**MACHINE LEARNING**

1.Which of the following methods do we use to find the best fit line for data in Linear Regression?

1. Least Square Error B) Maximum Likelihood C) Logarithmic Loss D) Both A and B

ANS. A) Least Square Error

2. Which of the following statement is true about outliers in linear regression?

A) Linear regression is sensitive to outliers B) linear regression is not sensitive to outliers

C) Can’t say D) none of these

ANS. A) Linear regression is sensitive to outliers

3. A line falls from left to right if a slope is \_\_\_\_\_\_?

A) Positive B) Negative C) Zero D) Undefined

ANS. B) Negative

4. Which of the following will have symmetric relation between dependent variable and independent variable?

A) Regression B) Correlation C) Both of them D) None of these

ANS. B) Correlation

5. Which of the following is the reason for over fitting condition?

A) High bias and high variance B) Low bias and low variance C) Low bias and high variance D) none of these

ANS. C) Low bias and high variance

6. If output involves label then that model is called as:

A) Descriptive model B) Predictive modal C) Reinforcement learning D) All of the above

ANS. B) Predictive model

7. Lasso and Ridge regression techniques belong to \_\_\_\_\_\_\_\_\_?

A) Cross validation B) Removing outliers C) SMOTE D) Regularization

ANS. D) Regularization

8. To overcome with imbalance dataset which technique can be used?

A) Cross validation B) Regularization C) Kernel D) SMOTE

ANS. D) SMOTE

9. The AUC Receiver Operator Characteristic (AUCROC) curve is an evaluation metric for binary classification problems. It uses \_\_\_\_\_ to make graph?

A) TPR and FPR B) Sensitivity and precision C) Sensitivity and Specificity D) Recall and precision

ANS. A) TPR and FPR

10. In AUC Receiver Operator Characteristic (AUCROC) curve for the better model area under the curve should be less.

A) True B) False

ANS. B) False

11. Pick the feature extraction from below:

A) Construction bag of words from a email B) Apply PCA to project high dimensional data

C) Removing stop words D) Forward selection

ANS. B) Apply PCA to project high dimensional data

12. Which of the following is true about Normal Equation used to compute the coefficient of the Linear Regression?

A) We don’t have to choose the learning rate. B) It becomes slow when number of features is very large. C) We need to iterate. D) It does not make use of dependent variable

ANS. B) It becomes slow when the number of features is very large.

13. Explain the term regularization?

ANS. **Regularization** is a crucial technique in machine learning and statistics used to enhance model performance by controlling model complexity and reducing the risk of overfitting.

14. Which particular algorithms are used for regularization?

ANS. Regularization techniques are applied across various algorithms to control model complexity, prevent overfitting, and improve generalization by incorporating additional penalties or constraints into the model training process.

15. Explain the term error present in linear regression equation?

ANS. In linear regression, the term **error** encompasses various measures of the discrepancies between the observed data and the model's predictions. Understanding and analyzing these errors is essential for assessing model performance, diagnosing issues, and improving the model's accuracy. Regularly evaluating errors helps in refining the model and ensuring that it generalizes well to new, unseen data.

**PYTHON – WORKSHEET 1**

1. Which of the following operators is used to calculate remainder in a division?

A) # B) & C) % D) $

ANS. C) %

2. In python 2//3 is equal to?

A) 0.666 B) 0 C) 1 D) 0.67

ANS. B) 0

3. In python, 6<<2 is equal to ?

A)36 B)10 C)24 D)45

ANS. C) 24

4. In python, 6&2 will give which of the following as output?

A) 2 B) True C) False D) 0

ANS. A) 2

5. In python, 6|2 will give which of the following as output?

A) 2 B) 4 C) 0 D) 6

ANS. D) 6

6. What does the finally keyword denotes in python?

A) It is used to mark the end of the code B) It encloses the lines of code which will be executed if any error occurs while executing the lines of code in the try block. C) the finally block will be executed no matter if the try block raises an error or not. D) None of the above

ANS. C) The finally block will be executed no matter if the try block raises an error or not.

7. What does raise keyword is used for in python?

A) It is used to raise an exception. B) It is used to define lambda function

C) it's not a keyword in python. D) None of the above

ANS. A) It is used to raise an exception.

8. Which of the following is a common use case of yield keyword in python?

A) in defining an iterator B) while defining a lambda function C) in defining a generator D) in for loop.

ANS. C) in defining a generator

9. Which of the following are the valid variable names?

A) \_abc B) 1abc C) abc2 D) None of the above

ANS. A) \_abc

10. Which of the following are the keywords in python?

A) yield B) raise C) look-in D) all of the above

ANS. D) all of the above

**STATISTICS WORKSHEET-1**

1. Bernoulli random variables take (only) the values 1 and 0.

a) True b) False

ANS. a) True

2. Which of the following theorem states that the distribution of averages of iid variables, properly normalized, becomes that of a standard normal as the sample size increases?

a) Central Limit Theorem b) Central Mean Theorem c) Centroid Limit Theorem d) All of the mentioned

ANS. a) Central Limit Theorem

3. Which of the following is incorrect with respect to use of Poisson distribution?

a) Modeling event/time data b) Modeling bounded count data c) Modeling contingency tables d) All of the mentioned

ANS. d) All of the mentioned

4. Point out the correct statement.

a) The exponent of a normally distributed random variables follows what is called the log- normal distribution

b) Sums of normally distributed random variables are again normally distributed even if the variables are dependent

c) The square of a standard normal random variable follows what is called chi-squared distribution

d) All of the mentioned

ANS. c) The square of a standard normal random variable follows what is called chi-squared distribution

5. \_\_\_\_\_\_ random variables are used to model rates. a) Empirical b) Binomial c) Poisson d) All of the mentioned

ANS. c) Poisson

6. 10. Usually replacing the standard error by its estimated value does change the CLT.

a) True b) False

ANS. b) False

7. 1. Which of the following testing is concerned with making decisions using data? a) Probability b) Hypothesis c) Causal d) None of the mentioned

ANS. b) Hypothesis

8. 4. Normalized data are centered at\_\_\_\_\_\_and have units equal to standard deviations of the original data. a) 0 b) 5 c) 1 d) 10

ANS. a) 0

9. Which of the following statement is incorrect with respect to outliers? a) Outliers can have varying degrees of influence b) Outliers can be the result of spurious or real processes c) Outliers cannot conform to the regression relationship d) None of the mentioned

ANS. c) Outliers cannot conform to the regression relationship

10. What do you understand by the term Normal Distribution?

ANS. The normal distribution is a key concept in probability and statistics due to its mathematical properties and its natural occurrence in various data sets. Its bell-shaped curve, defined by its mean and standard deviation, makes it a versatile tool for modeling and analyzing continuous data.

11. How do you handle missing data? What imputation techniques do you recommend?

ANS. Handling missing data effectively depends on the nature of the missingness, the type of data, and the goals of the analysis. Common imputation techniques include mean, median, or mode imputation, KNN, regression, and multiple imputation. Advanced methods like matrix factorization and deep learning approaches may be used for complex datasets. Always assess the impact of your chosen method on the overall analysis or model performance.

12. What is A/B testing?

ANS. A/B testing is a powerful method for comparing two versions of a variable to determine which performs better based on specific metrics. It involves randomizing participants, defining success metrics, and analyzing results to make informed decisions. It is widely used for optimization in various domains and helps in making data-driven improvements.

13. Is mean imputation of missing data acceptable practice?

ANS. Mean imputation is a simple and often used method for handling missing data, but it comes with limitations. It can be acceptable in certain scenarios, especially when the missing data is MCAR and the proportion of missing values is small. However, it may introduce bias and reduce variability. For more accurate and robust results, considering alternative methods such as multiple imputation or KNN imputation may be beneficial, especially when dealing with larger proportions of missing data or data that is not missing completely at random.

14. What is linear regression in statistics?

ANS. Linear regression is a statistical technique used to model the linear relationship between a dependent variable and one or more independent variables. It is characterized by its simplicity and effectiveness in predicting outcomes and understanding relationships. The method involves estimating coefficients that best fit the data, adhering to assumptions of linearity, independence, and constant variance of residuals. Linear regression is widely used across various fields for prediction, forecasting, and analyzing data relationships.

15. What are the various branches of statistics?

ANS. Each branch of statistics addresses different aspects of data analysis, from summarizing and visualizing data to making predictions and testing hypotheses. Understanding these branches helps in selecting appropriate methods for analyzing data, making informed decisions, and deriving meaningful insights from complex datasets.