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?**

   Goodness of fit model can be explained as closeness between the observed data point and predicted data points. If, the distance between fitted line created by predictions and observed data points are very less, it is said that model closely fitted and able to explain all possible variability to the response data around its mean.

   R-squared is equal to variability explained by the **regression**, divided by total variability. That is, Sum of total squares due to regression (SSR) divided by sum of total squares calculated from mean line (SST). It is a relative measure and takes values ranging from 0 to 1.

   An **R-squared** of zero means our regression line explains none of the variability of the data. An **R-squared** of 1 would mean our model explains the entire variability of the data.

   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. It is mare a valuee shown in above figure with **SSE** , but it has significance in **R-squared** test.

$$R^2 = \frac{\text{Sum of Squares due to Regression (SSR)}}{\text{Total sum of Squares (TSS)}} = \frac{(\text{TSS}) - (\text{SSE})}{(\text{TSS})} = \frac{(\text{TSS}) - (\text{RSS})}{(\text{TSS})}$$

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?**

TSS (Total Sum of Squares):

The sum of squares total, denoted as SST, is the squared differences between the observed dependent variable and its mean. It is also known as TSS or total sum of squares.

ESS (Explained Sum of Squares):

It is the sum of the differences between the predicted value and the mean of the dependent variable. It is also known as is ESS or explained sum of squares.

RSS (Residual Sum of Squares) :

The error is the difference between the observed value and the predicted value. It is also known as RSS or residual sum of squares. Residual as in: remaining or unexplained.

$$R^2 = \frac{\text{Sum of Squares due to Regression (SSR)}}{\text{Total sum of Squares (TSS)}} = \frac{(\text{TSS}) - (\text{SSE})}{(\text{TSS})} = \frac{(\text{TSS}) - (\text{RSS})}{(\text{TSS})}$$

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

The model will have a low accuracy if it is over fitting.This happenses due to high variance and high bias. Regularization will significantly reduce the variance of the model, without substantial increase in its bias.

4. **What is Gini–impurity index?**
   Gini Index, also known as Gini impurity, calculates the amount of probability of a specific feature that is classified incorrectly when selected randomly. If all the elements are linked with a single class then it can be called pure.

5. **Are unregularized decision-trees prone to overfitting? If yes, why?**
   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. This small sample could lead to unsound conclusions.

6. **What is an ensemble technique in machine learning?**
   Ensemble methods are techniques that aim at improving the accuracy of results in models by combining multiple models instead of using a single model. The combined models increase the accuracy of the results significantly. The goal of any machine learning problem is to find a single model that will best predict our wanted outcome. Rather than making one model and hoping this model is the best/most accurate predictor we can make, ensemble methods take a myriad of models into account, and average those models to produce one final model.

7. **What is the difference between Bagging and Boosting techniques?**
   Bagging, stands for bootstrap aggregating, is mainly applied in classification and regression. It increases the accuracy of models through decision trees, which reduces variance to a large extent. Whereas, Boosting is an ensemble technique that learns from previous predictor mistakes to make better predictions in the future. The technique combines several weak base learners to form one strong learner, thus significantly improving the predictability of models.

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

9. **What is K-fold cross-validation?**

   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. K refers to the number of groups the data sample is split into.

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

    Parameters which define the model architecture are referred to as hyper parameters and thus this process of searching for the ideal model architecture is referred to as hyper parameter tuning. Hyper parameter tuning is an essential part of controlling the behavior of a machine learning model. If we don't correctly tune our hyper parameters, our estimated model parameters produce suboptimal results, as they don't minimize the loss function. This means our model makes more errors. In order to reduce these errors, tuning is required.

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

    A learning rate that is too large can cause the model to converge too quickly to a suboptimal solution. Learning rate (λ) is one such hyper-parameter that defines the adjustment in the weights of our network with respect to the loss gradient descent. It determines how fast or slow we will move towards the optimal weights. If the learning rate is very large we will skip the optimal solution.

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

    Logistic regression is known and used as a linear classifier and cannot be used for classification of nonlinear data. Because it use to come up with a hyper *plane* in feature space to separate observations that belong to a class from all the other observations that do *not* belong to that class. The decision boundary is thus *linear*.

13. **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 modeling problem. This makes Gradient Boosting more flexible than AdaBoost.

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

There is a tradeoff between a model's ability to minimize bias and variance. Variance is the variability of model prediction for a given data point or a value which tells us spread of our data. Model with high variance pays a lot of attention to training data and does not generalize on the data which it hasn't seen before. If our model is too simple and has very few parameters then it may have high bias and low variance. On the other hand if our model has large number of parameters then it's going to have high variance and low bias. So we need to find the right/good balance without over fitting and under fitting the data.

This tradeoff in complexity is why there is a tradeoff between bias and variance. An algorithm can't be more complex and less complex at the same time.

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

The linear, polynomial and RBF or Gaussian kernel are simply different in case of making the hyper plane decision boundary between the classes.

The kernel functions are used to map the original dataset (linear/nonlinear) into a higher dimensional space with view to making it linear dataset.

A usually linear and polynomial kernel is less time consuming and provides less accuracy than the RBF or Gaussian kernels.