STATISTICS WORKSHEET-5

1) D) Expected
2) D) All of these
3) C) 6
4) B) Chi square distribution
5) A) Binomial Distribution
6) B) Hypothesis
7) A) Null Hypothesis
8) A) Two tailed
9) C) Research Hypothesis
10) A) np

MACHINE LEARNING

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?

Ans: Between these two R-squared measures the goodness of fit model in regression, because it determines the proportion of variance in dependent variable that can be explained by the independent variable.It ranges fro 0 to 1 higher the R-squared value better the goodness of fit model.

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.

Ans: TSS is the sum of the squared differences between each observed value of the dependent variable and the mean of the dependent variable.

ESS is the sum of the squares of the deviations of the predicted values from the mean value of a response variable.

RSS measures the level of variance in the error term or residuals of regression model.

ESS=TSS-RSS

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

Ans: While training a machine learning model , the model can be easily under fitted or over fitted to avoid this we use regularization in machine learning to fit a model onto our test set. It help us to get an optimal model.

4) What is Gini–impurity index?

Ans: Gini- impurity is measure of how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset.

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

Ans: Yes unregularized decision –trees prone to overfitting due to their high capacity to learn complex relationships in the data,they are sensitive to small changes in the training data,and they can grow deep trees that can capture intrivate pattern in training data including noise and outliers.

6) What is an ensemble technique in machine learning?

Ans: Ensemble technique is a machine learning paradigm where multiple models are trained to solve the same problem and combined to get better results.

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

Ans: In Bagging models are built independently. Training data subsets are drawn randomly with a replacement for the training dataset.

In boosting new models are affected by previously built model's performance.Every new subset comprises the elements that were misclassified by previous models.

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

Ans: Out-bag error is method of estimating performance of machine learning models . In ensemble methods like random forest trees are trained by ensemble bagging method at that time some data points are left out by training set for each individual tree. For each point you can calculate oob error.

9) What is K-fold cross-validation?

Ans. It is process used to access the performance of models. In this the set is devided into k equal folds and k times it is trained and validated and average of its performance is used to generalized the predicted model performance.

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

Ans: Hyper parameter is parameter whose value can influence the function, structure, performance of the model. The process of selecting these parameters is called hyper parameter tuning.

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

Ans: Choosing learning rate will effect performance of Gradient Descent. If we have large learning rate the algorithm may overshoot to minimum, instability, poor generalization, slow convergence will occur.

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

Ans: As we know Logistic Regression are used on linear data which interact with binary output, we can use with non linear data by applying some tecniques like feature engineering ect, but we will not get good accuracy.

13) Differentiate between Adaboost and Gradient Boosting

Ans: Adaboost ,shift was done by up weighting observations which are misclassified before.

It is more sensitive to noise and outlier.

It can be parallelized.

Gradient boost it creates builds model by combining previous weak learners.

It is less sensitive to noise and outlier as compared to adaboost.

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

Ans: It is fundamental concept in machine learning that deals with model performance and

Complexity.It refers to the trade of between variance and bias of model.

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

Ans: Linear is used with linearly separable data or which has high features. It does the dot product within the input features.

RBF(Radial basis function) is effective capturing complex and linearly non separable data.It maps the data into high dimensional form using gaussion function. Tuning of parameter is required here.

Polynomial is effective capturing nonlinear separable data. It does the dot product within input features by raising to certain power. It maps data into high dimensional using polynomial function. Tuning of parameter is required here.