Machine Learning

Answer- 4

Gini Impurity (Index) is a method that measures the impurity of a dataset. The more impure the dataset, the higher is the Gini index. The term “Impurity” indicates the number of classes present within a subset. The more distinct classes included in a subset, the higher the impurity.

Answer- 3

Regularization is a technique to prevent the model from overfitting by adding extra information to it. . It helps to reduce the generalization error and not the training error by ignoring less important features and decreasing the complexity of a model. This technique can be used in such a way that it will allow to maintain all variables or features in the model by reducing the magnitude of the variables. Hence, it maintains accuracy as well as a generalization of the model.

Answer- 2

RSS- The residual sum of squares is used to help you decide if a statistical model is a good fit for your data. It measures the overall difference between your data and the values predicted by your estimation model.

TSS- The total sum of squares is a variation of the values of a dependent variable from the sample mean of the dependent variable. Essentially, the total sum of squares quantifies the total variation in a sample.

ESS- The explained sum of squares is a statistical quantity used in describing how well a model, often a regression model, represents the data being modelled. It measures the amount of variation in the data that is explained by the model, as opposed to the variation that is unexplained or due to error.

TSS = ESS + RSS

Answer- 1

Both R-Squared and RSS are important metrics but they both serve different purposes.

R-Squared measures the proportion of the variance in the dependent variable (response) that is explained by the independent variables (predictors) in the regression model. R-squared is commonly used to evaluate linear regression models.

A higher R-squared value (closer to 1) indicates a better fit of the model to the data

- R- Squared Provides an overall assessment of model performance.
- Widely used in practice.

R-Squared can be misleading when the model is overfitting (fits noise rather than true patterns). Doesn't tell us about the quality of individual coefficient estimates.

RSS is the sum of the squared differences between the actual observed values and the predicted values (Residuals) from the regression model.

A lower RSS indicates a better fit because it measures how well the model fits the data points.

- Reflects the magnitude of prediction errors.
- Useful for assessing the fit of individual data points.

RSS alone doesn't provide a clear sense of overall model fit. It depends on the scale of the response variable.

R-squared is a better measure of overall goodness of fit, because R-squared focuses on the overall fit of the model, while RSS zooms in on individual prediction errors. RSS is used when we need to understand the magnitude of residuals for specific data points whereas R squared is used when we need a global assessment of model performance. Therefore RSS provides valuable information about individual prediction errors, so both are important metrics.

Answer-5

Yes, an unregularized decision-tree prone to overfitting. Overfitting occurs when a decision tree model becomes too complex, capturing noise and irrelevant patterns in the training data, instead of the underlying true patterns. When the available training data is limited, decision trees tend to overfit. Decision trees can grow to a considerable depth, resulting in intricate decision boundaries. As the tree becomes deeper, it becomes more susceptible to overfitting. Including irrelevant or noisy features in the training data can mislead the decision tree algorithm, leading to overfitting.

Answer- 6

Ensemble means 'a collection of things' and in Machine Learning terminology, Ensemble learning refers to the approach of combining multiple ML models to produce a more accurate and robust prediction compared to any individual model. The combined models increase the accuracy of the results significantly. Ensemble learning is a technique that takes the help of several base models and combines their output to produce an optimized model. This type of machine learning algorithm helps in improving the overall performance of the model.

Answer- 7

Bagging and boosting are two types of ensemble techniques that combine multiple models to improve the accuracy and stability of predictions. Bagging uses bootstrap sampling to create different subsets of data and trains the same type of model on each subset. Boosting uses weighted sampling to create different subsets of data and trains different types of models sequentially, adjusting the weights based on the previous errors.
Bagging reduces variance and overfitting, while boosting reduces bias and underfitting.

Answer- 8

A random forest is an ensemble machine-learning model that is composed of multiple decision trees. OOB (out-of-bag) errors are an estimate of the performance of a random forest classifier or regressor on unseen data. The OOB error is computed using the samples that were not included in the training of the individual trees. The OOB error can be useful for evaluating the performance of the random forest on unseen data.

Answer- 9

Cross-validation is a statistical method used to estimate the skill of machine learning models. There are common tactics that you can use to select the value of k for your dataset.
There are commonly used variations on cross-validation such as stratified and repeated. K-fold cross validation is a procedure used to estimate the skill of the model on new data. Cross-validation is primarily used in applied machine learning to estimate the skill of a machine learning model on unseen data. That is, to use a limited sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

Answer- 10

Hyperparameter tuning is the process of selecting the optimal values for a machine learning model's hyperparameters. Hyperparameters are settings that control the learning process of the model, such as the learning rate, the number of neurons in a neural network, or the kernel size in a support vector machine. The goal of hyperparameter tuning is to find the values that lead to the best performance on a given task. The purpose of hyperparameter tuning is to find the best set of hyperparameters for a given machine learning model. This can improve the model's performance on unseen data, prevent overfitting, and reduce training time.

This hyperparameter controls the step size taken by the optimizer during each iteration of training. Hypermeter tuning Reduce overfitting and underfitting.It is used for Enhancing Data Quality and Model Performance, Model Selection.

Answer- 11

Learning rate is crucial for successful convergence in Gradient Descent, too large a rate can cause instability and divergence. When using Gradient Descent, having a large learning rate can lead to several issues:

Overshooting the Minimum: The learning rate acts as a step size (denoted as η). If the learning rate is too large, it can "jump over" the minima we are trying to reach. . In other words, it overshoots the optimal point, leading to oscillations around the minimum or even outright divergence.

Instability: When the learning rate is too large, the optimization process becomes unstable. Coefficients can explode, and overflow errors may occur.

Divergence: A large learning rate can cause the algorithm to diverge instead of converging to the minimum.

Answer- 12

No, we cannot use Logistic Regression for classification of Non-Linear Data. Logistic regression is a supervised machine learning algorithm that can be used for binary classification, where the goal is to predict the probability that an instance belongs to a given class or not. Logistic regression uses a logistic function, also known as a sigmoid function, to map the predicted values to probabilities between 0 and 1. However, logistic regression cannot handle complex non-linear decision boundaries that cannot be approximated by a polynomial function. For example, if the data has a circular or spiral shape, logistic regression will not be able to separate the classes well.

Answer- 13

Adaboost and Gradient Boosting both enhance model performance by combining weak learners, but they differ in weight assignment, correlation, and handling of loss functions.

AdaBoost combines multiple **weak learners** (usually decision stumps) to create a **strong learner**, while Similar to AdaBoost, Gradient Boosting combines weak learners into a strong one.

In Adaboost each weak learner contributes to the final prediction with a weight based on its performance. Correctly classified samples receive lower weights, while misclassified ones receive higher weights, but Unlike AdaBoost, Gradient Boosting minimizes the loss function (e.g., mean squared error or cross-entropy) of the previous model using gradient descent.

AdaBoost trains models sequentially, where each subsequent model focuses on the mistakes made by the previous ones. In Gradient Boostuing : models are trained sequentially, with each new model fitting the residual errors of the previous one.

Adaboost minimizes the exponential loss function.Gradient Boosting minimizes the loss function directly.

Answer- 14

In statistics and machine learning, the bias–variance tradeoff describes the relationship between a model's complexity, the accuracy of its predictions, and how well it can make predictions on previously unseen data that were not used to train the model. In general, as we increase the number of tunable parameters in a model, it becomes more flexible, and can better fit a training data set. It is said to have lower error, or bias. However, for more flexible models, there will tend to be greater variance to the model fit each time we take a set of samples to create a new training data set. It is said that there is greater variance in the model's estimated parameters.

Answer-15

The linear kernel is employed when the data is linearly separable. It computes the inner product between two points in the original feature space. Essentially, it transforms the data into a higher-dimensional space where a linear decision boundary can be established.

The RBF kernel is the most commonly used type. It provides a localized and finite response along the entire x-axis. It transforms the data into a feature space using a Gaussian function, allowing for non-linear decision boundaries.

The polynomial kernel captures the similarity of vectors in the training data by considering polynomials of the original variables. It allows for non-linear decision boundaries by mapping the data to a higher-dimensional space.