 ASSIGNMENT - 5

MACHINE LEARNING

**Q1 to Q15 are subjective answer type questions, Answer them briefly.**

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 better measure of goodness of fit model in regression.

R-squared represents the proportion of the variance in your data. The residual sum of squares (RSS) is the sum of the squared distances between your actual versus predicted values, therefore the value of it depends largely on the **scale** of your response variable. Taken alone, RSS is not therefore as informative.

1. 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 : The Total SS (TSS or SST) tells you how much variation there is in the dependent variable.  
Total SS = Σ(Yi – mean of Y)2

The Explained SS tells you how much of the variation in the dependent variable your model explained.  
Explained SS = Σ(Y-Hat – mean of Y)2

Sum of squares is a measure of how a data set varies around a central number (like the mean).

*RSS =∑(yi​−y^​i​)2*

*where:yi​=Observed value*

*y^​i​=Value estimated by regression line​*

 total sum of squares (TSS) = explained sum of squares (ESS) + residual sum of squares (RSS)

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

Ans : We need regularization to minimize overfitting and underfitting.

1. What is Gini–impurity index?

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

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

Ans : Yes. Decision trees are prone to overfitting, especially when a tree is particularly deep. This is due to the amount of specificity we look at leading to smaller sample of events that meet the previous assumptions.

1. What is an ensemble technique in machine learning?

Ans: Ensemble methods are techniques that aim at improving the accuracy of results in models by combining multiple models instead of using a single model.

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

Ans: “bagging” algorithms that aim to reduce the complexity of models that overfit the training data. In contrast, boosting is an approach to increase the complexity of models that suffer from high bias, that is, models that underfit the training data

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

Ans : The out-of-bag error is the average error for each predicted outcome calculated using predictions from the trees that do not contain that data point in their respective bootstrap sample. This way, the Random Forest model is constantly being validated while being trained

1. What is K-fold cross-validation?

Ans: K-fold Cross-Validation is when the dataset is split into a K number of folds and is used to evaluate the model's ability when given new data.

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

Ans: Hyperparameter tuning consists of finding a set of optimal hyperparameter values for a learning algorithm while applying this optimized algorithm to any data set.

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

Ans: A learning rate that is too large can cause the model to converge too quickly to a suboptimal solution

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

Ans: No.

Logistic regression is neither linear nor is it a classifier.

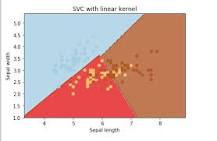
1. Differentiate between Adaboost and Gradient Boosting.

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.

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

“Bias and variance are complements of each other” The increase of one will result in the decrease of the other and vice versa. Hence, **finding the right balance of values** is known as the Bias-Variance Tradeoff. An ideal algorithm should neither underfit nor overfit the data.

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

linear kernels :

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.

RBF Kernel is popular because of its similarity to K-Nearest Neighborhood Algorithm. It has the advantages of K-NN and overcomes the space complexity problem as RBF Kernel Support Vector Machines just needs to store the support vectors during training and not the entire dataset.

Polynomial Kernel:It represents the similarity of vectors in the training set of data in a feature space over polynomials of the original variables used in the kernel

