MACHINE LEARNING

Assignment-39

1. To find the best fit line for data in Linear Regression we use **(A)Least Square Error** Method

2. Which of the following statement is true about outliers n Linear Regression?
   (A) A Linear Regression is Sensitive to Outliers

3. A line falls from left to right if a slope is **(B)Negative.**

4. **(B)Correlation** will have symmetric relation between dependent variable and independent variable.

5. **(C) Low Bias and High Variance** are the reason for Over fitting condition

6.

7. Lasso and Ridge regression techniques belong to **(D)Regularization.**
8. To overcome with imbalanced dataset **(D)SMOTE** technique can be used.
9. The AUC Receiver Operator Characteristic (AUCROC) is an evaluation metric for binary classification problems. It uses **(A)TPR and FPR** to make Graph.
10. In AUC Receiver Operator Characteristic (AUCROC) curve for the better model area under the curve should be less. **(B)FALSE**
11.
12. The following is true about the Normal equation used to compute the coefficient of he Linear Regression.

    (C)We need to iterate

    (D) It does not make the use of dependent variable

13. **Explain the term Regularization**?

Regularization is a technique used to reduce the errors by fitting the function appropriately on the given training set and avoid overfitting.
The commonly used regularization techniques are

1.L1 regularization \LASSO (Least Absolute Shrinkage and Selection Operator)

2. L2 regularization\Ridge

14.Which particular algorithm are used for regularization?

Ridge Regression and LASSO Regressions are the algorithms which are particularly used for regularization

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

The simple linear regression model is represented by: $y = \beta0 + \beta1x + \varepsilon$. The linear regression model contains an error term that is represented by ε. The error term is used to account for the variability in y that cannot be explained by the linear relationship between x and y.

PYTHON WORKSHEET 1

1. **(C) %** is the operator used to calculate remainder in a division.
2. In python 2//3 is equal to **(B) 0.**
3. In python 6<<2 is equal to **(C) 24.**
4. In python, 6&2 will give the output as **(A) 2.**
5. in python, 6|2 will give the output as **(D) 6.**
6. What does the finally keyword denotes in python?
7. What does the raise keyword used in python for?

   (A)It is used to raise an exception

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

   (C) in defining a Generator

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

   A) _abc

   C) abc2

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

    (A) yield

    (B) raise

STATISTIC WORKSHEET

1. Bernoulli random Variable take(only) the value 1 and 0: **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?
   c) Central Limit Theorem

3. Which of the following is incorrect with Poisson Distribution?

b) Modeling Bounded Count Data

4. Point out the correct statement.

d) All of the mentioned above

5. c) Poisson Random Variable are used to model rates

6. Usually replacing the standard error by its estimates value does change the CLT: **b) False.**

7. Which of the following testing is concerned with making decision using data?

b) Hypothesis

8. Normalized data are centered at **a) 0** and have units equal to standard deviations of the original data.

9. Which of the following statement Is incorrect with respect to outliers**?**

c) Outliers cannot conform to the regression relationship

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

Normal Distribution simply means normally distributed data or else properly distributed data. The normal distribution is the most widely known and used for all distributions. Because the normal distribution approximates many natural phenomena so well, it has developed into a standard of reference for many probability problems.

The normal distribution is easy to work mathematically. In many practical cases, the methods developed using normal theory work quite well even when the distribution is not normal. There is a strong connection between the size of the sample N and the extent to which a sampling distribution approaches the normal form. Many sampling distributions based on large N can be approximated by the normal distribution even though the population distribution itself is definitely not normal

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

We use two data imputation techniques to handle missing data: Average imputation and common-point imputation. Average imputation uses the average value of the responses from other data entries to fill out missing values. However, a word of caution when using this method – it can artificially reduce the variability of the dataset. Common-point imputation, on the other hand, is when the data scientists utilize the middle point or the most commonly chosen value. For example, on a five-point scale, the substitute value will be 3. Something to keep in mind when utilizing this method is the three types of middle values: mean, median and mode, which is valid for numerical data

12. What is A/B testing?

In statistical terms, A/B testing is a method of two-sample hypothesis testing. This means comparing the outcomes of two different choices (A&B) by running a controlled mini experiment. This method is also sometimes referred to as **split testing**

13. Is mean imputation of missing data acceptable for practice?

It is acceptable when the missing value proportion is not large enough. But, when the missing values are large enough and you impute them with the mean, the standard errors will be lesser

14. What is linear regression in statistics?

Linear regression is the simplest and most extensively used statistical technique for predictive modelling analysis. It is a way to explain the relationship between a dependent variable (target) and one or more explanatory variables(predictors) using a straight line. There are two types of linear regression - Simple and Multiple.

15. What are the various branches of statistics?

There are two branches of statistic that are Descriptive statistic and Inferential Statistics. These have a specific scientific approach which makes them equally essential.

Descriptive statistic:

Descriptive statistics is considered as the first part of statistical analysis which deals with collection and presentation of data. Scientifically, descriptive statistics can be defined as brief explanatory coefficients that are used by statisticians to summarize a given data set. Generally, a data set can either represent a sample of a population or the entire populations.

Inferential statistic:

inferential statistics are techniques that enable statisticians to use the gathered information from a sample to make inferences, decisions or predictions about a given population. Inferential statistics often talks in probability terms by using descriptive statistics. These techniques are majorly used by statisticians to analyze data, make estimates and draw conclusions from the limited information which is obtained by sampling and testing how reliable the estimates are.