MACHINE LEARNING
ASSIGNMENT - 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?

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?

4. What is Gini–impurity index?

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

6. What is an ensemble technique in machine learning?

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

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

9. What is K-fold cross-validation?

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

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?

13. Differentiate between Adaboost and Gradient Boosting.

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

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

1. R-squared is a better measure of goodness of fit in regression than Residual Sum of Squares (RSS) because it measures the proportion of variation in the dependent variable that is explained by the independent variables. R-squared provides a standardized measure of the goodness of fit that is independent of the scale of the dependent variable, while RSS does not take into account the degrees of freedom and can vary depending on the sample size.

2. TSS (Total Sum of Squares) represents the total variation in the dependent variable, ESS (Explained Sum of Squares) represents the variation in the dependent variable that is explained by the independent variables, and RSS (Residual Sum of Squares) represents the variation in the dependent variable that is not explained by the independent variables. The equation relating these three metrics is TSS = ESS + RSS, where TSS is the sum of squared deviations of the dependent variable from its mean, ESS is the sum of squared deviations of the predicted values from the mean of the dependent variable, and RSS is the sum of squared deviations of the actual values from the predicted values.

3. The need for regularization in machine learning arises to prevent overfitting of the model to the training data, which can lead to poor generalization performance on new data. Regularization techniques such as L1 (Lasso) and L2 (Ridge) regularization impose constraints on the model parameters to reduce their magnitude, thereby reducing the complexity of the model and preventing overfitting.

4. Gini-impurity index is a measure of the impurity or diversity of a set of data. In decision tree algorithms, it is used as a criterion for selecting the best split at each node of the tree. It measures the probability of misclassifying a randomly chosen data point in the set if it were labeled according to the distribution of class labels in the set.
5. Yes, unregularized decision trees are prone to overfitting because they have a high variance and can fit the noise in the data. They can also create complex, overfitted trees that may not generalize well to new data.

6. Ensemble techniques in machine learning combine the predictions of multiple models to improve the overall performance of the system. Examples of ensemble techniques include bagging, boosting, and stacking.

7. Bagging and boosting are two types of ensemble techniques. Bagging (Bootstrap Aggregating) involves training multiple models on different subsamples of the training data and then combining their predictions by averaging or voting. Boosting involves sequentially training models on weighted versions of the data, where the weights are adjusted to emphasize the misclassified data points in each iteration.

8. Out-of-bag error in random forests is an estimate of the generalization error of the model that is computed using the data points that were not used in the construction of each decision tree in the forest. It provides a measure of the model's performance on new data without the need for a separate validation set.

9. K-fold cross-validation is a technique for evaluating the performance of a machine learning model by partitioning the data into K subsets, or folds, and then training the model K times, each time using a different fold as the validation set and the remaining K-1 folds as the training set. The results are then averaged over the K iterations to obtain an estimate of the model's generalization performance.

10. Hyperparameter tuning in machine learning involves selecting the optimal values for the model's hyperparameters, which are parameters that are set before the training process and determine the

complexity of the model. Hyperparameter tuning is done to optimize the model's performance on the validation set and prevent overfitting.

11. If the learning rate in Gradient Descent is too large, it can lead to oscillations or instability in the training. The algorithm may fail to converge or overshoot the optimal solution, leading to poor performance on the validation set.

12. Logistic Regression can be used for classification of Non-Linear Data by transforming the features into a higher-dimensional space using basis functions or kernel methods, such as Polynomial or Radial Basis Function (RBF) kernels. However, if the data is highly non-linear, it may not be possible to find a suitable transformation that can linearize the data.

13. Adaboost and Gradient Boosting are two different types of boosting algorithms. Adaboost (Adaptive Boosting) sequentially trains a series of weak classifiers on different weighted versions of the data, where the weights are adjusted to emphasize the misclassified data points in each iteration. Gradient Boosting trains a series of decision trees on the residuals of the previous trees, where the goal is to minimize the loss function by adding the new tree to the ensemble.

14. The bias-variance trade off in machine learning is the balance between the model's ability to fit the training data (bias) and its ability to generalize to new data (variance). A model with high bias may underfit the data and have poor performance on both the training and validation sets, while a model with high variance may overfit the data and have good performance on the training set but poor performance on the validation set.

15. Linear, RBF, and Polynomial kernels are used in Support Vector Machines (SVMs) to transform the input features into a higher-dimensional space where the data may be linearly separable. A Linear kernel computes the dot product between the input vectors, a Polynomial kernel adds higher-order interactions between the features, and an RBF kernel measures the similarity between the input vectors using a Gaussian function.