PYTHON – WORKSHEET 1

Q1 to Q8 have only one correct answer. Choose the correct option to answer your question.

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

A) #

B) &

C) %

D) $

Answer: (C) %

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

A) 0.666

B) 0

C) 1

D) 0.67

Answer: (B) 0

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

A) 36

B) 10

C) 24

D) 45

Answer: (C) 24

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

A) 2

B) True

C) False

D) 0 5.

Answer: (A) 2

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

A) 2

B) 4

C) 0

D) 6

Answer: (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

Answer: (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

Answer: (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.

Answer: (c) in defining a generator

STATISTICS WORKSHEET-1

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

a) True b) False

Answer : 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

Answer: Centroid 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

Answer: Modeling bounded count data

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

Answer: d) All of the mentioned

5. _____ random variables are used to model rates.

a) Empirical

b) Binomial

c) Poisson

d) All of the mentioned

Answer: c) Poisson

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

a) True

b) False

Answer: 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

Answer: 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

Answer: 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

Answer: c) Outliers cannot conform to the regression relationship

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

Answer:

They were first called “normal” because the pattern occurred in many different types of common measurements. There are many normal curves. Even though all normal curves have the same bell shape, they vary in their center and spread. ... The mean of a normal distribution locates its center.

The standard deviation controls the spread of the distribution. A smaller standard deviation indicates that the data is tightly clustered around the mean; the normal distribution will be taller. A larger standard deviation indicates that the data is spread out around the mean; the normal distribution will be flatter and wider.

Properties of a normal distribution:

- The mean, mode and median are all equal.
- The curve is symmetric at the center (i.e. around the mean, μ).
- Exactly half of the values are to the left of center and exactly half the values are to the right.
- The total area under the curve is 1.

One way of figuring out how data are distributed is to plot them in a graph. If the data is evenly distributed, you may come up with a **bell curve**. A bell curve has a small percentage of the points on both tails and the bigger percentage on the inner part of the curve. In the **standard normal model**, about 5 percent of your data would fall into the “tails” and 90 percent will be in between.

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

Answer:

Answer:

There are some ways to handle missing data:

1.Delete the missing data.

2.Create a separate model to handle missing data

3. Using statistical methods like Mean , Median and mode

The best Imputation technique is to delete the row of missing data. But it is not recommended if there is huge data.

Q. What is A/B testing?

Answer:

A/B testing is the act of running a simultaneous experiment between two or more variants of a page to see which one performs the best.

By sending half your traffic to one version of the page and half to another, you can first gather evidence about which one works best before you commit to the change.

Essentially, A/B testing lets you play scientist and make decisions based on data about how people actually behave when they hit your page.

Q. Is mean imputation of missing data acceptable practice?

Answer:

There are three problems with using mean-imputed variables in statistical analyses:

- Mean imputation reduces the variance of the imputed variables.
- Mean imputation shrinks standard errors, which invalidates most hypothesis tests and the calculation of confidence interval.
- Mean imputation does not preserve relationships between variables such as correlations.

Q. What is linear regression in statistics?

Answer:

Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

Q. What are the various branches of statistics?

Answer:

The two main branches of statistics are descriptive statistics and inferential statistics.

Descriptive Statistics

Descriptive statistics deals with the presentation and collection of data. This is usually the first part of a statistical analysis. It is usually not as simple as it sounds, and the statistician needs to be aware of designing experiments, choosing the right focus group and avoid biases that are so easy to creep into the experiment.

Inferential Statistics

Inferential statistics, as the name suggests, involves drawing the right conclusions from the statistical analysis that has been performed using descriptive statistics. In the end, it is the inferences that make studies important and this aspect is dealt with in inferential statistics.

Machine learning:

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

B) Maximum Likelihood

C) Logarithmic Loss

D) Both A and B

Answer: 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

Answer: 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

Answer: c) zero

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

Answer: 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

Answer: (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

Answer: All of above

7.Lasso and Ridge regression techniques belong to _____?

A) Cross validation

B) Removing outliers

C) SMOTE

D) Regularization

Answer: (D) Regularization

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

A) Cross validation

B) Regularization

C) Kernel

D) SMOTE

Answer: Cross Validation

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

Answer: (D)Recall and precision

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

A) True

B) False

Answer: True

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 In

Answer:

A) Construction bag of words from a email

B) Apply PCA to project high dimensional data

C) Removing stop words

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?

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

Answer:

A) We don’t have to choose the learning rate.

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

3. Explain the term regularization?

This is a form of regression, that constrains/ regularizes or shrinks the coefficient estimates towards zero. In other words, *this technique discourages learning a more complex or flexible model, so as to avoid the risk of overfitting.*

4. Which particular algorithms are used for regularization?

Answer:

There are mainly two types of regularization techniques, which are given below:

- Ridge Regularization
- Lasso Regularization
- Dropout Regularization

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

An **error term** in statistics is a value which represents how observed data differs from actual population data. It can also be a variable which represents how a given statistical model differs from reality.