MACHINE LEARNING

1. R-squared, is a better measure of goodness of fit model in regression. It is also known as the coefficient of determination and statistical measure used to assess how well a regression model explains the variation in the dependent variable based on the independent variables. R-squared gives a measure of how predictive the regression is and how much variation is explained by the regression.
2. TSS- The total sum of squares is the sum of the squared deviations of each data point from the mean of the entire data set. The higher the TSS, the greater the variation in the response variable

ESS- The explained sum of squares measures the variation in the response variable that is explained by the predictors in the model .The higher the ESS, the better the fit of the model.

RSS- The residual sum of squares measures the difference between the actual value of the response variable and the predicted value of the response variable, the lower the RSS, the better the fit of the model.

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

3. Regularization is a critical aspect of Machine Learning models, ensuring they don't succumb to over fitting or under fitting. Essentially, it introduces a penalty term to the loss function, preventing the model from becoming too complex. Regularization techniques, such as L1 and L2 regularization, play a vital role in achieving a balanced and efficient machine learning model.

4. The Gini index is a measure of impurity in a set of data. It is calculated by summing the squared probabilities of each class. A lower Gini index indicates a more pure set of data

5. Yes, un-regularized decision –tree prone to over-fitting because, over-fitting occurs when the tree is designed so as to perfectly fit all samples in the training data set. Thus it ends up with branches with strict rules of sparse data. Thus this effects the accuracy when predicting samples that are not part of the training set.

6. Ensemble technique in machine learning helps improve machine learning results by combining several models. This approach allows the production of better predictive performance compared to a single model.

7. The bagging technique tries to resolve the issue of over-fitting training data, whereas Boosting tries to reduce the problem of Bias.

- 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 tends to decrease variance, not bias. In contrast, Boosting reduce Bagging tends to decrease variance, not bias.

8. Out-of-Bag Error, also known as OOB 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.

9. 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.

10. Hyper-parameter tuning is the process of selecting the optimal set of hyper-parameters for a machine learning model. It is an important step in the model development process, as the choice of hyper-parameters can have a significant impact on the model's performance. It is done because it involves searching for the optimal combination of hyper-parameters within a predefined set of possible values.

11. If we have large learning rate in gradient descent, the machine will not learning anything. If the learning rate is too high, the algorithm may overshoot the minimum, and if it is too low, the algorithm may take too long to converge.

12. No we cannot use Logistics Regression for classification of non linear data because logistic regression is a linear classifier and perfectly separates only linearly separable classes. Non-linear data could imply elements not associated with class separation, suggesting different sources of non-linearity.

13. (a) The technique of Boosting uses various loss functions. In case of Adaptive Boosting or AdaBoost, it minimises the exponential loss function that can make the algorithm sensitive to the outliers. With Gradient Boosting, any differentiable loss function can be utilised. Gradient Boosting algorithm is more robust to outliers than AdaBoost.

(b) 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.

14. In statistics and machine learning, the bias–variance trade-off describes the relationship between a model's complexity, the accuracy of its predictions, and how well it can make predictions on previously unseen data that were not used to train the model.

15. In machine learning, the polynomial kernel is a kernel functions commonly used with support vector machines (SVMs) and other kernelized models, that represents the similarity of vectors (training samples) in a feature space over polynomials of the original variables, allowing learning of non-linear models.

RBF (radial basis function) kernel is used when the boundaries are hypothesized to be curve-shaped.

RBF kernel uses two main parameters, gamma and C that are related to:

1. the decision region (how spread the region is), and
2. the penalty for misclassifying a data point.

A linear kernel is suitable for separable datasets

STATISTICS WORKSHEET 5

1. (d) expected

2. (c) frequencies

3. (c) 6

4. (b) Chi-square distribution

5. (c) F- distribution

6.(b) Hypothesis

7. (a) Null hypothesis

8. (a) Two – tailed

9. (b) Research Hypothesis

10. (a) np