FLIP ROBO TECHNOLOGIES

Worksheet Set – 2

Batch DS2301

Intern: Vishnukanth

Mail ID: vishnukh25@gmail.com

- ✓ **Machine Learning**
- ✓ **SQL**
- ✓ **Statistics**

Machine Learning

Worksheet – 2

Q1 to Q11 have only one correct answer. Choose the correct option to answer your question.

1. Movie Recommendation systems are an example of:

i) Classification

ii) Clustering

iii) Regression Options:

a) 2 Only

b) 1 and 2

c) 1 and 3

d) 2 and 3

Answer : b) 1 and 2

2. Sentiment Analysis is an example of:

i) Regression

ii) Classification

iii) Clustering

iv) Reinforcement Options:

a) 1 Only

b) 1 and 2

c) 1 and 3

d) 1, 2 and 4

Answer : b) 1 and 2

3. Can decision trees be used for performing clustering?

a) True b) False

Answer : b) False

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:

i) Capping and flooring of variables

ii) Removal of outliers Options:

a) 1 only

b) 2 only

c) 1 and 2

d) None of the above

Answer : d) None of the above

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

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

Answer : c) 2

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?

i) For a fixed number of iterations.

ii) Assignment of observations to clusters does not change between iterations. Except for cases witha bad local minimum.

iii) Centroids do not change between successive iterations.

iv) 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 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 : a) K-means clustering algorithm

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

i) Creating different models for different cluster groups.

ii) Creating an input feature for cluster ids as an ordinal variable.

iii) Creating an input feature for cluster centroids as a continuous variable.

iv) Creating an input feature for cluster size as a continuous variable. Options:

a) 1 only

b) 2 only

c) 3 and 4

d) All of the above

Answer : d) All of the above

11. 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) All of the above

Answer : d) All of the above

Q12 to Q14 are subjective answers type questions, Answers them in their own words briefly

12. Is K sensitive to outliers?

K-means is sensitive to outliers. K-means algorithm tries to minimize the variance within each cluster, it does this by computing the mean of all the points in each cluster. Outliers can have a big impact on the mean of a cluster, which can cause the cluster centroid to be shifted towards the outlier. As a result, the cluster boundaries can be affected, and other points that should be in the same cluster may be assigned to different clusters, which can lead to poor clustering results.

13. Why is K means better?

K-means is a popular clustering algorithm for several reasons:

Ease of implementation: K-means is a relatively simple algorithm that is easy to implement and understand.
Efficiency: K-means has a linear time complexity with respect to the number of data points and is relatively efficient for large datasets.
Scalability: K-means can be easily parallelized, making it well-suited for large-scale datasets.
Versatility: K-means can be used for a variety of types of data, including continuous and categorical variables.
Well-understood behaviour: The behaviour of the K-means algorithm is well-understood, which makes it easy to interpret the results of the clustering.
It assumes spherical cluster shape, it works well when the clusters are spherical in shape, meaning that all points in the cluster are closer to the centroid than to any other point in the cluster.

14. Is K means a deterministic algorithm?

K-means is a deterministic algorithm, which means that it will always produce the same result when run on the same dataset with the same initial conditions. This is because the algorithm follows a set of fixed steps to assign each data point to a cluster.

However, the initial conditions play a crucial role in the final result, if the initial centroids are chosen randomly, the algorithm may converge to different solutions on different runs, which are known as local optima. Therefore, to ensure reproducibility and to avoid the problem of local optima, a common practice is to run the algorithm multiple times with different initial centroids and choose the best solution.

Another common practice is to use some form of initialization methods like K-means++ which is a technique that addresses this problem by carefully selecting the initial centroids to be far from each other and it's proven to converge faster and produce a better solution.

Structured Querying Language

Worksheet – 2

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 : d) 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 : c) Null

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 : a) Each entry in the primary key uniquely identifies each entry or row in the table

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 : d) All of the above

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

A) Not Null

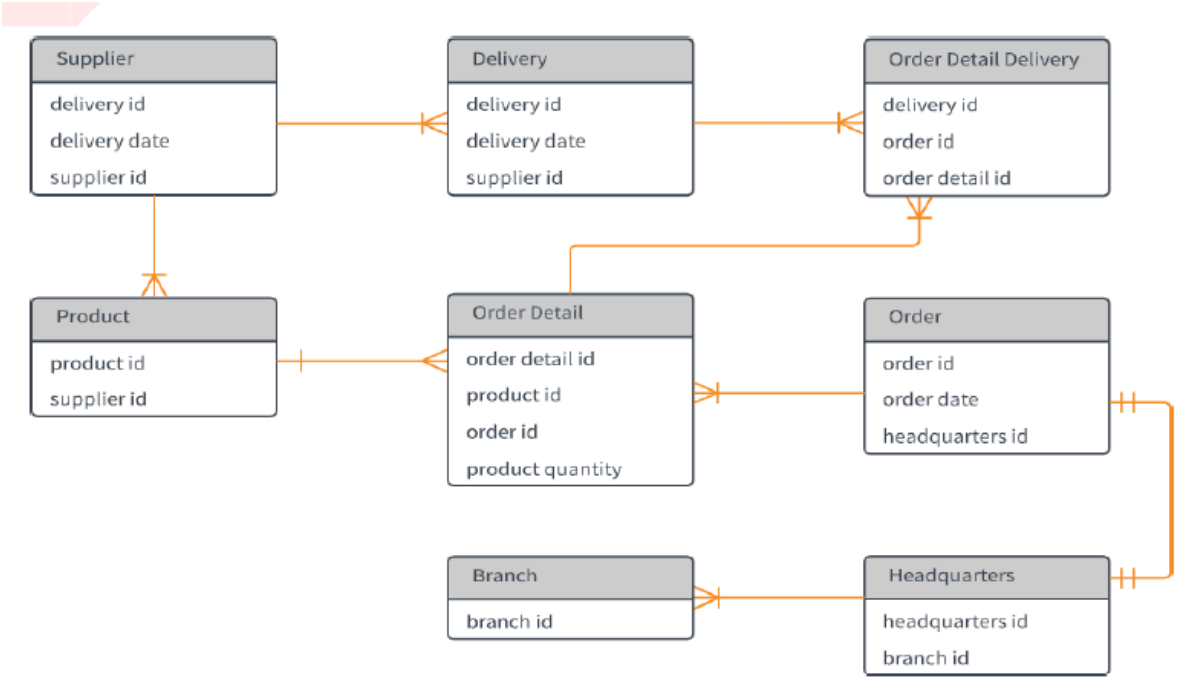
B) Foreign Key

C) Referential key

D) All of them

Answer : b) 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 : d) 1

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 : d) one to many

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

Answer : c) one to one

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

A) delivery id

B) supplier id

C) delivery date

D) None of them

Answer : b) supplier id

10. The number of foreign keys in order details is:

A) 0 B) 1 C) 3 D) 2

Answer : d) 2

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 : b) many 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

Answer : c) Table

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 : a) Insert in to

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 : b) Unique and **c) 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) A blood group can contain one of the following values - A, B, AB and O.

STATISTICS

Worksheet – 2

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 : c) 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 : c) 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 variance

D) All of the above

Answer : d) 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 : b) Mutually Exclusive

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 : d) 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 : b) Data set

7. In multiple regression,_____ independent variables are used

A) 2 or more

B) 2

C) 1

D) 1 or more

Answer : a) 2 or more

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 : b) Scatterplot

9. Two or more groups means are compared by using

A) analysis

B) Data analysis

C) Varied Variance analysis

D) Analysis of variance

Answer : d) 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 : a) Z-score

11. _____is the value calculated when you want the arithmetic average?

A) Median

B) mode

C) mean

D) All

Answer : c) 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 : d) 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 : d) Mean

14. _____ focuses on describing or explaining data whereas _____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 : a) Descriptive and inferences

15. What is the formula for range?

A) H+L

B) L-H

C) LXH

D) H-L

Answer : d) H – L