**MACHINE LEARNING**

ASSIGNMENT - 5

Q1 to Q15 are subjective answer type questions, Answer them briefly.

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: R-squared is preferred when evaluating the goodness of fit of a regression model because it provides a more holistic assessment of how well the model explains the dependent variable. RSS, on the other hand, is useful for understanding the accuracy of individual predictions and the residual errors of the 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 measures the total variation in the dependent variable (Y). It represents the sum of the squared differences between each observed dependent variable value and the mean of the dependent variable.
* ESS measures the variation in the dependent variable that is explained by the independent variables (X).
* RSS measures the variation in the dependent variable that is not explained by the independent variables, also known as the residuals or errors of the mode

Relationship among TSS, ESS, and RSS:

**TSS = ESS + RSS**

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

ANS: Regularization techniques play a crucial role in machine learning by balancing model complexity and improving generalization performance. They are essential tools for building models that are robust, interpretable, and capable of making accurate predictions on new, unseen data.

4. What is Gini–impurity index?

ANS : The Gini impurity index is a crucial metric used in decision tree algorithms and other classifiers to determine optimal splits that minimize impurity and improve the classification accuracy of the model.

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

ANS : Yes, unregularized decision trees are prone to overfitting because while decision trees are powerful models for learning from data, they are prone to overfitting when left unregularized.

6. What is an ensemble technique in machine learning?

ANS : Ensemble techniques in machine learning harness the power of combining multiple models to enhance predictive accuracy, improve robustness, and ensure reliable performance across various domains and datasets.

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

ANS: Bagging and Boosting are ensemble techniques that improve predictive performance, they differ fundamentally in how they generate and combine base models, as well as in their approach to training and correcting errors. Each technique has its strengths and is suited for different types of machine learning problems and datasets.

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

ANS: The out-of-bag (OOB) error is an estimate of the model's performance on unseen data. It is a method used to validate the Random Forest model without the need for an additional validation set or cross-validation.

9. What is K-fold cross-validation?

ANS: K-fold cross-validation is a technique used to evaluate the performance of a machine learning model by splitting the data into K subsets (folds) of approximately equal size. The model is then trained and evaluated K times, using a different fold for evaluation each time while the remaining K-1 folds are used for training.

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

ANS: Hyperparameter tuning is essential in machine learning because it ensures that models perform optimally on unseen data, generalize well, and meet the specific requirements of the problem domain. By selecting the right hyperparameters through systematic tuning methods, practitioners can build models that achieve higher accuracy, reliability, and efficiency in real-world applications.

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

ANS: Having a large learning rate in Gradient Descent can lead to several issues, primarily related to the optimization process and the convergence of the model.

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

ANS: Logistic Regression is not suitable for classification of non-linear data due to its inherent limitation of assuming a linear decision boundary.

13. Differentiate between Adaboost and Gradient Boosting.

ANS:

 **Approach**: Adaboost focuses on correcting misclassifications by adjusting instance weights, while Gradient Boosting optimizes an ensemble of trees by minimizing the gradient of the loss function.

 **Model Type**: Adaboost typically uses shallow decision trees as weak learners, whereas Gradient Boosting often uses decision trees but can incorporate other types of models.

 **Training Strategy**: Adaboost trains models sequentially by adjusting weights, while Gradient Boosting trains models sequentially by fitting to residual errors.

 **Predictions**: Adaboost combines predictions using weighted voting, whereas Gradient Boosting sums predictions from all trees.

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

ANS: The bias-variance trade-off is crucial in machine learning because it guides the selection of appropriate models and hyperparameters to achieve the best possible predictive performance. It emphasizes the need to balance model complexity to avoid underfitting and overfitting, thereby ensuring that the model generalizes well to new, unseen data.

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

ANS:

* The linear kernel is the simplest kernel function used in SVMs.
* The RBF kernel is a popular choice for SVMs due to its ability to handle non-linear decision boundaries.
* The polynomial kernel computes the dot product of feature vectors in a higher-dimensional space using polynomial functions.

**STATISTICS WORKSHEET-5**

Q1 to Q10 are MCQs with only one correct answer. Choose the correct option.

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

a) Mean

b) Actual

c) Predicted

d) Expected

ANS : d) Expected

2. Chisquare is used to analyse

a) Score

b) Rank

c) Frequencies

d) All of these

ANS: c) Frequencies

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

a) 4

b) 12

c) 6

d) 8

ANS: c) 6

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

a) Normal distribution

b) Chisqared distribution

c) Gamma distribution

d) Poission distribution

ANS: b) Chisqared distribution

5. Which of the following distributions is Continuous

a) Binomial Distribution

b) Hypergeometric Distribution

c) F Distribution

d) Poisson Distribution

ANS: c) F Distribution

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

a) Statistic

b) Hypothesis

c) Level of Significance

d) TestStatistic

ANS: b) Hypothesis

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

a) Null Hypothesis

b) Statistical Hypothesis

c) Simple Hypothesis

d) Composite Hypothesis

ANS: a) Null Hypothesis

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

a) Two tailed

b) One tailed

c) Three tailed

d) Zero tailed

ANS: a) Two tailed

9. Alternative Hypothesis is also called as?

a) Composite hypothesis

b) Research Hypothesis

c) Simple Hypothesis

d) Null Hypothesis

ANS: b) 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

a) np

b) n

ANS: a) np