Machine Learning Assignment

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?

R-Squared is a better measure of goodness of fit in Regression models, because when there are multiple variables in the model, it would allow us to compare models with differing numbers of independent variables, it is useful for understanding the overall proportion of variance explained by the model, whereas RSS provides an absolute measure of fit but is not as interpretable or comparable across different models.

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.

The Total SS tells you how much variation there is in the dependent variable. It is the sum of squared differences between the observed dependent variables and the overall mean.

Total SS = $\Sigma(Yi - \text{mean of } Y)^2$

The Explained SS tells you how much of the variation in the dependent variable your model explained. It is the sum of the differences between the predicted value and the mean of the dependant variable. In other words, it describes how well our line fits the data.

Explained SS = $\Sigma(Y\text{-Hat} - \text{mean of } Y)^2$

The residual sum of squares RSS 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 and the predicted. Residual Sum of Squares = Σ e2

The relationship between them is TSS= ESS+RSS.

The total variability of the dataset is equal to the variability explained by the regression line plus the unexplained variability, known as error.

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

Regularization is a technique used to prevent overfitting or underfitting and improve the performance of models. Regularization helps control model complexity; it can help balance the trade-off between model bias and variance which leads to improved performance.

It also helps in feature selection when it drives out some feature coefficients to zero. This automatically selects important feature while excluding the less important ones. When features are highly correlated, regularization can stabilize the model by reducing coefficient sensitivity to small data changes.

Regularisation models learn underlying patterns of data for better understanding of new data.

4. What is Gini–impurity index?

The Gini impurity measure is one of the methods used in decision tree algorithms to decide the optimal split from a root node, and subsequent splits. Gini Impurity tells us what the probability is of misclassifying an observation. Low the Gini index the better the split. The range of the Gini index is [0, 1], where 0 indicates perfect purity and 1 indicates maximum impurity.

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

Yes, unregularized Decision tree models are prone to overfitting. In decision trees, In order to fit the data (including outliers), the model keeps generating new nodes and ultimately the tree becomes too complex to interpret. The decision tree predicts well for the training data but can be inaccurate for new data. If a decision tree model is allowed to train to its full potential, it can overfit the training data.

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 several base models instead of using a single model to produce one optimal predicted model. The combined models increase the accuracy of the results significantly. This has increased the usage of ensemble methods in machine learning.

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

In Bagging each model receives equal weight whereas in Boosting models are weighted according to their performance. In Bagging each model is built independently whereas in Boosting new models are influenced by performance of previously built models. Bagging model aims to reduce variance whereas boosting model aims to reduce bias.

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

Out-of-Bag Error is a concept used in ensemble machine learning algorithms such as random forests. When building a random forest model, each tree is trained using a subset of the original data, known as the bootstrap sample. During the training process, some observations are left out or out of bag (OOB) for each tree.

The OOB observations that were not used in the training of a particular tree can be considered as a testing or validation set for that tree. The model's prediction accuracy on the OOB observations can then be calculated and averaged across all the trees to obtain the OOB Error. Out-of-Bag Error works by utilizing the OOB observations to estimate the model's performance on unseen data.

9. What is K-fold cross-validation?

K-fold cross-validation is a technique for evaluating predictive models. The dataset is divided into k subsets or folds. The model is trained and evaluated k times, using a different fold as the validation set each time. Performance metrics from each fold are averaged to estimate the model’s generalization performance. This method aids in model assessment, selection and hyperparameter tuning providing a more reliable measure of a model’s effectiveness.

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

Hyperparameter tuning is the process of selecting the optimal values for a machine learning model’s hyperparameters. Hyperparameters are settings that control the learning process of the model, such as the learning rate or the kernel size in a support vector machine. The goal of hyperparameter tuning is to find the values that lead to the best performance on a given task.

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

The learning rate is an important hyperparameter that greatly affects the performance of gradient descent. It determines how quickly or slowly our model learns, and it plays an important role in controlling both convergence and divergence of the algorithm. When the learning rate is too large, gradient descent can suffer from divergence. This means that weights increase exponentially, resulting in exploding gradients which can cause problems such as instabilities and overly high loss values.

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

Logistics regression is a type of linear model that predicts the probability of a binary outsome such as yes or no, true or false or 0 or 1. Logistics regression assumes a linear relationship between the features and the label. This means that it cannot capture the complexity and non-linearity of the data. Another point is that it is too sensitive to outliers, which can affect the accuracy and stability of the model. It has a limited capacity to learn from multiple features as it can only combine them linearly, so logistics regression is not suited for classification of nonlinear data.

13. Differentiate between Adaboost and Gradient Boosting.

Adaboost and gradient boosting are both ensemble learning techniques that combine multiple weak learners to create a stronger model, but they differ in their approach to build the ensemble and in updating the weights of the features in the dataset.

In Adaboost, the weights of the samples are adjusted at each iteration whereas no reweighting of the samples are done in gradient boosting.

The final model in Adaboost is formed by combining the predictions from individual trees through a weight sum whereas the final model in gradient boost is an equal weighted sum of all the individual trees.

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

The bias-variance trade-off describes the relationship between a model's complexity, the accuracy of its prediction and how well it can make predictions based on previously unseen data that were not used to train the model. The bias-variance trade off is the delicate equilibrium between underfitting and overfitting model.

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

Support Vector Machine are a set of supervised learning methods used for classification, regression and outliers' detection. Some popular kernel functions used in Kernel SVR are the radial basis function (RBF) kernel, polynomial kernel, and sigmoid kernel. The choice between Linear SVR and Kernel SVR depends on the characteristics of the data and the desired level of complexity in the model.

Linear Kernels: A linear kernel can be used as normal dot product any two given observations. The product between two vectors is the sum of the multiplication of each pair of input values.

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

Radial Basis Function Kernel: RBF is more complex and efficient at the same time that it can combine multiple polynomial kernels multiple times of different degrees to project the non-linearly separable data into higher dimensional space so that it can be separable using a hyperplane. The RBF kernel works by mapping the data into a high-dimensional space by finding the dot products and squares of all the features in the dataset and then performing the classification using the basic idea of Linear SVM.