**MACHINE LEARNING – 5**

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:** The residual sum of squares (RSS) is the absolute amount of explained variation, whereas R-squared is the absolute amount of variation as a proportion of total variation. Typically, however, a smaller or lower value for the RSS is ideal in any model since it means there's less variation in the data set. In other words, the lower the sum of squared residuals, the better the regression model is at explaining 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.

   **ANS:**

   a. **Total sum of squares:** TSS represents the total sum of squares. It is the squared values of the dependent variable to the sample mean. In other words, the total sum of squares measures the variation in a sample.

   $$\text{Total Sum of squares TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

   yi is the one of the value in the sample

   $\bar{y}$ is the sample mean

   b. **ESS (Explained Sum of Squares):** Explained sum of square (ESS) or Regression sum of squares or Model sum of squares is a statistical quantity used in modeling of a process. ESS gives an estimate of how well a model explains the observed data for the process. It tells how much of the variation between observed data and predicted data is being explained by the model proposed. Mathematically, it is the sum of the squares of the difference between the predicted data and mean data.

      Let $y_i = a + b_1x_{1i} + b_2x_{2i} + ... + \varepsilon_i$ is regression model, where:

      $y_i$ is the $i^{th}$ observation of the response variable

      $x_{ji}$ is the $i^{th}$ observation of the $j^{th}$ explanatory variable

      a and bi are coefficients

      i indexes the observations from 1 to n

      $\varepsilon_i$ is the $i^{th}$ value of the error term

      then

$$\mathrm{ESS} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 .$$

This is usually used for regression models. The variation in the modelled values is contrasted with the variation in the observed data (total sum of squares) and variation in modelling errors (residual sum of squares). The result of this comparison is given by ESS as per the following equation:

**ESS = total sum of squares – residual sum of squares**

c. **RSS (Residual Sum of Squares):** Residual Sum of Squares (RSS) is a statistical method that helps identify the level of discrepancy in a dataset not predicted by a regression model. Thus, it measures the variance in the value of the observed data when compared to its predicted value as per the regression model. Hence, RSS indicates whether the regression model fits the actual dataset well or not. Also referred to as the Sum of Squared Errors (SSE), RSS is obtained by adding the square of residuals. Residuals are projected deviations from actual data values and represent errors in the regression model's estimation. A lower RSS indicates that the regression model fits the data well and has minimal data variation

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

   **ANS:** The term "regularisation" 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 particular test set using regularisation, which lowers the mistakes in the test set.

4. What is Gini–impurity index?

   **ANS**: The Gini Index, commonly referred to as Gini impurity, determines the likelihood that a certain feature would be erroneously classified when chosen at random. It can be said to as pure if all of the elements are connected by a single class.

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

   **ANS:** Yes unregularized decision tree prone to overfitting. This is because the level of specificity we consider results in a smaller sample of events that are consistent with the prior hypotheses. Conclusions drawn from this small sample may not be valid.

6. What is an ensemble technique in machine learning?

   **ANS:** By combining numerous models rather than relying just on one, ensemble approaches seek to increase the accuracy of findings in models. The integrated models considerably improve the results' accuracy. As a result, ensemble approaches for machine learning have gained prominence.

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

   **ANS:** By creating additional data for training from a dataset by combining repeats and combinations to create multiple sets of the original data, the bagging strategy lowers prediction variance. Boosting is an iterative technique for modifying the weight of an observation in accordance with the previous classification. It tries to give an observation more weight if it was incorrectly classified. In general, boosting produces accurate predictive models.

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

   **ANS:** The average error for each derived using predictions from the trees that do not contain in their respective bootstrap sample is known as the out-of-bag (OOB) error. This enables fitting and validating the Random Forest Classifier while it is being trained.

9. What is K-fold cross-validation?

   **ANS:** When the dataset is divided into K folds, K-fold cross-validation is used to assess the model's performance when presented with new data. K is the number of groups into which the data sample is divided. For instance, we can refer to this as a 5-fold cross-validation if the k-value is 5.

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

    **ANS:** Finding a set of optimal hyper parameter values for a learning algorithm and using this tuned algorithm on any data set is hyper parameter tuning. The model's performance is maximised by using that set of hyper parameters, which minimises a predetermined loss function and resulting in better results with fewer errors.

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

    **ANS:** If learning rate is too large, gradient descent can overshoot the minimum. It may fail to converge and even diverge. In order for Gradient Descent to work, we must set the learning rate to an appropriate value. This parameter determines how fast or slow we will move towards the optimal weights. If the learning rate is very large we will skip the optimal solution

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

    **ANS:** It can only be used to predict discrete functions. Hence, the dependent variable of Logistic Regression is bound to the discrete number set. It is very fast at classifying unknown records. Non-linear problems can't be solved with logistic regression because it has a linear decision surface.

13. Differentiate between Adaboost and Gradient Boosting.

**ANS:** Adaboost and gradient boosting are types of ensemble techniques applied in machine learning to enhance the efficacy of week learners. The concept of boosting algorithm is to crack predictors successively, where every subsequent model tries to fix the flaws of its predecessor. Boosting combines many simple models into a single composite one. By attempting many simple techniques, the entire model becomes a strong one, and the combined simple models are called week learners. So the adaptive boosting and gradient boosting increases the efficacies of these simple model to bring out a massive performance in the machine learning algorithm.

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

**ANS:** In statistics and machine learning, the bias–variance tradeoff is the property of a model that the variance of the parameter estimated across samples can be reduced by increasing the bias in the estimated parameters. If the algorithm is too simple (hypothesis with linear eq.) then it may be on high bias and low variance condition and thus is error-prone. If algorithms fit too complex (hypothesis with high degree eq.) 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 Trade-off or Bias Variance Trade-off.

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

**ANS:**

a. Linear Kernel

Linear Kernel is used when the data is linearly separable, that is, it can be separated using a single Line. It is one of the most common kernels to be used. It is mostly used when there are a large number of Features in a particular Data Set.

b. RBF Kernel

Radial Basis Function (RBF) Kernel is Machine Learning algorithm like Support Vector Machines for non-linear datasets and you can't seem to figure out the right feature transform. The RBF kernel function for two points $X_1$ and $X_2$ computes the similarity or how close they are to each other. This kernel can be mathematically represented as follows:

$$K(X_1, X_2) = exp(-\frac{||X_1 - X_2||^2}{2\sigma^2})$$

Where,

1. ‘σ’ is the variance and our hyper parameter
2. $||X_1 - X_2||$ is the Euclidean ($L_2$-norm) Distance between two points $X_1$ and $X_2$

c. Polynomial kernel: 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.