MACHINE LEARNING WORKSHEET-5

1. A residual sum of squares (RSS), also known as the sum of squared residuals (SSR), is a statistical technique used to measure the amount of variance in a data set that is not explained by a regression model. R2 (R Squared) is another statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable in a regression model.R-Squared is the better measure of goodness of fit compared to RSS. R-squared explains to what extent the variance of one variable explains the variance of the second variable. So, if the R2 of a model is 0.50, then approximately half of the observed variation can be explained by the model's inputs.
2. Total Sum of Squares-->The Total SS (TSS or SST) tells you how much variation there is in the dependent variable.TSS = Σ(Yi – mean of Y)^2.
   Explained Sum of Squares--> The Explained SS tells you how much of the variation in the dependent variable your model explained.ESS = Σ(Y-Hat – mean of Y)^2.
   Residual Sum of Squares---> The residual sum of squares tells you how much of the dependent variable's variation your model did not explain. It is the sum of the squared differences between the actual Y and the predicted Y: Residual Sum of Squares = Σ e2
   The relationship between the three types of sum of squares can be summarized by the following equation: Relationship Formula TSS=SSR+SSE
3. Regularisation is a technique used to reduce the errors by fitting the function appropriately on the given training set and avoid overfitting. It is a technique which makes slight modification to the learning algorithm such that model generalizes better.
4. 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 Tress or random forest.
5. Yes, decision tress 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 to 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. 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. 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. Out of sample is a technique to verify the performance of a boostrapping model without 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. 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 cros validation will create 5 subsets in both training ad testing dataset, train and predict are output 5 accuracy values. Averagin those values would gives us a grater idea of how good the model is.
10. In machine learning, hyperparameter optimization or tuning is the problem of choosing a set of optimal hyperparameters for a learning algorithm. A hyperparameter is a parameter whose value is used to control the learning process. By contrast,the values of other parameters (typically node weights) are learned. Hyperparameters are important because they directly control the behaviour of the training algorithm and have a significant impact on the performance of the model is being trained. Easy to manage a large set of experiments for hyperparameter tuning.
11. When the learning rate is too large, gradient descent can inadvertently increase rather than decrease the training error. When the learning rate is too small, training is not only slower, but may become permanently stuck with a high training error.

12. No because Logistic regression is considered a generalized linear model because the outcome always depends on the sum of the inputs and parameters. Or in other words, the output cannot depend on the product (or quotient, etc.) of its parameters.

13. Adaboost is more about 'voting weights' and Gradient boosting is more about 'adding gradient optimization'. Adaboost increases the accuracy by giving more weightage to the target which is misclassified by the model. At each iteration, Adaptive boosting algorithm changes the sample distribution by modifying the weights attached to each of the instances. It increases the weights of the wrongly predicted instances and decreases the ones of the correctly predicted instances.

Gradient boosting calculates the gradient (derivative) of the Loss Function with respect to the prediction (instead of the features). Gradient boosting increases the accuracy by minimizing the Loss Function (error which is difference of actual and predicted value) and having this loss as target for the next iteration.Gradient boosting algorithm builds first weak learner and calculates the Loss Function. It then builds a second learner to predict the loss after the first step. The step continues for third learner and then for fourth learner and so on until a certain threshold is reached.

14. Bias: Amount of error introduced by approximating real world phenomena with a simplified model. Variance: It shows how much your model's test error changes based on variation in the training data.Trade-off: It is tension between the error introduced by the bias and variance.

Bias-Variance Trade off-> It is the property of a set of predictive models where by models with a lower bias in parameter estimation have higher variance of the parameter estimates across samples and vice versa.

15. Linear kernel: A linear kernel is used as normal dot product of any two given observation. The product between two vectors is the sum of the multiplication of each pair of input values.It is mostly used when there are a Large number of features in a particular Data Set.Training a SVM with a Linear Kernel is faster than with any other Kernel.

k(x,x1) = sum(x*x1)

RBF(Radial Basis Function): In machine learning, the radial basis function kernel, or RBF kernel, is a popular kernel function used in various kernelized learning algorithms.RBF can map an input space in infinite dimensional space. In particular, it is commonly used in support vector machine classification. k(x,x1) = exp(-gamma*sum(x-x1^2)) Here gamma is a parameter which ranges from 0 to 1.

Polynomial: A polynomial kernel is a more generalized form of the linear kernel. The polynomial kernel can distinguish curved or non linear input space.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.

k(x,x1) = 1+sum(x*x1)^d where d = degree of polynomial