PYTHON – WORKSHEET 1

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

Ans - C) %

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

Ans - B) 0

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

Ans - C) 24

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

Ans - A) 2

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

Ans - D) 6

6. What does the finally keyword denotes in python?

Ans - A) It is used to mark the end of the code

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

Ans - A) It is used to raise an exception.

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

Ans - B) while defining a lambda function

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

Ans - A) _abc

C) abc2

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

Ans - A) yield

B) raise

STATISTICS WORKSHEET-1

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

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?

Ans - a) Central Limit Theorem

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

Ans - b) Modeling bounded count data

4. Point out the correct statement.

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

5. ______ random variables are used to model rates.

Ans - c) Poisson

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

Ans - b) False

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

Ans - b) Hypothesis

8. Normalized data are centered at______and have units equal to standard deviations of the original data.

Ans - a) 0

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

Ans - c) Outliers cannot conform to the regression relationship.

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

Ans - A normal distribution of data is a data in which the most of data points are relatively similar which means that they appear within a small range of values with less outliers on the high and low ends of the data range. When the data are normally distributed , plotting them on a graph results in a bell like shaped uniform image which is called as Bell curve. In such a distribution of data mean, median and mode are all of the same value and agree with the peak of the curve.

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

Ans - While dealing with missing data, we can use two primary methods, imputation or the removal of data. The imputation method helps in progress appropriate guesses for missing data. It is more useful when the percentage of missing data is low. If the percentage of missing data is very high, the results lack natural variation that could result in an effective model.The other option is to remove data.

When dealing with data is random, related data can be used to reduce bias. Removing data may not be the best option if there are not enough observations to result in a reliable analysis. Multiple imputation is considered a good approach for data sets with a large amount of missing data.

12. What is A/B testing?

Ans - An A/B testing is an example of statistical hypothesis testing, a process whereby a hypothesis is made about the relationship between two data sets and those data sets are then compared against each other to determine if there is statistically significant relationship or not.

13. Is mean imputation of missing data acceptable practice?

Ans - Mean imputation is considered to be a terrible practice since it ignores feature correlation. The process of replacing null values in a data collection with the data's mean is known as mean imputation. Also, mean imputation decreases the variance of our data while increasing bias. As a result of the reduced variance, the model is less accurate and the confidence interval is narrower.

14. What is linear regression in statistics?

Ans - Linear regression is one of the most fundamental algorithms in the machine learning world. It is a supervised learning algorithm and simplest form of regression used to study the relationship between the variables. Linear regression is used to decide the character and strength between relationship between the dependent variable and other independent variables. It helps makes models to predict such as crop yield for the next year.

linear regression tries to pattern the relationship between two variables by applying a linear equation to the observed data. A linear regression line can be represented using the equation of a straight line.

Y = mx+c

Wherein,

- y is the estimated dependant variable (or the output)
- m is the regression coefficient (or the slope)
- x is the independent variable (or the input)
- c is the constant (or the y-intercept)

15. What are the various branches of statistics?

Ans - There are two branches of statistics that is Descriptive Statistics and Inferential Statistics. Descriptive Statistics : In this type of statistics, the data is summed up through the given observations. The summing up is one from a sample of population using parameters such as the mean or standard deviation. Descriptive statistics is a wat\y to organise, represent and describe a collection of data using tables, graphs etc.

Descriptive statistics are also categorised into four different types:

- Measure of frequency
- Measure of dispersion
- Measure of central tendency
- Measure of position

The frequency measurement decribes the number of times a particular data occurs.

Range, Variance, Standard Deviation are measures of dispersion. It identifies the spread of data.

Central tendencies are the mean, median and mode of the data.

Inferential Statistics:

This type of statistics is used to interpret the meaning of Descriptive statistics. That means once the data has been collected, analysed and summarised then we use these stats to describe the meaning of the collected data.

Inferential Statistics is a method that allows us to use information collected from a sample to make decisions, predictions or inferences from a population

MACHINE LEARNING

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

Ans - A) Least Square Error

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

Ans - A) Linear regression is sensitive to outliers

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

Ans - B) Negative

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

Ans - C) Both of them

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

Ans - A) High bias and high variance

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

Ans - B) Predictive modal

7. Lasso and Ridge regression techniques belong to _________?

Ans - B) Removing outliers

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

Ans - D) SMOTE

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

Ans - A) TPR and FPR

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

Ans - B) False

11. Pick the feature extraction from below:

Ans - A) Construction bag of words from a email

B) Apply PCA to project high dimensional data

C) Removing stop words

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

Ans - B) It becomes slow when number of features is very large.

C) We need to iterate.

13. Explain the term regularization?

Ans - Regularizations are techniques used to reduce the error by fitting a function appropriately on the given training set and avoid overfitting. The word regularize means to make things regular or acceptable. Regularization refers to techniques that are used to calibrate machine learning models in order to minimize the adjusted loss function and prevent overfitting or under fitting.

There are two main types of regularization techniques: Ridge and Lasso.

Ridge :

Also known as Ridge Regression, it modifies the over-fitted or under fitted models by adding the penalty equivalent to the sum of the squares of the magnitude of coefficients. This means that the mathematical function representing our machine learning model is minimized and coefficients are calculated. The magnitude of coefficients is squared and added. Ridge Regression performs regularization by shrinking the coefficients present.

Lasso :

It modifies the over-fitted or under-fitted models by adding the penalty equivalent to the sum of the absolute values of coefficients. Lasso regression also performs coefficient minimization, but instead of squaring the magnitudes of the coefficients, it takes the true values of coefficients. This means that the coefficient sum can also be 0, because of the presence of negative coefficients.

14. Which particular algorithms are used for regularization?

Ans - Ridge :

Also known as Ridge Regression, it modifies the over-fitted or under fitted models by adding the penalty equivalent to the sum of the squares of the magnitude of coefficients. This means that the mathematical function representing our machine learning model is minimized and coefficients are calculated. The magnitude of coefficients is squared and added. Ridge Regression performs regularization by shrinking the coefficients present.

Lasso :

It modifies the over-fitted or under-fitted models by adding the penalty equivalent to the sum of the absolute values of coefficients. Lasso regression also performs coefficient minimization, but instead of squaring the magnitudes of the coefficients, it takes the true values of coefficients. This means that the coefficient sum can also be 0, because of the presence of negative coefficients.

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

Ans - An error term is the term in a model regression that calculates and is responsible for the unexplained difference between the actually observed values of the independent values and the results predicted by the model. The error term can point out that the model can be made better by joining in an another independent variable that describes some or all of the difference, meaning that the dependent and independent variables are not co related to any greater degree.There are two types of errors used in regression analysis: Absolute error and Relative error.