PYTHON WORKSHEET 1

Q1. Which on the following operation is used to calculate remainder in a division ?

C) %

Q2. In python 2//3 is equal to?

B) 0

Q3. In python, 6<<2 is equal to ?

C) 24

Q4. In python, 6&2 will give which of the following as output?

A) 2

Q5. In python, 6|2 will give which of the following as output?

D) 6

Q6. What does the finally keyword denotes in python?

D) None of the above

Q7. What does raise keyword is used for in python?

A) It is used to raise an exception.

Q8. Which of the following is a common use case of yield keyword in python?

B) 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

STATISTICS WORKSHEET 1

Q1. Bernoulli random variables take (only) the values 1 and 0.

a) True

Q2. 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

Q3. Which of the following is incorrect with respect to use of Poisson distribution?

b) Modeling bounded count data

Q4. Point out the correct statement.

c) All of the mentioned

Q5. _____ random variables are used to model rates.

d) Poisson

Q6. Usually replacing the standard error by its estimated value does change the CLT.

b) False

Q7. Which of the following testing is concerned with making decisions using data?

c) Hypothesis

Q8. Normalized data are centered at_____and have units equal to standard deviations of the original data.

a) 0

Q9. Which of the following statement is incorrect with respect to outliers?

b) Outliers cannot conform to the regression relationship

Q10. What do you understand by the term Normal Distribution?

The normal distribution is also called as Gaussian distribution. It is most commonly seen continuous distribution in nature, just as in binomial distribution every event is independent from each other. Here the mean, median and mode all line up such that mean comes in centre of distribution, because of this exactly half of the result fall on either side of the mean. It is also identified by its bell-shaped curve which is called as bell curve.

Q11. How do you handle missing data? What imputation techniques do you recommend?

How we deal with missing data mainly depends on the percentage of missing data, if the data missing is more than 5% there definitely needs any imputation technique to be followed for replacement of missing data and neglect if it is less than 5%.

The most common and easy imputation techniques I would recommend are :

1. Zero Replacement: Here, you replace the missing value with zero irrespective of everything.

2. Min or Max Replacement: Replace the missing value with the minimum or maximum value of a feature.
3. Mean/ Median/ Mode Replacement: Replace missing value with mean or median or most frequent feature value. (Since mean imputation method has serious drawbacks we can go with some advanced imputation methods like predictive mean matching or stochastic regression imputation)

Q12. What is A/B testing?

A/B testing is a basic randomized control experiment. It is a way to compare the two versions of a variable to find out which performs better in the controlled environment.

Ex: Let us consider ABC company which wants to increase the sales of their product and let us apply A/B testing tool. (Test can be done on a sample of customers instead of considering whole population)

- Take the product of the company and divide it into 2 parts – A and B.
- Let A remain same and make some significant changes in B's packaging.
- Now on the basis of the response from customer group who used A and B respectively, we can decide which is performing better.

Q13. Is mean imputation of missing data acceptable practice?

Mean imputation is a bad idea.

Q14. What is linear regression in statistics?

Linear regression models the relationships between at least one explanatory variable and an outcome variable. These variables are known as the independent(input) and dependent(output) variables, respectively. When there is one independent variable (IV), the procedure is known as simple linear regression. When there are more IVs, statisticians refer to it as multiple regression.

In general linear regression model can be mathematically represented as,

$y = a + bx + e$

where,

y = Dependent variable

x = Independent variable

a = intercept

b = co-efficient of x

e = error

Q15. What are the various branches of statistics?

Statistics can be majorly branched as a. Descriptive statistics b. Inferential statistics

a. Descriptive statistics

Descriptive statistics describes the properties of sample population data which mostly focus on the central tendency, variability, and distribution of sample data

b. Inferential statistics
Inferential statistics is one which uses those properties to test hypotheses and draw conclusions. It is drawn from the characteristics of a sample, and to decide how certain they can be of the reliability of those conclusions.

MACHINE LEARNING WORKSHEET

Q1. Which of the following methods do we use to find the best fit line for data in Linear Regression?

A) Least Square Error

Q2. Which of the following statement is true about outliers in linear regression?

A) Linear regression is sensitive to outliers

Q3. A line falls from left to right if a slope is _____?

B) Negative

Q4. Which of the following will have symmetric relation between dependent variable and independent variable?

C) Correlation

Q5. Which of the following is the reason for over fitting condition?

D) Low bias and high variance

Q6. If output involves label then that model is called as:

E) All of the above

Q7. Lasso and Ridge regression techniques belong to _____?

C) Regularization

Q8. To overcome with imbalance dataset which technique can be used?

D) SMOTE

Q9. 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 Specificity

Q10. In AUC Receiver Operator Characteristic (AUCROC) curve for the better model area under the curve should be less.

A) True

11. Pick the feature extraction from below:

A) Construction bag of words from a email

Q12. 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.

Q13. Explain the term regularization ?

Regularization is a technique to prevent the model from overfitting by adding extra information to it. Sometimes the machine learning model performs well with the training data but does not perform well with the test data.

Q14. Which particular algorithms are used for regularization ?

a. Ridge regression
b. LASSO regression

Q15. Explain the term error present in linear regression equation?

An error term in linear regression is a value which represents the difference between observed data and actual population data.