Name:- Priti Varma

Batch :- DS2401

Q 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 measure that represents the proportion of the variance in the dependent variable which is explained by the independent variables in the model, RSS measures the total squared difference between the actual values of the dependent variable and the predicted values by the regression model.

Both the models are useful, but R-squared is more commonly used for assessing overall model fit because it provides a standardized measure that is easy to interpret and can be compared across models. On the other hand RSS can be used to understand the magnitude of errors or residuals in the model.

Q 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:- TSS (Total Sum of Squares), ESS (Explained Sum of Squares), and RSS (Residual Sum of Squares) are the important metrics which are used to evaluate the performance of a regression model and understand the distribution of variance in the data.

TSS represents the total variability in the dependent variable (y) without considering any predictors.

ESS shows the variability in the dependent variable (y) that is explained by the regression model with the help of the predictors.

RSS shows the variability in the dependent variable (y) that is not dependant on any predictors in the regression model.

Following is the equation:

TSS = ESS+RSS

Q 3) What is the need of regularization in machine learning?

Ans:- In Machine Learning, sometimes the predicted data or model gives prediction away from the expected target. This happens due to the Bias and the variance happening into the model. Regularization helps in preventing from underfitting or overfitting of variables and minimize the errors. The model is the best fit when it can find all necessary patterns in the data and avoid random data points and unnecessary patterns called noise.

Two types of regularization taught are Lasso Regularization and Ridge Regularization. Elastic net is a combination of Lasso and Ridge.

Q 4) What is Gini–impurity index?

Ans:- The Gini impurity index is a measure used in decision tree algorithms, to find out the impurities of the set of data points. It ranges from 0 to 0.5, where 0 stands for absolute purity that means all the data points belong to the similar class. 0.5 stands for absolute impurity that means all data points are distributed in all types of classes. We use Gini-impurity the minimize the impurity and to change it into more pure and accurate data.

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

Ans:- Unregularized decision-trees are prone to overfitting. Overfitting happens when the model captures all the random data or noise in the training data instead of understanding the said pattern. Following may be some of the reasons:- High variance, lack of generalization, complexity etc…. hence to minimize it we use regularization methods.

Q 6) What is an ensemble technique in machine learning?

Ans:- In Machine Learning, Ensemble technique is used to combine the predictions from the individual models and create more perfect robust model. The mean output or aggregate of the predictions of multiple models are used considering different aspects and achieve a better performance. Some of the techniques taught are Bagging, Boosting and Stacking.

Q 7) What is the difference between Bagging and Boosting techniques?

Ans:- Bagging is a parallel technique where the entire data set is divided into different subsets . These subsets in decision tree classifiers helps us with the mean accuracy parallelly

Boosting is a sequential technique where data set is learned by one of the models and the model will learn from another sequenced model. In short, the base model learns from the previous 'n' models and so on.

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

Ans:- The out-of-bag error in Random Forest is an estimate of the model's performance on unseen data, calculated using the data points that were not included in the bootstrap samples used to train each individual tree in the forest.

Q 9) What is K-fold cross-validation?

Ans:- K-fold cross-validation is one of the techniques apart from HOLD ON and LEAVE ONE OUT, used to assess the performance of a machine learning model by splitting the dataset into K equal folds or subsets. Every time we use different fold as validation set and the remaining folds as training set and hence the model is trained providing more accurate estimates.

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

Ans:- In Machine Learning, Hyperparameter tuning is the process of optimizing the hyperparameters of a machine learning model to improve its performance on unseen data. Hyperparameters are external to the model and cannot be learned from the training data. It is done for several reasons like improving performance, overfitting or underfitting of model, producing an accurate model.

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

Ans:- A large learning rate in Gradient Descent can lead to several issues like:

a. Overshooting the minimum:- This can cause the algorithm to overshoot or surpass the minimum of the loss function which results in instability and oscillation around making it hard to find the right spot.
b. Divergence: The algorithm gets out of control and the situation gets worse instead of better.
c. Instability: The training process becomes shaky and uncertain.
d. Difficulty in finding optimal solution: The algorithm might settle for a less valuable estimate or get stuck without finding the best one.

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

Ans: Logistic Regression is one of the linear classification algorithm. It assumes a linear relationship between the input features and the target variable or labels . Therefore, it may not perform well non-linear data is provided because it cannot capture complex, non-linear relationships between the features (x value) and the target (y value). If the non-linear data can be transformed into Linear feature, only the Logistic Regression can be used.

Q 13) Differentiate between Adaboost and Gradient Boosting?

Ans:- Adaboost and Gradient Boost are the two techniques used under Boosting in ensemble techniques.

Adaboost:-

1) Adaboost Sequentially trains weak learners, adjusting instance weights to focus on misclassified instances.
2) Uses weighted majority voting.
3) Uses simple weak learners.

Gradient Boosting:-

1) Builds trees sequentially, fitting each to the residual errors of the previous trees.
2) Sums predictions.
3) Uses fully grown decision trees.

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

Ans:- Bias refers to the error introduced by the assumptions made by the model. It is unable to capture the true relationship between the features and the target variable. Variance refers to the error introduced by the model's sensitivity to fluctuations in the training data. It performs well on the training data but fails to generalize to new, unseen data. The bias-variance trade-off arises because reducing bias typically increases variance and vice versa. It is important to find the optimal trade-off for building a perfect model.

Q 15) Give short description each of Linear, RBF, Polynomial kernels used in SVM?

Ans:- Linear Kernels:- It is the simplest kernel function used in SVM. It measures the similarity between data points in the original feature space.

RBF Kernels:- It is effective in capturing complex, non-linear relationships in the data.

Polynomial Kernels:- It is another non-linear kernel used in SVM. It is capable of capturing non-linear relationships in the data and is particularly useful for data that exhibits polynomial patterns.