Worksheet set 5 Machine Learning

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- RSS is better measure of goodness 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-

   TSS is the sum of square of difference of each data point from the mean value of all the values of target variable (y). Here, the line is with intercept ('c' in y = mx + c) equal to Y mean; it means that this line does not include any influence of independent variable.

   Explained sum of squares is a statistical measure of deviation from the mean. It is also known as variation. It is calculated by adding together the squared differences of each data point. To determine the sum of squares, square the distance between each data point and the line of best fit, then add them together.

   The residual sum of squares (RSS) is a statistical technique used to measure the amount of variance in a data set that is not explained by a regression model itself. Instead, it estimates the variance in the residuals, or error term.

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

   Ans- Regularization refers to techniques that are used to calibrate machine learning models in order to minimize the adjusted loss function and prevent overfitting or underfitting. Using Regularization, we can fit our machine learning model appropriately on a given test set and hence reduce the errors in it.

4. What is Gini–impurity index?

   Ans- Gini Impurity is a measurement used to build Decision Trees to determine how the features of a dataset should split nodes to form the tree.

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

Ans- Yes, Out of all machine learning techniques, decision trees are amongst the most prone to overfitting. No practical implementation is possible without including approaches that mitigate this challenge. In this module, through various visualizations and investigations, you will investigate why decision trees suffer from significant overfitting problems. Using the principle of Occam's razor, you will mitigate overfitting by learning simpler trees. At first, you will design algorithms that stop the learning process before the decision trees become overly complex. In an optional segment, you will design a very practical approach that learns an overly-complex tree, and then simplifies it with pruning. Your implementation will investigate the effect of these techniques on mitigating overfitting on our real-world loan data set.

6. What is an ensemble technique in machine learning?
   Ans-
   An ensemble method is a technique which uses multiple independent similar or different models/weak learners to derive an output or make some predictions

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

   Ans-
   Bagging is a technique for reducing prediction variance by producing additional data for training from a dataset by combining repetitions with combinations to create multi-sets of the original data. Boosting is an iterative strategy for adjusting an observation's weight based on the previous classification.

8. What is out-of-bag error in random forests?
   Ans- Out-of-bag (OOB) error, also called out-of-bag estimate, is a method of measuring the prediction error of random forests, boosted decision trees, and other machine learning models utilizing bootstrap aggregating (bagging). Bagging uses subsampling with replacement to create training samples for the model to learn from.

9. What is K-fold cross-validation?
   Ans- Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation
10. What is hyper parameter tuning in machine learning and why it is done?
    Ans-
    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.

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

The learning rate can seen as step size, η. As such, gradient descent is taking successive steps in the direction of the minimum. If the step size η is too large, it can (plausibly) "jump over" the minima we are trying to reach, ie. we overshoot.

12. Can we use Logistic Regression for classification of Non-Linear Data? If not, why?
    Ans- Logistic regression is neither linear nor is it a classifier. The idea of a "decision boundary" has little to do with logistic regression, which is instead a direct probability estimation method that separates predictions from decision.
13. Differentiate between Adaboost and Gradient Boosting.
    Ans- AdaBoost is the first designed boosting algorithm with a particular loss function. On the other hand, Gradient Boosting is a generic algorithm that assists in searching the approximate solutions to the additive modelling problem. This makes Gradient Boosting more flexible than Ada Boost.

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

linear kernels-

It is the most basic type of kernel, usually one dimensional in nature. It proves to be the best function when there are lots of features. The linear kernel is mostly preferred for text-classification problems as most of these kinds of classification problems can be linearly separated.

Linear kernel functions are faster than other functions.

Linear Kernel Formula

F(x, xj) = sum(x.xj)

Here, x, xj represents the data you're trying to classify.

RBF-

it is one of the most preferred and used kernel functions in svm. It is usually chosen for non-linear data. It helps to make proper separation when there is no prior knowledge of data.

Gaussian Radial Basis Formula

$$F(x, xj) = \exp(-\text{gamma} * ||x - xj||^2)$$

The value of gamma varies from 0 to 1. You have to manually provide the value of gamma in the code. The most preferred value for gamma is 0.1.

Polynomial kernels-

It is a more generalized representation of the linear kernel. It is not as preferred as other kernel functions as it is less efficient and accurate.

Polynomial Kernel Formula

$$F(x, xj) = (x.xj+1)^d$$

Here '.' shows the dot product of both the values, and d denotes the degree.

F(x, xj) representing the decision boundary to separate the given classes.