MACHINE LEARNING WORKSHEET 5

1. R-squared is a better measure of goodness of fit model in regression. This statistic indicates the percentage of the variance in the dependent variable that the independent variables explain collectively. R-squared measures the strength of the relationship between the model and the dependent variable on a convenient 0 – 100% scale. The higher the R-squared, the better the model fits our data.

2. TSS- The total sum of squares measures how much variation there is in the observed data

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

ESS- The Explained Sum of Squares tells us how much of the variation in the dependent variable your model explained

Explained SS = Σ(Y-Hat – mean of Y)2.

RSS- The residual sum of squares measures the variation in the error between the observed data and modelled values

$RSS = \sum^{n}_{i=1} (y^i - f(x_i))^2$

3. Regularisation refers to techniques that are used to calibrate machine learning models in order to minimise 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.

   In the context of machine learning, regularization is the process which regularizes or shrinks the coefficients towards zero. In simple words, regularization discourages learning a more complex or flexible model, to prevent overfitting

4. 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.
   It helps us to determine which splitter is best so that we can build a pure decision tree. Gini impurity ranges values from 0 to 0.5.

5. 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.

MACHINE LEARNING WORKSHEET 5

6. An ensemble method is a technique which uses multiple independent similar or different models/weak learners to derive an output or make some predictions. Ensemble methods usually produces more accurate solutions than a single model would.

7. • The Bagging technique is a simple way of combining predictions of the same kind, whereas boosting combines predictions that belong to different types.
   • In Bagging, each model is created independent of the other, But in boosting new models, the results of the previously built models are affected.
   • Bagging gives equal weight to each model, whereas in Boosting technique, the new models are weighted based on their results.
   • In boosting, new subsets of data used for training contain observations that the previous model misclassified. Bagging uses randomly generated training data subsets.
   • Bagging tends to decrease variance, not bias. In contrast, Boosting reduces bias, not variance.
   • The bagging technique tries to resolve the issue of overfitting training data, whereas Boosting tries to reduce the problem of Bias.
   • In Bagging, training data subsets are drawn randomly with a replacement for the training dataset. In Boosting, every new subset comprises the elements that were misclassified by previous models.
   • Bagging is usually applied where the classifier is unstable and has a high variance. Boosting is usually applied where the classifier is stable and simple and has high bias.

8. The out-of-bag (OOB) error is the average error for each calculated using predictions from the trees that do not contain in their respective bootstrap sample. This allows the RandomForestClassifier to be fit and validated whilst being trained

9. K-fold cross-validation is defined as a method for estimating the performance of a model on unseen data. This technique is recommended to be used when the data is scarce and there is an ask to get a good estimate of training and generalization error thereby understanding the aspects such as underfitting and overfitting. This technique is used for hyperparameter tuning such that the model with the most optimal value of hyperparameters can be trained. It is a resampling technique without replacement. The advantage of this approach is that each example is used for training and validation (as part of a test fold) exactly once. This yields a lower-variance estimate of the model performance than the holdout method.

MACHINE LEARNING WORKSHEET 5

As mentioned earlier, this technique is used because it helps to avoid overfitting, which can occur when a model is trained using all of the data. By using k-fold cross-validation, we are able to “test” the model on k different data sets, which helps to ensure that the model is generalizable.

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. That combination of hyperparameters maximizes the model's performance, minimizing a predefined loss function to produce better results with fewer errors.

11. What issues can occur if we have a large learning rate in Gradient Descent?
This is the hyperparameter that determines the steps the gradient descent algorithm takes. Gradient Descent is too sensitive to the learning rate. If it is too big, the algorithm may bypass the local minimum and overshoot and can cause the model to converge too quickly to a suboptimal solution and due to this it can inadvertently increase the training error.

12. Can we use Logistic Regression for classification of Non-Linear Data? If not, why?
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.). Non-Linear Classification refers to categorizing those instances that are not linearly separable, i.e. It is not easy to classify data with a straight line.

13. Differentiate between Adaboost and Gradient Boosting. Gradient boosting and Adaboost are ensemble methods applied in machine learning(ML) modeling to improve the effectiveness of weak learners to increase algorithm performance. Both use a combination of weak learners to predict a target variable. However, they do that differently.
While gradient boosting trains the learners and reduces the weak learners' loss functions by training the model's residues, Ada boost focuses on training the prior miscalculated observations and alters the data distribution to improve sample

weight values. And while boosting algorithms are less prone to overfitting, they can get complex and overfit the training data.

14. What is bias-variance trade off in machine learning?
It helps optimize the error in our model and keeps it as low as possible. An optimized model will be sensitive to the patterns in our data, but at the same time will be able to generalize to new data. In this, both the bias and variance should be low so as to prevent overfitting and underfitting.

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