Assignment -5

Ques 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 and Residual Sum of Squares (RSS) are both measures of the goodness of fit model in regression. R-squared is useful for comparing different models or for determining the proportion of the variability in the dependent variable that is explained by the model.

Ques 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 sum of squares (TSS) measures how much variation there is in the observed data.

Explained sum of squares (ESS) is the sum of the differences between the predicted value and the mean of the dependent variable.

The residual sum of squares measures the variation in the error between the observed data and modeled values.

TSS = ESS + RSS

Ques 3: What is the need of regularization in machine learning?

While training a machine learning model, the model can easily be overfitted or under fitted. To avoid this, we use regularization in machine learning to properly fit a model onto our test set. Regularization techniques help reduce the chance of overfitting and help us get an optimal model.

Ques 4: What is Gini–impurity index?

Gini impurity is a measure used in decision tree algorithms to quantify a dataset's impurity level or disorder.

Ques5: Are unregularized decision-trees prone to overfitting? If yes, why?

Decision trees are a popular and powerful method for data mining, as they can handle both numerical and categorical data, and can easily interpret the results. However, decision trees can also suffer from overfitting, which means that they learn too much from the training data and fail to generalize well to new data.

Ques 6: What is an ensemble technique in machine learning?

Ensemble methods in machine learning are **techniques that create multiple models and then combine them to produce improved results.**

Ques 7:What is the difference between Bagging and Boosting techniques?

Bagging is the simplest way of combining predictions that belong to the same type while Boosting is a way of combining predictions that belong to the different types. Bagging aims to decrease variance, not bias while Boosting aims to decrease bias, not variance.

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

Out-of-bag (OOB) error, also called out-of-bag estimate, is a method of measuring the prediction error of random forests, boosted decision trees, and other machine learning models utilizing bootstrap aggregating.

Ques 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.

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

When you are training machine learning models, each dataset and model needs a different set of hyperparameters, which are a kind of variable. The only way to determine these is through multiple experiments, where you pick a set of hyperparameters and run them through your model. This is called hyperparameter tuning.

It is done because it allows data scientists to tweak model performance for optimal results. This process is an essential part of machine learning, and choosing appropriate hyperparameter values is crucial for success

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

The choice of learning rate can significantly impact the performance of gradient descent. If the learning rate is too high, the algorithm may overshoot the minimum.

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

No, we cannot use Logistic Regression for classification o Non linear Data because Logistic Regression has traditionally been used as a linear classifier, i.e. when the classes can be separated in the feature space by linear boundaries. That can be remedied however if we happen to have a better idea as to the shape of the decision boundary.

Ques 13: Differentiate between Adaboost and Gradient Boosting.

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. This makes Gradient Boosting more flexible than AdaBoost.

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

Finding the right balance of bias and variance is key to creating an effective and accurate model. This is called the bias-variance tradeoff.

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

- **Linear :** used when data is linearly separable.
  - **Polynomial Kernel:** It represents the similarity of vectors in the training set of data in a feature space over polynomials of the original variables used in the kernel.
- RBF :In machine learning, the radial basis function kernel, or RBF kernel, is a popular kernel function used in various kernelized learning algorithms. In particular, it is commonly used in support vector machine classification.

STATISTICS WORKSHEET-5

1 . Using a goodness of fit,we can assess whether a set of obtained frequencies differ from a set of frequencies.

Ans : Expected

2. Chisquare is used to analyse

Ans : All od these

3.What is the mean of a Chi Square distribution with 6 degrees of freedom?

Ans : 6

4. Which of these distributions is used for a goodness of fit testing?

Ans : Chisquared distribution

5.Which of the following distributions is Continuous

Ans :F Distribution

6. A statement made about a population for testing purpose is called?

Ans: Teststatistic

7. If the assumed hypothesis is tested for rejection considering it to be true is called?

Ans : Null Hypothesis

8. If the Critical region is evenly distributed then the test is referred as?

Ans : Two Tailed

9. Alternative Hypothesis is also called as?

Ans : Research Hypothesis

10. In a Binomial Distribution, if ‘n’ is the number of trials and ‘p’ is the probability of success, then the mean value is given by

Ans: np