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: R-squared is a statistical measure that represents the goodness of fit of a regression model.
- The ideal value for r-square is 1. The closer the value of r-square to 1, the better is the model fitted. R-square is a comparison of the residual sum of squares (SSres) with the total sum of squares (SStot).

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: Regression sum of squares or Model sum of squares is a statistical quantity used in modelling of a process. ESS gives an estimate of how well a model explains the observed data for the process.
- Residual Sum of Squares (RSS) is a statistical method used to measure the deviation in a dataset unexplained by the regression model. Residual or error is the difference between the observation's actual and predicted value.
- the value estimated by the regression line . In some cases (see below): total sum of squares (TSS) = explained sum of squares (ESS) + residual sum of squares (RSS).

3. What is the need of regularization in machine learning?
   Regularization describes methods for calibrating machine learning models to reduce the adjusted loss function and avoid overfitting or underfitting.

   We can properly fit our machine learning model on a given test set using regularization, which lowers the errors in the test set.

A penalty or complexity term is added to the complex model during regularization. Let's consider the simple linear regression equation:

$y= \beta0+\beta1\times1+\beta2\times2+\beta3\times3+\cdots+\beta nxn +b$

In the above equation, Y represents the value to be predicted

Features for Y are X1, X2, and Xn.

β0,β1,…..βn are the weights or magnitude attached to the features, respectively. Here, stands for the model's bias, and b stands for the intercept.

Now, in order to create a model that can accurately predict the value of Y, we will add a loss function and optimize a parameter. The loss function for the linear regression is called **RSS or residual square sum.**

4. What is Gini–impurity index?
   Ans: Gini index or Gini impurity measures the probability of a particular variable to be wrongly classified when chosen randomly. This measures is calculated where the modeling contains Tree Algorithms like Decision Trees or random forest.
5. Are unregularized decision-trees prone to overfitting? If yes, why?
   Ans: Yes, decision trees are prone to overfitting. But unlike other algorithms decision tree does not use regularization to fight against overfitting. Instead it uses pruning. There are mainly two types of pruning performed:

Pre-pruning that stop growing the tree earlier, before it perfectly classifies the training set.
Post-pruning that allows the tree to perfectly classify the training set, and then post prune the tree.

6. What is an ensemble technique in machine learning?
   Ans: Ensemble techniques are the algorithms created combining multiple weak learners to a strong learning model. Random Forest, XG Boosts, Gradient Boosting are some examples of ensemble learning techniques. These are 2 types of ensemble techniques, Bagging and Boosting.
7. What is the difference between Bagging and Boosting techniques?
   Ans: Bagging, which is also known as bootstrap aggregating sits on top of the majority voting principle. Boosting is another ensemble procedure to make a collection of predictors. In other words, we fit consecutive trees, usually random samples, and at each step, the objective is to solve net error from the prior trees.
8. What is out-of-bag error in random forests?
   Ans: Out of sample is a technique to verify the performance of a boostrapping model with out having to use a validations set. This is an advantage if:
   Your data set is to small to split in to training, validation and test. Gives a second validation on the model allowing.
9. What is K-fold cross-validation?
   Ans: K Fold cross validation means training and testing with different subset of the training and testing data so that the model wont be biased over some cords in the dataset. The K in K fold is the integer defining how any times does the subset should be created and trained and tested.
   For example a 5 Fold cross validation will create 5 subsets in both training and testing dataset, train and predict are output 5 accuracy values. Averaging those values would gives us a greater idea of how good the model is.
10. What is hyper parameter tuning in machine learning and why it is done?
    Ans: Hyper Parameters are the parameters of the model algorithms which are to be tunes in order to get maximum accuracy from the machine learning model.
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 inadvertently increase rather than decrease the training error.
12. Can we use Logistic Regression for classification of Non-Linear Data? If not, why?
    Ans: 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.
13. Differentiate between Adaboost and Gradient Boosting.
    Ans: Gradient boosting defines boosting as a numerical optimization problem where the objective is to minimize the loss function of the model by adding weak learners using gradient descent. Where as, method focuses on training upon misclassified observations. Alters the distribution of the training dataset to increase weights on sample observations that are difficult to classify.
14. What is bias-variance trade off in machine learning?
    Ans: There is a tradeoff between a model's ability to minimize bias and variance. Understanding these two types of errors can help us diagnose model results and avoid the mistake of over- or under- fitting. This is known as bias- variance tradeoff.
15. Give short description each of Linear, RBF, Polynomial kernels used in SVM.
    Ans: SVM also known as support vector machine is a supervised machine learning algorithm which can be used for both classification or regression challenges.
    SVM uses different kernels for different types of questions.

    A linear kernel allows you to use linear functions, which are really impoverished. As you increase the order of the polynomial kernel, the size of the function class increases. In the polynomial kernel, we simply calculate the dot product by increasing the power of the kernel. Gaussian RBF is another popular kernel method uses in SVM models for more. RBF kernel is a function whose value depends on the distance from the origin or from some point.