.

15) SVM Kernels

Linear Kernel

* Simplest kernel.

* Assumes data is linearly separable.

* Efficient for large datasets.

Polynomial Kernel

* Captures non-linear relationships between data points.

* Degree of polynomial determines complexity.

* Can be computationally expensive for high-degree polynomials.

RBF Kernel

* Most commonly used kernel.

* Handles complex non-linear relationships.

* Often a good starting point for many problems.

* Has a hyperparameter (gamma) that controls the kernel's behavior.