WORKSHEET

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

Answer - c) Predicted

2. Chisquare is used to analyse

a) Score

b) Rank

c) Frequencies

d) All of these

Answer - 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

Answer - 6 The mean of a Chi-Square distribution is equal to its degrees of freedom.

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

Answer – b) Chi-Squared Distribution

5. Which of the following distributions is Continuous

a) Binomial Distribution

b) Hypergeometric Distribution

c) F Distribution

d) Poisson Distribution

Answer - 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

Answer - 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

Answer - 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

Answer - a) Two tailed

9. Alternative Hypothesis is also called as?

a) Composite hypothesis

b) Research Hypothesis

c) Simple Hypothesis

d) Null Hypothesis

Answer - 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

Answer - a) np

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?

Answer – R-squared is generally a better measure of goodness of fit in regression.

R-squared provides the proportion of the variance in the dependent variable explained by the independent variables. It ranges from 0 to 1 where higher values indicate a better fit. Residual Sum of Squares (RSS) on the other hand measures the total squared differences between observed and predicted values. It doesn't account for the total variability in the data which makes it less comprehensive than R-squared.

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.

Answer – In regression –

TSS (Total Sum of Squares) is the total variability in the dependent variable.

ESS (Explained Sum of Squares) represents the variability explained by the regression model.

RSS (Residual Sum of Squares) is the unexplained variability or residuals.

The equation relating these three metrics with each other is:

TSS = ESS + RSS.

The above equation signifies that the total variability can be decomposed into the explained variability and the unexplained variability captured by the residuals.

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

Answer – Regularization in machine learning is needed to prevent overfitting. In regularization, a model learns the training data too well and performs poorly on new or unseen data. Regularization helps in achieving a balance between fitting the training data and generalizing well to new data. It helps in improving the model's performance on unseen instances.

4. What is Gini–impurity index?

Answer – The Gini- impurity index is a measure of how often a randomly selected element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the set.

In decision tree algorithms, it is used to evaluate the impurity or disorder of a set of data points based on their class labels. A lower Gini impurity indicates a more homogeneous set.

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

Answer – Yes, unregularized decision - trees are prone to overfitting.

Unregularized decision-trees can become excessively complex and capturing noise in the training data and fitting it too closely. Without constraints, decision trees can memorize the training data rather than generalize patterns that leads to poor performance on new or unseen data.

Regularization techniques like pruning help prevents overfitting by preventing the model from becoming overly complex.

6. What is an ensemble technique in machine learning?

Answer – An ensemble technique in machine learning combines the predictions of multiple individual models to improve overall performance and generalization.

Some of the common ensemble methods include bagging e.g., Random Forest and boosting which leverage the strength of diverse models to enhance predictive accuracy and robustness.

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

Answer – The main difference between Bagging and Boosting techniques lies in how they combine individual models:

Bagging :

1. It builds multiple models independently on different subsets of the training data, often through bootstrapping and then averages or takes a vote to make predictions.

Random Forest is a common example.

2. Bagging creates diverse models in parallel.

Boosting :

1. It sequentially builds models each correcting the errors of its predecessor. It assigns weights to training instances emphasizing the misclassified ones.

Example - AdaBoost and Gradient Boosting are popular boosting algorithms.

2. Boosting builds models sequentially focusing on instances with higher errors.

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

Answer – The out-of-bag (OOB) error in Random Forests is the prediction error calculated on the observations not used in the training of a particular decision tree within the ensemble. Each tree is trained on a bootstrap sample.

The OOB error is computed by evaluating each tree on the samples it did not use during training. OOB error provides a convenient estimate of the model's performance without the need for a separate validation set.

9. What is K-fold cross-validation?

Answer – K-fold cross-validation is a technique used in machine learning to assess the performance of a model.

a. The dataset is divided into K subsets,

b. The model is trained K times, each time using K-1 folds for training and the remaining fold for validation.

This process helps evaluate the model's performance more robustly by utilizing different subsets for training and validation in each iteration.

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

Answer – Hyper parameter tuning in machine learning involves optimizing the settings of hyperparameters which are configuration settings external to the model.

It is done because of the following reason -

It is done to find the best combination of hyperparameters that maximizes the model's performance. This process helps improve the model's accuracy, generalization and ability to handle new data by fine-tuning aspects like learning rates, regularization strengths and other configuration parameters.

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

Answer – A large learning rate in Gradient Descent can lead to issues such as overshooting the minimum and causing the algorithm to fail to converge.

It may result in unstable and oscillatory behaviour, preventing the model from reaching an optimal solution.

This can hinder convergence, make the algorithm diverge and lead to poor generalization on new data. Choosing an appropriate learning rate is crucial for effective training.

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

Answer –

Logistic Regression is inherently a linear classifier, so it might not effectively handling non-linear data. In the case of non-linear data, the decision boundary may not be accurately captured by a linear function which leads to suboptimal performance.

Models capable of capturing non-linear relationships are more sophisticated models like support vector machines or decision trees which are better suited for such scenarios. These can be applied to Logistic Regression to make it more adaptable to non-linear patterns.

13. Differentiate between Adaboost and Gradient Boosting.

Answer –

1. Weight Assignment:

Adaboost: Adjusts weights of misclassified instances at each iteration, assigning higher weights to those misclassified.

Gradient Boosting: Builds trees sequentially each one focusing on the errors of the previous trees.

2. Model's Focus:

Adaboost: Gives more emphasis to difficult-to-classify instances by adjusting weights.

Gradient Boosting: Focuses on minimizing errors by optimizing the loss function through sequential tree building.

3. Learning Rate:

Adaboost: Utilizes a learning rate which is a factor applied to the weights to control the contribution of each weak learner.

Gradient Boosting: Employs a learning rate to control the step size in the direction of minimizing the loss function.

4. Base Learners:

Adaboost: Often uses shallow trees as its base models.

Gradient Boosting: Typically employs decision trees as weak learners but can use other base learners.

5. Weighted Average:

Adaboost: Combines weak learners through a weighted average based on their performance.

Gradient Boosting: Combines weak learners by adding them sequentially, with each tree correcting the errors of the previous ones.

6. Loss Function Optimization:

Adaboost: Focuses on reducing the classification error.

Gradient Boosting: Optimizes a specified loss function, often mean squared error for regression problems or deviance for classification.

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

Answer – The bias-variance trade off in machine learning refers to the delicate balance between model simplicity (bias) and flexibility (variance).

A model with high bias may oversimplify and consistently make the same mistakes often called underfitting.

While on the other hand, a high-variance model may be too flexible, capturing noise in the training data and performing poorly on new data often called overfitting.

So achieving an optimal trade off in machine learning involves finding a model complexity that minimizes both bias and variance for better generalization to new unseen data.

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

Answer –

1. **Linear Kernel:**

- Description: The linear kernel is the simplest SVM kernel. It represents the dot product between feature vectors in the original feature space.

- Use Case: Suitable for linearly separable data where a straight line can effectively divide the classes.

2. **RBF (Radial Basis Function) Kernel:**

- Description: The RBF kernel measures the similarity between data points in a high-dimensional space. It is effective in capturing non-linear relationships.

- Use Case: Useful for complex, non-linear data where the decision boundary is not a straight line.

3. **Polynomial Kernel:**

- Description: The polynomial kernel calculates the similarity between points based on polynomial functions. It introduces non-linearity and can be adjusted using the polynomial degree parameter.

- Use Case: Suitable for data with non-linear patterns and the degree parameter controls the level of non-linearity introduced.

Each of these kernels serves a specific purpose in SVM, allowing the algorithm to handle various types of data structures and complexities.