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: Residual Sum of Squares (RSS) is better as used to measure **the variance** in a data set that is not explained by the regression model.

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 : The sum of squares is used as a mathematical way to find the function that best fits (varies least) from the data in regression. When the sum of squares is done for all data points it is called TSS, if its done for residual data it is called RSS and

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

Answer: To reduce the variance – bias problem regularization is used in Machine Learning.

4. What is Gini–impurity index?

Answer : 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?

Answer : Yes , because overfitting can happen in any ML algorithm in spite of all our efforts.

6. What is an ensemble technique in machine learning?

Answer : In ensemble technique the problem is divided in multiple parts (or trees) and multiple algorithms are applied on those data to get best output by combining the results based on votes of the algorithms.

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

Answer : 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 out-of-bag (OOB) error is the average error for each calculated using predictions from the trees that do not contain in their respective bootstrap sample. This allows the RandomForestClassifier to be fit and validated whilst being trained

9. What is K-fold cross-validation?

Answer : K-fold Cross-Validation is when the dataset is split into a K number of folds and is used to evaluate the model's ability when given new data. K refers to the number of groups the data sample is split into. For example, if you see that the k-value is 5, we can call this a 5-fold cross-validation.

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

Answer : hyper parameter tuning is process of searching the best parameter for a particular algorithms to get highest accuracy in prediction.

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

Answer : A learning rate that is too large **can cause the model to converge too quickly to a suboptimal solution**, whereas a learning rate that is too small can cause the process to get stuck. The challenge of training deep learning neural networks involves carefully selecting the learning rate.

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

Answer : Non-linear problems can't be solved with logistic regression because it has a linear decision surface.

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 : When the model does not learn from data point and gives result on its own assumptions it is called High Bias (Underfitting) and where as when the model learns all the data point then it is called high variance (overfittings) this problem is called bias-variance trade off. To reduce the problem regularization techniques are used.

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

Answer : A **kernel is a function** that takes the original non-linear problem and transforms it into a linear one within the higher-dimensional space. Linear Kernel is used when the data is Linearly separable.

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.