

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?

Ans. both measures are important and should be considered together when evaluating the goodness of fit of a model. However, R-squared is often considered to be a better measure of goodness of fit than RSS because it provides a single number that summarizes the proportion of variance in the dependent variable that is explained by the model, which is more interpretable and easier to compare across models.

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.

Ans. Total sum of squares (TSS) → it is the sum of squared differences between the observed dependent variables and the overall mean.

Sum of squares (ESS) → it is the sum of the squares of the deviations of the predicted values from the mean value of a response variable, in a standard regression model.

Residual sum of squares (RSS) → measures the level of variance in the error term, or residuals, of a regression model.

Equation → $TSS = ESS + RSS$

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

Ans. When trained a machine learning model, the model can easily be overfitted or under fitted. To avoid this, we use regularization in machine learning to properly fit a model onto our test set. Regularization techniques help reduce the chance of overfitting and help us get an optimal model

4. What is Gini-impurity index?

Ans. Gini Impurity is a measurement used to build Decision Trees to determine how the features of a dataset should split nodes to form the tree.

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

Ans. Yes, 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. This small sample could lead to unsound conclusion.

6. What is an ensemble technique in machine learning?

Ans. Ensemble technique is to create multiple models and then combine them to produce improve result. Ensemble methods in machine learning usually produce more accurate solutions than a single model.

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

Ans. Bagging technique combines multiple models trained on different subsets of data, while boosting trains models sequentially, focusing on the error made by the previous model.

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

Ans. It is a method of measuring the prediction error of random forests, boosted decision trees, and other machine learning models utilizing bootstrap aggregating (bagging).

9. What is K-fold cross-validation?

Ans. K-fold cross-validation is a technique for evaluating predictive models.

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

Ans. When we trained machine learning models then each dataset and model needs different set of hyper parameters, which is a kind of variable. For this way to determine these is through multiple experiments, where we pick a set of hyper parameters and run them through our model. This is called hyperparameter tuning. It is important because it have a direct impact on the behavior of training algorithms and also have a significant impact on the performance of the model being trained.

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

Ans. When the learning rate is too large, gradient descent can suffer from divergence it means that weights increase exponentially, resulting in exploding gradients which can cause problems such as instabilities and overly high loss values.

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

Ans. No. Because By employing the logistic function, which maps any real valued number into a range between 0 and 1 logistic regression models can effectively handle non-linear relationships.

13. Differentiate between Adaboost and Gradient Boosting.

Ans. Gradient Boost and Ada Boost lies in what it does with the underfitted values of its predecessor. Contrary to Adaboost, which tweaks the instance weights at every interaction, this method tries to fit the new predictor to the residual errors made by the previous predictor.

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

Ans. the bias variance tradeoff describe the relationship between a models 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.

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

Ans. Kernel Function is a method used to take data as input and transform it into the required form of processing data. "Kernel" is used due to a set of mathematical functions used in Support Vector Machine providing the window to manipulate the data. So, Kernel Function generally transforms the training set of data so that a non-linear decision surface is able to transform to a linear equation in a higher number of dimension spaces. Basically, It returns the inner product between two points in a standard feature dimension.

