Statistics Worksheet 5

Q1) D → Expected

Q2) C → Frequencies

Q3) C → 6

Q4) B → Chi-Square

Q5) C → F Distribution

Q6) B → Hypothesis

Q7) A → Null Hypothesis

Q8) A → 2 Tailed

Q9) B → Research Hypothesis

Q10) A → np

Machine Learning Assignment 5

Q1) R Squared is a measure of RSS with TSS. RSS estimates the variance in the residuals or error terms. Since R Squared takes into account RSS and TSS hence R Squared is a better measure of goodness of fit model in regression.

Q2) TSS = Tells how much is the variation in the dependent variables

ESS = Tell how much of the variation in dependent variables is explained by the model

RSS = Tells how much of the dependent variables variation your model did not explain

Relation between TSS, ESS and RSS is TSS = ESS + RSS

Q3) Regularization helps in reducing multicollinearity and used to calibrate ML Model in order to minimize adjusted loss function and prevent under and over fitting.

Q4) Gini Impurity indicates likelihood of the new random data being misclassified if it was given a random class label according to class distribution in the dataset. Its value is from 0 - 0.5

Q5) Yes, decision tree is prone to overfitting. Decision trees can learn a training set to a point of high granularity that makes them prone to overfitting if the decision tree is allowed to split to a granular degree.

Q6) Ensemble learning technique is machine learning technique which takes the help of several models and combines their output to produce an optimized model. This helps in improving the overall performance of the model

Q7)

- Bagging is a method of merging the same type of predictions whereas boosting is a method of merging different type of predictions.
- Bagging decreases variance and solves over-fitting issues whereas boosting decreases bias and not variance.

- Bagging is applied where the classifier is unstable and has a high variance whereas boosting is applied where the classifier is stable and has high bias.
- In Bagging each model receives equal weight whereas in boosting models are weighted based on their performance.
- Models are built independently in bagging whereas New Models are affected by a previously built model's performance in boosting.

Q8) 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.

Q9) K Fold Cross Validation is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the no of groups that a given data sample is to split into.

Q10) Hyperparameter Tuning is the process of picking a collection of optimal parameters for a model learning algorithm. Hyper Parameter is also called model predictor since its value is used at the starting point for model learning algo. Hyperparamtric function is just a mathematical tool and not a specification or a rule to solve any design problem. This means that even the most optimized hyper parameter in a given training set will not produce desired accuracy. This problem is addressed by tuning the hyper parameter. After each tuning the model can make corrections to hyper parameter in order to bring back the model with an accurate prediction.

Q11) Gradient Descend takes successive steps in the direction of minimum. If the learning rate is high it can jump over the minima we are trying to reach, i.e. overshooting. This can lead to osculation around the minimum or in some cases outright divergence.

Q12) No Logistic Regression cannot be used for Classification of Non Linear Data as it can only handle binary classification.

Q13)

- Adaboost minimises the exponential loss function which makes algo sensitive to outliers whereas in case of gradient boosting any differentiable loss function can be utilized and is more robust to outliers.
- Gradient Boosting is more flexible than adaboost as the former is generic algorithm that assist in searching approx. solutions to modelling problem.
- Adaboost minimizes loss function related to any classification error and best used with weak learners and majorly used for Binary classification problem. Gradient Boosting is used to solve differential loss function problem and used for both classification and regression.

Q14) Bias is the difference between our actual and predicted value. Bias is simple assumptions that our model makes about our data to be able to predict new data. If bias is high our model has not learned the data properly and hence will not perform well on testing data due to underfitting.

If variance is high our model will capture all features of the data including the un important and tune itself to the data and work well on test data as well. But when new data is given it cannot predict on it due to overfitting.

Bias Variance Tradeoff for any model is to find a perfect balance between bias and variance. This ensures that we capture the essential patterns in the model while ignoring the noise present in it. This helps to optimize the error in the model and keeps it as low as possible.

Q15) Linear SVM: Is used for linearly separable data i.e. if the data can be classified into 2 classes by using a straight line such data is termed as linearly separable and classifier used is Linear SVM Classifier.

Polynomial SVM: Used with support vector machines and other kernelized models that represents similarity of vectors in a feature space over polynomials of the original variables allowing learning of non linear models. It non only on given features of input samples to determine their similarity but also combination of these