

1 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?
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.
3. What is the need of regularization in machine learning?

Answer= While training 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 on our test set. Regularization techniques help reduce the chance of overfitting and help us get an optimal model.

4. What is Gini-impurity index?

Answer= the Gini Impurity of a dataset is a number between 0-0.5, which indicates the chance of new, random data being misclassified if it were given a random class label according to the class distribution in the dataset

5. Are unregularized decision-trees prone to overfitting? If yes, why?
6. What is an ensemble technique in machine learning?

Answer= Ensemble learning is a technique in machine learning which 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. Here the base model which is most commonly used is the Decision tree classifier

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

Answer= difference between bagging and boosting are-

Bagging

- >The simplest way of combining prediction that belongs to the same type**
- >Aim to decrease variance not bias**
- >each model is built independently**
- >bagging tries to solve the overfitting problem**
- >if the classifier is unstable apply bagging**

Boosting

- >the simplest way of combining prediction data that belongs to different type**
- >Aim to decrease bias not variance**

>New model are influenced by the performance of previous built model

>Boosting tries to reduce bias

>if the classifier is stable and simple apply bagging

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

Answer= The out-of-bag error is the average error for each predicted outcome calculated using predictions from the trees that do not contain that data point in their respective bootstrap sample. This way, the Random Forest model is constantly being validated while being trained.

9. What is K-fold cross-validation?

Answer= K-fold cross-validation improves the model by validating the data. This technique ensures that the model's score does not relate to the technique we use to choose the test or training dataset. K-fold cross-validation method divides the data set into subsets as K number.

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

Answer= Hyperparameters are the knobs or settings that can be tuned before running a training job to control the behavior of an ML algorithm. They can have a big impact on model training as it relates to training time, infrastructure resource requirements (and as a result cost), model convergence and model accuracy.

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

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

Answer= Logistic Regression is not suitable for complex non-linear relationships between the dependent variable and independent variables. It is also not recommended for multi-class classification problems, as it can only handle binary classification. Live model training for Diabetics prediction using logistic regression

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= If the algorithm is too simple then it may be on high bias and low variance condition and thus is error-prone. If algorithms fit too complex then it may be on high variance and low bias. In the latter condition, the new entries will not perform well. Well, there is something between both of these conditions, known as Bias Variance Trade-off.

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

Answer=

Linear in SVM=A simple linear SVM classifier works by making a straight line between two classes. That means all of the data points on one side of the line will represent a category and the data points on the other side of the line will be put into a different category. This means there can be an infinite number of lines to choose from.

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