STATISTICS WORKSHEEET-1

Q1 to Q9 have only one correct answer. objective type

1. Bernaoulli random variables take (only) values 1 and 0

Ans-1 TRUE

2 Which of the following theorem states that the distibution of averages of iid variables, properply normalized , becomes that of a standard nornal as the sample size increases

Ans-2 Central Limit Theorem

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

Ans-3 b. Modelling bounded count data

a. Modeling event/time data

b. Modeling bounded count data

c. Modeling contingency tables

d. All of the mentioned

4. Point out the correct statement.

Ans-4. d- All of the meenetioned

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

5. _____random variables are used to model rates.

Ans-5 poisson

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

Ans-6 . False

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

Ans 7. Hypothesis

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

Ans-8 0

9. Which of the following statement is incorrect with respect to outliers?
Ans 9. c) Outliers cannot conform to the regression relationship

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

Q 10 to Q 15 subjective type questions

10. what do you understand by the term normal distribution?
Ans= The normal distribution, also known as the Gaussian distribution, also called Bell curve ,is the most important probability distribution in statistics for independent, random variables.
The Normal distribution is a continuous probablity distribution that is symmetrical arounf its mean
moat of observation cvalue around the centrel peak It is the most important probability distributin in statistics, beacause it describe very accurately describe the distribution of values for many phenomena.
Using 1 stan dard deviation, Approximately 68% of the data falls within one standard deviation of the mean. (i.e., Between Mean - 1 std and Mean + 1 std). Approximately 95% of the data falls within 2 standard_deviation of the mean. (i.e., Between Mean- 2 std and Mean + 2 std)
eg. height of student

11. How do you handle missing data? what imputation techniques do you recommend?
Ans= Handling missing data is complex problem and totally depend dataset type.
we have many way to fill missing data , mean , median, mode method we apply to fill missing value and Sklearn imputation technique to fill missing value. But isn't simple handle missing value is depend how data is distributed and we analysis dataset type and then we choose which method is best for fill missing value those method we select

12. What is A/B testing?
Ans= A/B Testing consist of a randomized experiment with two varients, A and B. A/B testing is a way to compare two versions of a single variable
A/B tedting is widely used in e_commerce industry, web_ design industry, and much more, some industry start compaign through online eg.(one product new launching which design take to make prodect best and like large no. of customer) such condition we try A/B testing , statistical analysis

13 Is mean imputation of missing data acceptable practice?
Ans=In feature engineering mean imputation technique we doesn't say 100% acceptable pracitce in all dataset .
we first take dataset and see distribution , analysis missing value and then we try impute mean method to fill missing value. after filling missing value we plot 'kde' on dataset.feature see distribution and does'nt much more changing distrution in dataset than we process forward

14. What is linear regression in statistics?
Ans= Linear regression is a statistical model technique used to show relation ship between dependent and independent variable. It is one of the most common technique used to predict dependent variable in the linear distributed data points
The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.
Linear-regression models are relatively simple and provide an easy-to-interpret mathematical formula that can generate predictions. Linear regression can be applied to various areas in business and academic study.

15. What are the various branches of statistics?

Ans= The two main Branches of Statistics

a. Descriptive Statistics b. Inferential Statistics

= Descriptive Statistics is a branch of statistics which deals with collection of data, its presentation and organization in various forms, such as distribution tables, graphs ,diagrams (e.g., pie charts) and finding measures of central tendency and measures of dispersion or spread which are used in the description of data. Managers, CEOs. etc. make use of descriptive statistics in presenting their annual reports, financial accounts and bank statements.

Descriptive Statistics is used to present the continuous data in understandable way and then Business Organisation make some decision at understandable data

= Inferential Statistics branch of statistics which deals with technique which used in analysis of data , making estimation based on limited information those infformation taken on sample basis. Inferential statistics make prediction about larger group or population, we using on gathered sample part of population is called sample

PYTHON - WORKSHEET 1

Q1 to Q8 have objective type and one correct answer

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

Ans= % operator

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

Ans= 0

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

Ans= 24

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

Ans= 2

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

Ans= 6

6. What does the finally keyword denotes in python?

Ans= 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?

Ans=It is used to raise an exception

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

Ans= In defining a generator

Q9 and Q10 have multiple correct answers. Choose all the correct options to answer your question.

9. Which of the following are the valid variable names?
Ans= a) _abc , b) Iabc , c) abc2

10. Which of the following are the keywords in python?
Ans= A) yield , B) raise

Q11 to Q15 are programming questions. Answer them in Jupyter Notebook.

.. load in jupyter file Q11 to Q15

MACHINE LEARNING worksheet

Q1 to Q11 is objective and one answer is correct:

Q1= Which of the following methods do we use to find the best fit line for data in Linear Regression?
Ans= Least Square Error

Q2= Which of the following statement is true about outliers in linear regression?
Ans= A) Linear regression is sensitive to outliers

Q3= A line falls from left to right if a slope is _____?
Ans= Negative

Q4= Which of the following will have symmetric relation between dependent variable and independent
Ans= Correlation

Q5= Which of the following is the reason for over fitting condition?
Ans= Low Bias and High Variance

Q6= If output involves label then that model is called as:
Ans = Predictive model

Q7= Lasso and Ridge regression techniques belong to _____?
Ans= Rgularisation technique

Q8= To overcome with imbalance dataset which technique can be used?
Ans= SMOTE

Q9= The AUC Receiver Operator Characteristic (AUCROC) curve is an evaluation metric for binary classification problems. It uses _____ to make graph?
Ans= TPR and FPR

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

Q11= Pick the feature extraction from below:

Ans= Apply PCA to project high dimensional data

A) Construction bag of words from a email

B) Apply PCA to project high dimensional data

C) Removing stop words

D) Forward selection

Q12, more than one options are correct, choose all the correct options:

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

Ans= A). We don't have to choose the learning rate.

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

Q13 and Q15 are subjective answer type questions, Answer them briefly.

Q13= Explain the term regularization?

Ans= Regularization are technique used to reduce error by fitting a function appropriately on given training dataset and avoid overfitting by adding extra information to it

In Machine Learning the term 'regularisation' refers to a a set of techniques that help the machine to learn more than just of memorize. let's discuss differnce between 'learning' and 'memorizing

When you train your machin learning model and its give accurate result of training data, but provides poor result on testing data then you say your model would memorise more than generalizing

The commonnly used regularisation technique

. L1 regularization is called Lasso Regression , L2 regularization is called Ridge regression

Q14= Which particular algorithms are used for regularization?

Ans= three different type of algoritms used in regularisation

1. Ridge regression 2. Lasso Regression 3. Elastic_Net Regression

1.Ridge Regression shrinks the coefficients as it help to reduce the model complexity and multi_collinearity , Ridge adds a penalty to loss function that is equivalent to the square of magnitude of the coefficient

2. Lasso regression analysis method that perform both feature selection and regularization to enhance the prediction accuracy of model, Lasso adds penalty to loss function is equivalent to the magnitude of coefficient

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

Ans= The error term is the difference between the predicted value and the actual value.

This can range from relatively small to huge. This error term helps in the calculation of the R-squared value, that is tells us how good the model is overall. If the R-squared value of the model is 0.8, then your model explains 80% of the variation in your target variable

other aspects of the error terms give us some help in improving our model.

If the error terms follow certain patterns, it's a warning that we might be using the wrong modelling technique.