1.The k-means algorithm...

(0/1 Point)

always converges to a clustering that minimizes the mean-square vector-representative distance

can converge to different final clustering, depending on initial choice of representatives

is widely used in practice

is typically done by hand, using paper and pencil

2. True-False: Is it possible to apply a logistic regression algorithm on a 3-class Classification problem?

(1/1 Point)

A)TRUE

B)FALSE

3.A statement made about a population for testing purpose is called?

(1/1 Point)

a) Statistic

b) Hypothesis

c) Level of Significance

d) Test-Statistic

4.True-False: Is Logistic regression a supervised machine learning algorithm?

(1/1 Point)

A)TRUE

B)FALSE

5.An independent t-test can be used to assess which of the following?

(1/1 Point)

It assesses differences between two groups of participants

It assesses relationships between two interval data sets

It assesses differences between scores obtained on two separate occasions from the same participants

6.A result is called “statistically significant” whenever

(1/1 Point)

The null hypothesis is true.

The alternative hypothesis is true.

The p-value is less or equal to the significance level.

The p-value is larger than the significance level.

7.Consider a hypothesis H0 where Mean μ = 5 against H1 where Mean μ > 5. The test is?

(1/1 Point)

a) Right tailed

b) Left tailed

c) Center tailed

d) Cross tailed

8.Suppose that we have N independent variables (X1,X2… Xn) and dependent variable is Y. Now Imagine that you are applying linear regression by fitting the best fit line using least square error on this

data. You found that correlation coefficient for one of it's variable(Say X1) with Y is -0.95. Which of the following is true for X1?

(1/1 Point)

A) Relation between the X1 and Y is weak
B) Relation between the X1 and Y is strong
C) Relation between the X1 and Y is neutral
D) Correlation can't judge the relationship

9.Analysis of variance in short form is?

(1/1 Point)

a) ANOV
b) AVA
c) ANOVA
d) ANVA

10.Considering the K-median algorithm, if points (0, 3), (2, 1), and (-2, 2) are the only points which are assigned to the first cluster now, what is the new centroid for this cluster?

(1/1 Point)

(0,2)
(2,1)
(2,0)
(1,2)

11.True-False: Linear Regression is mainly used for Regression.

(1/1 Point)

A) TRUE
B) FALSE

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

(0/1 Point)

A) Least Square Error
B) Maximum Likelihood
C) Logarithmic Loss
D) Both A and B

13.When there are more than one independent variables in the model, then the linear model is termed as _____

(1/1 Point)

a) Unimodal
b) Multiple model
c) Multiple Linear model
d) Multiple Logistic model

14.Which of the following methods do we use to best fit the data in Logistic Regression?

(1/1 Point)

Least Square Error
Maximum Likelihood
Jaccard distance

Both A and B

15.Consider a hypothesis where H0 where μ = 23 against H1 where μ < 23. The test is?

(0/1 Point)

a) Right tailed
b) Left tailed
c) Center tailed
d) Cross tailed

16.The confidence level for a confidence interval for a mean is

(0/1 Point)

The probability the procedure provides an interval that covers the sample mean.
The probability of making a Type 1 error if the interval is used to test a null hypothesis about the population mean.
The probability that individuals in the population have values that fall into the interval.
The probability the procedure provides an interval that covers the population mean.

17.True-False: Is Logistic regression mainly used for Regression?

(1/1 Point)

A)TRUE
B)FALSE

18. Type 1 error occurs when?

(1/1 Point)

a) We reject H0 if it is True
b) We reject H0 if it is False
c) We accept H0 if it is True
d) We accept H0 if it is False

19.It is known that for right-handed people, the dominant (right) hand tends to be stronger. For left-handed people who live in a world designed for right-handed people, the same may not be true. To test this, muscle strength was measured on the right and left hands of a random sample of 15 left-handed men and the difference (left - right) was found. The alternative hypothesis is one-sided (left hand stronger). The resulting t-statistic was 1.80. This is an example of:

(1/1 Point)

A two-sample t-test.
A paired t-test.
A pooled t-test.
An unpooled t-test.

20.The rejection probability of Null Hypothesis when it is true is called as?

(0/1 Point)

a) Level of Confidence
b) Level of Significance
c) Level of Margin
d) Level of Rejection

21. In practice, Line of best fit or regression line is found when _____

(1/1 Point)

a) Sum of residuals (∑(Y – h(X))) is minimum
b) Sum of the absolute value of residuals (∑|Y-h(X)|) is maximum
c) Sum of the square of residuals (∑ (Y-h(X))2) is minimum
d) Sum of the square of residuals (∑ (Y-h(X))2) is maximum

22.K-means is an iterative algorithm, and two of the following steps are repeatedly carried out in its inner-loop. Which two?

(1/1 Point)

Assign each point to its nearest cluster
Test on the cross-validation set
Update the cluster centroids based the current assignment
Using the elbow method to choose K

23.The goal of clustering a set of data is to

(1/1 Point)

divide them into groups of data that are near each other
choose the best data from the set
determine the nearest neighbors of each of the data
predict the class of data

24._____ is a clustering procedure where all objects start out in one giant cluster. Clusters are formed by dividing this cluster into smaller and smaller clusters.

(1/1 Point)

Non-hierarchical clustering
Divisive clustering
Agglomerative clustering
K-means clustering

25.In hypothesis testing, a Type 2 error occurs when

(1/1 Point)

The null hypothesis is not rejected when the null hypothesis is true.
The null hypothesis is rejected when the null hypothesis is true.
The null hypothesis is not rejected when the alternative hypothesis is true.
The null hypothesis is rejected when the alternative hypothesis is true.

26. The point where the Null Hypothesis gets rejected is called as?

(1/1 Point)

a) Significant Value
b) Rejection Value
c) Acceptance Value
d) Critical Value

27. In a simple linear regression model (One independent variable), If we change the input variable by 1 unit. How much output variable will change?

(1/1 Point)

a) by 1
b) no change
c) by intercept

d) by its slope

28. If the Critical region is at both the sides of distribution then the test is referred as?

(1/1 Point)

a) Two tailed
b) One tailed
c) Three tailed
d) Zero tailed

29.The probability of Type 1 error is referred as?

(1/1 Point)

a) 1-α
b) β
c) α
d) 1-β

30.Which of the following is true about Residuals ?

(1/1 Point)

A) Lower is better
B) Higher is better
C) A or B depend on the situation
D) None of these

31.1) True-False: Linear Regression is a supervised machine learning algorithm.

(1/1 Point)

A) TRUE
B) FALSE

32.Which of the following evaluation metrics can be used to evaluate a model while modeling a continuous output variable?

(1/1 Point)

A) AUC-ROC
B) Accuracy
C) Logloss
D) Mean-Squared-Error

33.For which of the following tasks might clustering be a suitable approach?

(1/1 Point)

Given sales data from a large number of products in a supermarket, estimate future sales for each of these products.
Given a database of information about your users, automatically group them into different market segments.
From the user's usage patterns on a website, identify different user groups.
Given historical weather records, predict if tomorrow's weather will be sunny or rainy.

34.A significance test based on a small sample may not produce a statistically significant result even if the true value differs substantially from the null value. This type of result is known as

(1/1 Point)

the significance level of the test.

the power of the study.
a Type 1 error.
a Type 2 error.

35.How many coefficients do you need to estimate in a simple linear regression model (One independent variable)?

(1/1 Point)

a) 1
b) 2
c) 3
d) 4

36. It is known that for right-handed people, the dominant (right) hand tends to be stronger. For left-handed people who live in a world designed for right-handed people, the same may not be true. To test this, muscle strength was measured on the right and left hands of a random sample of 15 left-handed men and the difference (left - right) was found. The alternative hypothesis is one-sided (left hand stronger). The resulting t-statistic was 1.80. Assuming the conditions are met, based on the t-statistic of 1.80 the appropriate conclusion for this test using α = .05 is: (Refer t-Table)

(0/1 Point)

Df = 14, so p-value < .05 and the null hypothesis can be rejected.
Df = 14, so p-value > .05 and the null hypothesis cannot be rejected.
Df = 28, so p-value < .05 and the null hypothesis can be rejected.
Df = 28, so p-value > .05 and the null hypothesis cannot be rejected.

37.Question

(-/1 Point)

Option 1
Option 2

38.A random sample of 25 college males was obtained and each was asked to report their actual height and what they wished as their ideal height. A 95% confidence interval for μd = average difference between their ideal and actual heights was 0.8" to 2.2". Based on this interval, which one of the null hypotheses below (versus a two-sided alternative) can be rejected?

(0/1 Point)

H0: μd = 0.5
H0: μd = 1.0
H0: μd = 1.5
H0: μd = 2.0

39. What are the two types of Hierarchical Clustering

(1/1 Point)

Top-Down Clustering (Divisive)
Bottom-Top Clustering (Agglomerative)
Dendrogram
K-means

40. If the null hypothesis is false then which of the following is accepted?

(1/1 Point)

a) Null Hypothesis
b) Positive Hypothesis
c) Negative Hypothesis
d) Alternative Hypothesis.

41. In the mathematical Equation of Linear Regression Y = β1 + β2X + ϵ, (β1, β2) refers to _____

(1/1 Point)

a) (X-intercept, Slope)
b) (Slope, X-Intercept)
c) (Y-Intercept, Slope)
d) (slope, Y-Intercept)

42.Null and alternative hypotheses are statements about: A.population parameters. B.sample parameters. C.sample statistics. D.it depends - sometimes population parameters and sometimes sample statistics.

(1/1 Point)

population parameters.
sample parameters.
sample statistics.
it depends - sometimes population parameters and sometimes sample statistics.

43.A hypothesis test is done with 5% significance level in which the alternative hypothesis is that more than 10% of a population is left-handed. The p-value for the test is calculated to be 0.25. Which statement is correct?

(1/1 Point)

We can conclude that more than 10% of the population is left-handed.
We can conclude that more than 25% of the population is left-handed.
We can conclude that exactly 25% of the population is left-handed.
We cannot conclude that more than 10% of the population is left-handed.

44.Which of the following metrics can be used for evaluating regression models? i) R Squared ii) Adjusted R Squared iii) F Statistics iv) RMSE / MSE / MAE

(1/1 Point)

a) ii and iv
b) i and ii
c) ii, iii and iv
d) i, ii, iii and iv

45. If the assumed hypothesis is tested for rejection considering it to be true is called?

(1/1 Point)

a) Null Hypothesis
b) Statistical Hypothesis
c) Simple Hypothesis

d) Composite Hypothesis

1.A publisher of college textbooks claims that the average price of all hardbound college textbooks is Rs.150.50. A student group believes that the actual mean is not equal to Rs. 150.50 and wishes to test their belief. The relevant null and alternative hypotheses are ____ and _____ respectively.

(1/1 Point)

a) $H_0 : \mu =150.50$ and $Ha : \mu >=150.50$
b) $H_0 : \mu =150.50$ and $Ha : \mu !=150.50$
c) $H_0 : \mu =150.50$ and $Ha : \mu <=150.50$
d) $H_0 : \mu =150.50$ and $Ha : \mu < 150.50$

2.A sports governing body wants to test whether a supplement used by professional athlet testosterone in the body. The levels of testosterone in picograms/millilitre of ten athletes were tested before and after taking the supplement. The paired sample hypothesis test is performed on data for 10 observations using a 1% significance level to decide whether or not the supplement should be banned. The test statistics shows the p-value of .00457. So _____

(1/1 Point)

a) So null hypothesis is true which means there is no significance difference in the sample sets. So supplement should not be banned.
b) So reject null hypothesis and accept alternative hypothesis which means there is significance difference in the sample sets. So supplement should be banned.
c) Insufficient information and t-score is required for conclusion.
d) None of these

3.A _______ is a set of values for the test statistic for which the null hypothesis is accepted.

(1/1 Point)

a. p-value
b. observations
c. confidence interval
d.critical value for given DoF and significance level

4.The population mean of the heights of five-year old boys is 100cm. A teacher measures the height of her twenty five students, obtaining a mean height of 105 cm and standard deviation 18. This means xbar (sample mean) =105, μ=100,s=18,n=25. A one sided Student’s t-test is performed with a 5% significance level to calculate whether the true mean is actually greater than 100 cm. The critical value for 95 %

significance level with 24 DoF is t0.05,24 = 1.7109. The test result shows that calculated t-statistic score is t = 1.3889, So the conclusion is _____

(1/1 Point)

a) Null hypothesis can not be rejected
b) Null hypothesis is rejected

5.N=20 clinical observations are given, and it is divided into four subsets i.e. k = 4 with observations n1=7, n2 = 4, n3 = 4, n4 = 5 based on attributes Low Calorie,Low Fat,Low Carbohydrate,Control respectively. What is DFwithin and DFbetween for ANOVA test.

(1/1 Point)

a. DFwithin = 19, DFbetween = 16
b. DFwithin = 16, DFbetween = 3
c. DFwithin = 3, DFbetween = 16
d. DFwithin = 4, DFbetween = 20

6.The assumption in using t-tests is that_____ is unknown, while z-tests assume it is known.

(1/1 Point)

a) Distribution
b) Median
c) Standard Deviation
d) Mean

7.Cluster the following eight points (with (x, y) representing locations) into three clusters: A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9). Initial cluster centers are: A1(2, 10), A4(5, 8) and A7(1, 2). The distance function between two points a = (x1, y1) and b = (x2, y2) is defined as- P(a, b) = |x2 – x1| + |y2 – y1| Use K-Means Algorithm to find the three cluster centers after the second iteration.

(4/4 Points)

Correct answers: C1(3, 9.5), C2(6.5, 5.25), C3(1.5, 3.5)

8._____ and _____ among fastest growing sources of big data and examples of untraditional data sources.

(1/1 Point)

1) medical imaging and video surveillance
2) social media and genetic sequencing
3) mobile data and video rendering
4) pdf and text documents

9._____ is example of quasi structured data and _____ is example of unstructured data.

(1/1 Point)

a) JSON , XML
b) XML , Video
c) RDBM, Images

d) Webclick stream data ,pdf/text documents

10.In which critical phase of analytics project lifecycle ,it's common for teams to spend at least 50% of a data science project's time.

(1/1 Point)

1) Discovery
2) Data Preparation
3) Model Planning
4) Model Building
5) Commutations results

11._____ data repository enable flexible , high performance and robust analysis.

(0/1 Point)

1) Analytic Sandbox
2) Data Marts
3) Data Warehouses
4) Spread Marts

12.In the emerging bigdata eco system, the Government, Medical, Retail, internet etc. are examples of _____ .

(1/1 Point)

1)Data Devices
2)Data Collectors
3)Data Aggregators
4) Data Buyers/Users

13.The GINA case study provides an example of how a team applied the Data Analytics lifecycle to analyze innovation data at EMC. Write GINA case study along with phases and advanced analytical methods which had appiled to identify key innovators within the company.

(3/5 Points)

14.In the mathematical Equation of Linear Regression $Y = \beta1 + \beta2X + \epsilon$, (β1, β2) refers to _____.

(1/1 Point)

a) (slope, Y-Intercept)
b) (Slope, X-Intercept)
c) (X-intercept, Slope)
d) (Y-Intercept, Slope)

15.As per the method of "least squares", we choose a regression line where the sum of the square of deviations of the points from the line is:

(1/1 Point)

a) Negative
b) Positive
c) Minimum
d) Maximum

16.What is the relation between candidate and frequent itemsets?

(1/1 Point)

a) A frequent itemset must be a candidate itemset
b) A candidate itemset is always a frequent itemset
c) A frequent itemset is a super set of a candidate itemset

d) No relation between the two

17.Logistic regression assumes a:

(1/1 Point)

a) Linear relationship between observations
b) Linear relationship between continuous predictor variables and the logit of the outcome variable
c) Linear relationship between continuous predictor variables and the outcome variable
d) Linear relationship between continuous predictor variables

18.In logistic regression the logit is:

(1/1 Point)

a) the natural logarithm of the odds ratio
b) the cube root of the sample size
c) a logarithm of a digit
d) the natural logarithm of sample size

19.In a simple linear regression model (one dependent and one independent variable) if we change the input variable by 1 unit. How much output variable will change?

(1/1 Point)

a) by intercept
b) by its slope
c) no change
d) by 1 unit

20.Consider the following set of transactions: 1. I1, I2, I3, I4, I5, I6 2. I7, I2, I3, I4, I5, I6 3. I1, I8, I4, I5 4. I1, I9, I10, I4, I6 5. I10, I2, I4, I11, I5 Find all strong association rules given the support is 0.6 and confidence is 0.8.

(4/4 Points)

1.What is statistics?

(1/1 Point)

The study of data
Defining numbers
Comparing Theory vs. Experimental
Measuring how far apart two sets of numbers are

2.A defined collection of individuals or objects about which we want to draw conclusions.

(1/1 Point)

Population
Census
Sample
Survey

3.The _____ of a data set is the sum of the data entries divided by the number of entries.

(1/1 Point)

Mean
Median
Mode
Range

4.A subset of the population which we want to collect information from. This must be random to avoid a bias result.

(1/1 Point)

Population
Census
Sample
Survey

5.This word means the middle of the data.

(1/1 Point)

Interquartile Range
Median
Mean
Mode

6. Find the midpoint: 3 , 11 , 7 , 1, 9 , 2, 13, 8, 10

(1/1 Point)

there is no midpont for this group
8.5
8
9

7.Which of the following is a measure of variability?

(1/1 Point)

mean
median
mode
standard deviation

8.All of the below sets have the same mean. Set 1-Standard Deviation=3.1 Set 2-Standard Deviation=4.9 Set 3-Standard Deviation=1.7 Set 4-Standard Deviation=3.2 Which set of data probably has the points closest to the mean?

(1/1 Point)

1
2
3
4

9.Four data sets are shown below. Set 1: {10, 19, 38, 50, 51} Set 2: {5, 21, 26, 39, 51} Set 3: {9, 38, 50, 50, 51} Set 4: {5, 28, 28, 28, 51} Which data set has the largest standard deviation?

(1/1 Point)

Set 1
Set 2
Set 3
Set 4

10.What is the standard deviation for the data given: 5, 10, 7, 12, 0, 20, 15, 22, 8, 2

(1/1 Point)

6.89
10.1
7.26
9

11.How do you find range?

(1/1 Point)

Subtract the highest and lowest value
Add all the numbers and divide
Find the middle
It is the highest number

12.This is the value that is most frequently occurring.

(1/1 Point)

Range
Mean
Median
Mode

13.Mode for data 23,45,40,33,44,23,32,49,23

(1/1 Point)

None
44
49
23

14.14. Range of the data:7,9,19,22,27,29,35,42,56.

(1/1 Point)

27
49
28
None

15.Outliers can be what from the other numbers?

(1/1 Point)

Close
Distant
Exactly
None

16.Identify the outlier for the given data? 23, 34, 27, 4, 30, 26, 28, 31, 34

(1/1 Point)

23
4
31
34

17.In a boxplot Q2 or Quartile 2, can be also known as the what?

(1/1 Point)

- Range
- Mean
- Mode
- Median

18.Which of the following lists all the five numbers needed to make a box plot?

(1/1 Point)

- Mean, Median, Mode, Range, and Total
- Minimum, Quartile 1, Median, Quartile 3, and Maximum
- Smallest, Q1, Q2, Q3, and Q4
- Minimum, Maximum, Range, Mean, and Median

19.____ is used to show central tendancy in the data when values are categorical in nature.

(0/1 Point)

- Median
- Mode
- Mean
- None of these

20.When data is highly skewed or has outliers_____ is used to show central tendancy of data.

(1/1 Point)

- Mean
- Median
- Mode
- Variance