MACHINE LEARNING

Q1 to Q15 are subjective answer type questions, Answer them briefly.

1. R-squared or Residual Sum of Squares (RSS) which one of these two is a better measure of goodness of fit model in regression and why?

Answer: R-squared and Residual Sum of Squares are both commonly used measures to know the goodness of fit of a regression model, but they capture different aspects of model performance and analysis, and the choice between them depends on the context and what you want to evaluate.

Because R-squared is a statistical measure used to evaluate the goodness of fit of a regression model. It provides insights into how well the model explains the variability in the data.

2. What are TSS (Total Sum of Squares), ESS (Explained Sum of Squares) and RSS (Residual Sum of Squares) in regression. Also mention the equation relating these three metrics with each other.

Answer :- **TSS** (Total Sum of Squares) is describe the sum of squared differences between the observed dependent variables and the overall mean **.**

(ESS) (Explained Sum of Squares) is the sum of the differences between the predicted value and the mean of the dependent variable. In other words, it describes how well our line fits the data.

RSS(Residual Sum of Squares) is describe The residual sum of squares is a key metric that gives a numerical representation of how well your regression model fits the data. Specifically, RSS indicates how well an independent variable predicts a dependent variable. This technique works for long term as well as forecasting with limited data.

The relationship between the three metrics is as follows:

SST = SSR + RSS

3. What is the need of regularization in machine learning?

Answer : - Regularization is a set of methods for reducing overfitting in machine learning models and it is needed when we used to minimize the problem of overfitting ,the result is that the model generalizes well on the unseen data once overfitting is minimized.

4. What is Gini–impurity index?

Answer :- The Gini impurity measure is one of the methods used in decision tree algorithms to decide the optimal split from a root node, and subsequent splits and also a powerful measure of the randomness or the impurity or entropy in the values of a dataset.
Gini Index aims to decrease the impurities from the root nodes (at the top of decision tree) to the leaf nodes (vertical branches down the decision tree) of a decision tree model

5. Are unregularized decision-trees prone to overfitting? If yes, why?

Answer :- Overfit condition arises when the model memorizes the noise of the training data and fails to capture important patterns. A perfectly fit decision tree performs well for training data but performs poorly for unseen test data.

Decision trees are prone to overfitting, especially when a tree is particularly deep.This is due to the amount of specificity we look at leading to smaller sample of events that meet the previous assumptions.

6. What is an ensemble technique in machine learning?

Answer:- Ensemble methods are techniques that create multiple models and then combine them to produce improved results. Ensemble methods in machine learning usually produce more accurate solutions than a single model would.

7. What is the difference between Bagging and Boosting techniques?

Answer:- Bagging is the simplest way of combining predictions that belong to the same type while Boosting is a way of combining predictions that belong to the different types. Bagging aims to decrease variance, not bias while Boosting aims to decrease bias, not variance.

8. What is out-of-bag error in random forests?

Answer:- The OOB Error provides an unbiased estimate of the model's performance without the need for a separate validation set .It serves as an internal validation mechanism within the random forest algorithm.

9. What is K-fold cross-validation?

Answer:- K-fold cross-validation is a technique for evaluating predictive models. The dataset is divided into k subsets or folds. The model is trained and evaluated k times, using a different fold as the validation set each time. Performance metrics from each fold are averaged to estimate the model's generalization performance.

10. What is hyper parameter tuning in machine learning and why it is done?

Answer:- When you're training machine learning models, each dataset and model needs a different set of hyperparameters, which are a kind of variable. The only way to determine these is through multiple experiments, where you pick a set of hyperparameters and run them through your model. This is called hyperparameter tuning.

11. What issues can occur if we have a large learning rate in Gradient Descent?

Answer:- learning rate impact the performance of gradient descent.If the learning rate is too high, the algorithm may overshoot the minimum, and if it is too low, the algorithm may take too long to converge.

12. Can we use Logistic Regression for classification of Non-Linear Data? If not, why?

Answer:- Logistic Regression has traditionally been used as a linear classifier, i.e. when the classes can be separated in the feature space by linear boundaries. That can be remedied however if we happen to have a better idea as to the shape of the decision boundary.

13. Differentiate between Adaboost and Gradient Boosting.

Answer:- AdaBoost is the first designed boosting algorithm with a particular loss function. On the other hand, Gradient Boosting is a generic algorithm that assists in searching the approximate solutions to the additive modelling problem. This makes Gradient Boosting more flexible than AdaBoost.

14. What is bias-variance trade off in machine learning?

Answer :- In machine learning, as we are try to minimize one component of the error (e.g.,bias) the other component (e.g.variance) tends to increase, and vice versa. Finding the right balance of bias and variance is key to creating an effective and accurate model. This is called the bias-variance tradeoff.

15. Give short description each of Linear, RBF, Polynomial kernels used in SVM.

Answer:- The linear kernel is the simplest SVM kernel, representing a linear decision boundary. It computes the dot product of two input vectors.

In machine learning, the radial basis function kernel, or RBF kernel, is a popular kernel function used in various kernelized learning algorithms. In particular, it is commonly used in support vector machine classification.

In machine learning, the polynomial kernel is a kernel function commonly used with support vector machines (SVMs) and other kernelized models, that represents the similarity of vectors (training samples) in a feature space over polynomials of the original variables, allowing learning of non-linear models.

FLIP ROBO