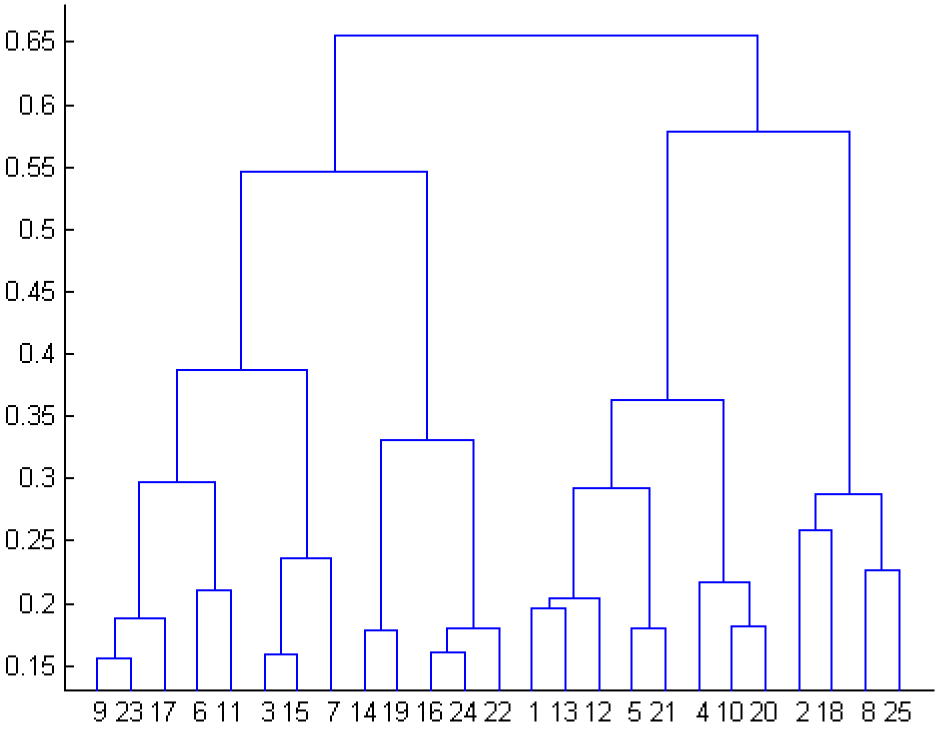
MACHINE LEARNING – WORKSHEET (CLUSTERING)

Q1 to Q12 have only one correct answer. Choose the correct option to answer your question.

1. What is the most appropriate no. of clusters for the data points represented by the following dendrogram:



a. 2
b. 4
c. 6
d. 8

Answer = b. 4

2. In which of the following cases will K-Means clustering fail to give good results?

1. Data points with outliers
2. Data points with different densities
3. Data points with round shapes
4. Data points with non-convex shapes

Options:

a. 1 and 2
b. 2 and 3
c. 2 and 4
d. 1, 2 and 4
e. 1, 2, 3 and 4

Answer = e. 1, 2, 3 and 4

3. The most important part of______is selecting the variables on which clustering is based.

a. interpreting and profiling clusters
b. selecting a clustering procedure
c. assessing the validity of clustering
d. formulating the clustering problem

Answer = formulating the clustering problem

4. The most commonly used measure of similarity is the______or its square.

a. euclidean distance
b. city-block distance
c. Chebyshev's distance
d. Manhattan distance

Answer = a . eucledian distance

5. ______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.

a. Non-hierarchical clustering
b. Divisive clustering
c. Agglomerative clustering
d. K-means clustering

Answer = Agglomerative clustering

6. Which of the following is required by K-means clustering?

a. defined distance metric
b. number of clusters
c. initial guess as to cluster centroids
d. all answers are correct

answer = all answers are correct

7. The goal of clustering is to-

a. Divide the data points into groups
b. Classify the data point into different classes
c. Predict the output values of input data points
d. All of the above

Answer = a . Divide the data points into groups

8. Clustering is a-

a. Supervised learning
b. Unsupervised learning
c. Reinforcement learning
d. None

Answer – b. unsupervised learning

9. Which of the following clustering algorithms suffers from the problem of convergence at local optima?

a. K- Means clustering
b. Hierarchical clustering
c. Diverse clustering
d. All of the above

Answer = d. All of the above

10. Which version of the clustering algorithm is most sensitive to outliers?

a. K-means clustering algorithm
b. K-modes clustering algorithm

c. K-medians clustering algorithm
d. None

Answer a. K-means clustering

11. Which of the following is a bad characteristic of a dataset for clustering analysis-

a. Data points with outliers
b. Data points with different densities
c. Data points with non-convex shapes
d. All of the above

Answer = d. All of the above

12. For clustering, we do not require-

a. Labeled data
b. Unlabeled data
c. Numerical data
d. Categorical data

Answer = a. Labeled Data

Q13 to Q15 are subjective answers type questions, Answers them in their own words briefly.

13. How is cluster analysis calculated?

Answer= The hierarchical cluster analysis follows three basic steps: 1) calculate the distances, 2) link the clusters, and 3) choose a solution by selecting the right number of clusters. First, we have to select the variables upon which we base our clusters

14. How is cluster quality measured?

Answer = Here you have a couple of measures, but there are many more: SSE: sum of the square error from the items of each cluster. Inter cluster distance: sum of the square distance between each cluster centroid.
Intra cluster distance for each cluster: sum of the square distance from the items of each cluster to its centroid.

15. What is cluster analysis and its types?

Answer = Cluster analysis is the task of grouping a set of data points in such a way that they can be characterized by their relevancy to one another. ... These types are Centroid Clustering, Density Clustering Distribution Clustering, and Connectivity Clustering

MACHINE LEARNING – WORKSHEET (CLUSTERING)

Q1 to Q12 have only one correct answer. Choose the correct option to answer your question.

1. Movie Recommendation systems are an example of:
1. Classification
2. Clustering
3. Reinforcement Learning
4. Regression

Options:
- a. 2 Only
- b. 1 and 2
- c. 1 and 3
- d. 2 and 3
- e. 1, 2 and 3
- f. 1, 2, 3 and 4

Answer = d. 2 and 3

2. Sentiment Analysis is an example of:
1. Regression
2. Classification
3. Clustering
4. Reinforcement Learning

- a. 1 Only
- b. 1 and 2
- c. 1 and 3
- d. 1, 2 and 3
- e. 1, 2 and 4
- f. 1, 2, 3 and 4

Answer = e. 1, 2 and 4

3. Can decision trees be used for performing clustering?
- a. True
- b. False

Answer = a. True

4. Which of the following is the most appropriate strategy for data cleaning before performing clustering analysis, given less than desirable number of data points:
- a. Capping and flooring of variables
- b. Removal of
outliers Options:
  - a. 1 only
  - b. 2 only

c. 1 and 2

d. None of the above

Answer =a. Capping and flooring of variables

5. What is the minimum no. of variables/ features required to perform clustering?

a. 0
b. 1
c. 2
d. 3

Answer = b. 1

6. For two runs of K-Mean clustering is it expected to get same clustering results?

a. Yes
b. No

Answer = b. NO

7. Is it possible that Assignment of observations to clusters does not change between successive iterations in K-Means

a. Yes
b. No
c. Can't say
d. None of these

Answer = a. Yes

8. Which of the following can act as possible termination conditions in K-Means?

1. For a fixed number of iterations.
2. Assignment of observations to clusters does not change between iterations. Except for cases with a bad local minimum.
3. Centroids do not change between successive iterations.
4. Terminate when RSS falls below a threshold.

Options:

a. 1, 3 and 4
b. 1, 2 and 3
c. . 1, 2 and 4
d. . All of the above

Answer = d. All of the above

9. Which of the following can act as possible termination conditions in K-Means?

1. K- Means clustering algorithm
2. Agglomerative clustering algorithm
3. Expectation-Maximization clustering algorithm
4. Diverse clustering algorithm

a. 1 only
b. 2 and 3
c. 2 and 4
d. 1 and 3
e. 1,2 and 4
f. All of the above

Answer = All of the above

10. Which of the following algorithms is most sensitive to outliers?

a. K-means clustering algorithm
b. K-medians clustering algorithm
c. K-modes clustering algorithm
d. K-medoids clustering algorithm

Answer = K – Means Clustering

11. How can Clustering (Unsupervised Learning) be used to improve the accuracy of Linear Regression model (Supervised Learning):

1. Creating different models for different cluster groups.
2. Creating an input feature for cluster ids as an ordinal variable.
3. Creating an input feature for cluster centroids as a continuous variable.
4. Creating an input feature for cluster size as a continuous variable.

a. 1 only
b. 1 and 2
c. 1 and 4
d. 3 only
e. 2 and 4
f. All of the above

Answer = f. All of the above

12. What could be the possible reason(s) for producing two different dendrograms using agglomerative clustering algorithms for the same dataset?

a. Proximity function used
b. of data points used
c. of variables used
d. B and c only
e. All of the above

Answer = e. All of the above

Q13 to Q15 are subjective answers type questions, Answers them in their own words briefly

13. Is K sensitive to outliers?

Answer = The K-means clustering algorithm is sensitive to outliers, because a mean is easily influenced by extreme values. ... Mean is greatly influenced by the outlier and thus cannot represent the correct cluster center, while medoid is robust to the outlier and correctly represents the cluster center.

14. Why is K means better?

Answer = Other clustering algorithms with better features tend to be more expensive. In this case, k-means becomes a great solution for pre-clustering, reducing the space into disjoint smaller sub-spaces where other clustering algorithms can be applied

15. Is K means a deterministic algorithm?

Answer = The basic k-means clustering is based on a non-deterministic algorithm. This means that running the algorithm several times on the same data, could give different results. However, to ensure consistent results, FCS Express performs k-means clustering using a deterministic method

WORKSHEET

SQL

Q1 and Q2 have one or more correct answer. Choose all the correct option to answer your question.

1. Which of the following is/are DDL commands in SQL?

   A) Create   B) Update

   C) Delete   D) ALTER

   Answer = Create and Alter

2. Which of the following is/are DML commands in SQL?

   A) Update   B) Delete

   C) Select   D) Drop

   Answer = update and delete

3. Full form of SQL is:

   A) Strut querying language

   C) Simple Query Language   B) Structured Query Language

   D) None of the

   Answer = Structured query language

4. Full form of DDL is:

   A) Descriptive Designed Language

   C) Data Descriptive Language   B) Data Definition Language

   D) None of the above

   Answer = Data Definition Language

5. DML is:

   A) Data Manipulation Language

   C) Data Modeling Language   B) Data Management Language

   D) None of these

   Answer = Data Manipulative language

6. Which of the following statements can be used to create a table with column B int type and C float type?

   A) Table A (B int, C float)

   C) Create Table A (B int,C float)   B) Create A (b int, C float)

   D) All of them

   Answer= Create Table A (B int,C float)

7. Which of the following statements can be used to add a column D (float type) to the table A created above?

A) Table A (D float)
C) Table A(B int, C float, D float)
B) Alter Table A ADD COLUMN D float
D) None of them

Answer = Alter Table A ADD COLUMN D float

8. Which of the following statements can be used to drop the column added in the above question?

A) Table A Drop D
C) Delete D from A
B) Alter Table A Drop Column D
D) None of them

Answer = Alter Table A Drop Column D

9. Which of the following statements can be used to change the data type (from float to int) of the column D of
table A created in above questions?

A) Table A (D float int)
C) Alter Table A D float int
B) Alter Table A Alter Column D int
D) Alter table A Column D float to int

Answer = Alter table A Column D float to int

10. Suppose we want to make Column B of Table A as primary key of the table. By which of the following statements we can do it?

A) Alter Table A Add Constraint Primary Key B
B) Alter table (B primary key)
C) Alter Table A Add Primary key B
D) None of them

Answer = Alter Table A Add Constraint Primary Key B

Q11 to Q15 are subjective answer type questions, Answer them briefly.

11. What is data-warehouse?

Answer = A data warehouse is a type of data management system that is designed to enable and support business intelligence (BI) activities, especially analytics. Data warehouses are solely intended to perform queries and analysis and often contain large amounts of historical data

12. What is the difference between OLTP VS OLAP?

Answer = OLTP and OLAP both are the online processing systems. OLTP is a transactional processing while OLAP is an analytical processing system. ... The basic difference between OLTP and OLAP is that OLTP is an online database modifying system, whereas, OLAP is an online database query answering system

13. What are the various characteristics of data-warehouse?

Answer = The key characteristics of a data warehouse are as follows:

Some data is denormalized for simplification and to improve performance.

Large amounts of historical data are used.

Queries often retrieve large amounts of data.

Both planned and ad hoc queries are common.

The data load is controlled

14. What is Star-Schema??

Answer = In computing, the star schema is the simplest style of data mart schema and is the approach most widely used to develop data warehouses and dimensional data marts. The star schema consists of one or more fact tables referencing any number of dimension tables

15. What do you mean by SETL?

Answer= SETL (SET Language) is a very high-level programming language based on the mathematical theory of sets.

WORKSHEET

SQL

Q1 to Q13 have only one correct answer. Choose the correct option to answer your question.

1. Which of the following constraint requires that there should not be duplicate entries?
   A) No Duplicity B) Different
   C) Null D) Unique

   Answer = Unique

2. Which of the following constraint allows null values in a column?

   A) Primary key B) Empty Value C) Null D) None of them

Answer = Primary key

3. Which of the following statements are true regarding Primary Key?
   A) Each entry in the primary key uniquely identifies each entry or row in the table
   B) There can be duplicate values in a primary key column
   C) There can be null values in Primary key
   D) None of the above.

Answer = There can be null values in primary key

4. Which of the following statements are true regarding Unique Key?
   A) There should not be any duplicate entries
   B) Null values are not allowed
   C) Multiple columns can make a single unique key together
   D) All of the above

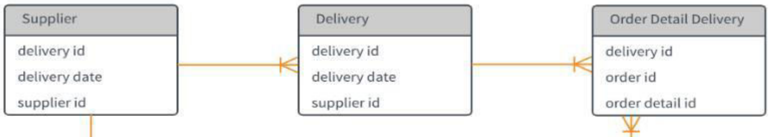
Answer = All of the above

5. Which of the following is/are example of referential constraint?

   A) Not Null B) Foreign Key C) Referential key
   D) All of them

   Answer=Foreign key

For Questions 6-13 refer to the below diagram and answer the questions:



6. How many foreign keys are there in the Supplier table?
   A) 0 B) 3
   C) 2 D) 1

   Answer= 2

7. The type of relationship between Supplier table and Product table is:

   A) one to many B) many to one C) one to one D) many to many

   Answer = many to one

8. The type of relationship between Order table and Headquarter table is:
   A) one to many B) many to one

C) one to one D) many to many

9. Which of the following is a foreign key in Delivery table?

Answer = One to one

A) delivery id B) supplier id C) delivery date D) None of them

10. The number of foreign keys in order details is:
A) 0 B) 1
C) 3 D) 2

Answer = 3

11. The type of relationship between Order Detail table and Product table is:

A) one to many B) many to one C) one to one D) many to many

Answer = One to one

12. DDL statements perform operation on which of the following database objects?

A) Rows of table B) Columns of table C) Table D) None of them

13. Which of the following statement is used to enter rows in a table?
A) Insert in to B) Update
C) Enter into D) Set Row

Answer = insert into

Q14 and Q15 have one or more correct answer. Choose all the correct option to answer your question.

14. Which of the following is/are entity constraints in SQL?
A) Duplicate B) Unique
C) Primary Key D) Null

Answer = Primary key

15. Which of the following statements is an example of semantic Constraint?
A) A blood group can contain one of the following values - A, B, AB and O.
B) A blood group can only contain characters
C) A blood group cannot have null values
D) Two or more donors can have same blood group

Answer = A blood group can contain one of the following values - A, B, AB and O

Statistics– WORKSHEET

Q1 to Q9 have only one correct answer. Choose the correct option to answer your question.

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 = Central 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 = All of the mentioned

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 = all of the mentioned

5. random variables are used to model rates.
   a) Empirical
   b) Binomial
   c) Poisson
   d) All of the mentioned

Answer = Poisson

6. Usually replacing the standard error by its estimated value does change the CLT.
   a) True
   b) False

Answer = 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 = 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 = Outliers cannot conform to the regression relationship

Q10and Q15 are subjective answer type questions, Answer them in your own words briefly.

10. What do you understand by the term Normal Distribution?
    Answer = The normal distribution is a probability function that describes how the values of a variable are distributed. It is a symmetric distribution where most of the observations cluster around the central peak and the probabilities for values further away from the mean taper off equally in both directions

11. How do you handle missing data? What imputation techniques do you recommend?
    Answer: We can handle missing data using imputer functions . We can either drop the rows and columns completely if the missing values are very less. Or we can impute them with mean, median or mode using numpy. It depends on the type of data we are presented with. Knn imputation is the imputation technique which can be used majorly.

12. What is A/B testing?
    Answer = 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 a statistically significant relationship or not.

13. Is mean imputation of missing data acceptable practice?
    Answer = No

13. What is linear regression in statistics?

Answer = Linear regression attempts to establish the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable.

14. What are the various branches of statistics?

Answer = Descriptive and Inferential

STATISTICS – WORKSHEET

Q1 to Q15 have only one correct answer. Choose the correct option to answer your question.

1. What represent a population parameter?
   A) SD
   B) mean
   C) both
   D) none\

Answer = both

2. What will be median of following set of scores (18,6,12,10,15)?
   A) 14
   B) 18
   C) 12
   D) 10

Answer = 12

3. What is standard deviation?
   A) An approximate indicator of how number vary from the mean
   B) A measure of variability
   C) The square root of the varience
   D) All of the above

Answer = All of the above

4. The intervals should be in a grouped frequency distribution
   A) Exhaustive
   B) Mutually exclusive
   C) Both of these
   D) None

Answer = Both A and B

5. What is the goal of descriptive statistics?
   A) Monitoring and manipulating a specific data
   B) Summarizing and explaining a specific set of data
   C) Analyzing and interpreting a set of data
   D) All of these

Answer = All of these

6. A set of data organized in a participant by variables format is called
   A) Data junk
   B) Data set
   C) Data view
   D) Data dodging

Answer = Dataset

7. In multiple regression, dependent variables are used
   A) 2 or more

B) 2
C) 1
D) 1 or more

Answer = 1

8. Which of the following is used when you want to visually examine the relationship between 2 quantitative variables?
A) Line graph
B) Scatterplot
C) Bar graph
D) Pie graph

Answer = Lineplot

9. Two or more groups means are compared by using_
A) analysis
B) Data analysis

C) Varied Variance analysis
D) Analysis of variance

Answer = Analysis of variance

10. is a raw score which has been transformed into standard deviation units?
A) Z-score
B) t-score
C) e-score
D) SDU score

Answer = Z score

11. is the value calculated when you want the arithmetic average?
A) Median
B) mode
C) mean
D) All

Answer = mean

12. Find the mean of these set of number (4,6,7,9,2000000)?
A) 4
B) 7
C) 7.5
D) 400005.2

Answer = 400005.2

13. is a measure of central tendency that takes into account the magnitude of scores?
A) Range
B) Mode
C) Median
D) Mean

Answer = Mean

14. Is focuses on describing or explaining data wheras involves going beyond immediate data and making inferences
A) Descriptive and inferences
B) Mutually exclusive and mutually exhaustive properties
C) Positive skew and negative skew
D) Central tendency

Answer = Descriptive and inferences

15. What is the formula for range?
A) H+L
B) L-H
C) LXH
D) H-L

Answer = H-L