FLIP ROBO

NAME OF THE PROJECT

WORKSHEET-2 ON MACHINE LEARNING SQL AND STATISTICS

SUBMITTED BY:

VARSHA.V.SHINDE

WORKSHEET-2 ML

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

Q1. 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:-Only 2

Q2. 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: - D 1, 2 and 4

Q3. Can decision trees be used for performing clustering?

a) True
b) False
ANSWER:-True

Q4. 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: - 1 only

Q5. What is the minimum no. of variables/ features required to perform clustering?

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

ANSWER: - b) 1

Q6. For two runs of K-Mean clustering is it expected to get same clustering results?

a) Yes
b) No
ANSWER: - b) No

Q7. 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
ANSWER: - a) Yes

Q8. 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 with a 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

Q9. 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

Q10. 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

Q11. 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

Q12. Is K sensitive to outliers?
ANSWER:-

- The K-mean algorithm updates the cluster centres by taking the average of all the data points that are closer to each cluster centre.
- When all the points are packed nicely together, the average make sense.however, when you have outliers, this can affects the average calculation of the whole cluster.
- As a result; this will push your cluster centre closer to the outlier.

Q13. Why K is means better?
ANSWER: -
K- Means is a technique for data clustering that may be used for unsupervised machine learning. It is capable of classifying unlabeled data into a predetermined number of clusters based on similarities (k)

Advantages of K-means:-

- Relatively simple to implement.
- Scales to Large data sets
- Guarantees convergence.
- Can warm-start the positions of centroids.
- Easily adapts to new examples.4r4545
- Generalizes to cluster of different shapes and sizes, such as elliptical clusters.

Q14. Is K means a deterministic?
ANSWER:-

- Clustering Algorithms with steps involving randomness usually give different results on different executions for the same dataset.
- This non-deterministic nature of algorithms such as the K-Means clustering algorithm limits their applicability in areas such as cancer subtype prediction using gene expression data.
- It is hard to sensibly compare the results of such algorithms with those of other algorithms.
- The non-deterministic nature of K-Means is due to its random selection of data points as *initial centroids*.

The main goal of the K-Means algorithm is to minimize the sum of distances between the points and their respective cluster centroid. Also, to assign each data point to its closet k-centre. Points that are near to a particular K centre create a cluster.

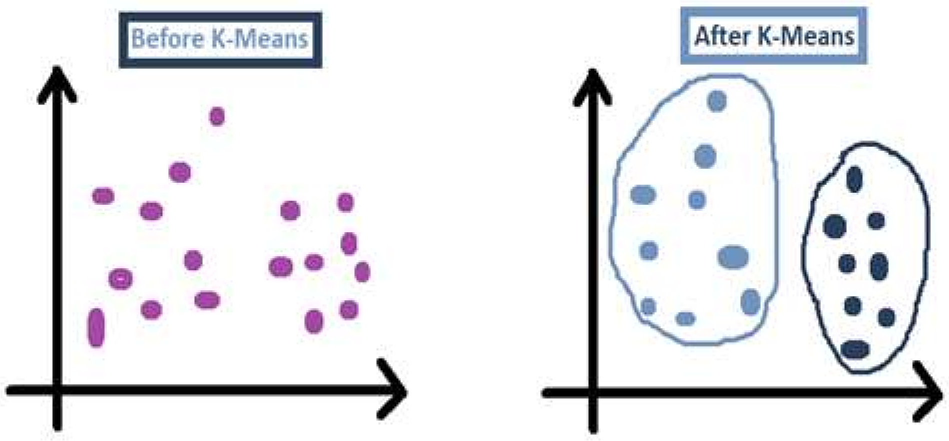
- It is an **iterative algorithm** that divides the unlabeled datasets into k clusters so that each data set belongs only to one group with similar properties.

- It is a **centroid based algorithm**, where each cluster is associated with a centroid.

- Here K defines the number of pre-defined clusters that need to be created. If we have K=2, then there will be 2 clusters or 2 different categories, if K=3, there will be 3 clusters.

- The value of K is predetermined in this algorithm.

- This algorithm takes input as an unlabeled dataset, divides the dataset into K-number of clusters, and repeats the process until we find the best cluster.



WORKSHEET-2 STATISTICS

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

Q1. What represent a population parameter?

A) SD
B) Mean
C) Both
D) None
ANSWER: - C) Both

Q2. What will be median of following set of scores (18, 6, 12, 10, and 15)?

A) 14
B) 18
C) 12
D) 10
ANSWER:- C) 12

Q3. What is standard deviation?

A) An approximate indicator of how number varies 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

Q4. The intervals should be ______ in a grouped frequency distribution

A) Exhaustive
B) Mutually exclusive
C) Both of these
D) None
ANSWER: - C) Both of these

Q5. 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: - B) Summarizing and explaining a specific set of data

Q6. 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

Q7. In multiple regressions, _______ independent variables are used

A) 2 or more
B) 2
C) 1
D) 1 or more
ANSWER: - C) 1

Q8. Which of the following is used when you want to visually examine the relationship between 2 quantitative variables?

A) Line graph
B) Scatter plot
C) Bar graph
D) Pie graph
ANSWER: - B) Scatter plot

Q9. Two or more group's means are compared by using

A) Analysis
B) Data analysis
C) Varied Variance analysis
D) Analysis of variance
ANSWER: - D) Analysis of variance

Q10. _______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

Q11. _______is the value calculated when you want the arithmetic average?

A) Median
B) Mode
C) Mean
D) All
ANSWER: - C) Mean

Q12. 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

Q13. _______ 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

Q14. ______ 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

Q15. What is the formula for range?

A) H+L
B) L-H
C) LXH
D) H-L
ANSWER: **- D) H-L**

WORKSHEET-2 SQL

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

Q1. 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

Q2. Which of the following constraint allows null values in a column?

A) Primary key
B) Empty Value
C) Null
D) None of them
ANSWER: - D) none of them

Q3. 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: - None of the above

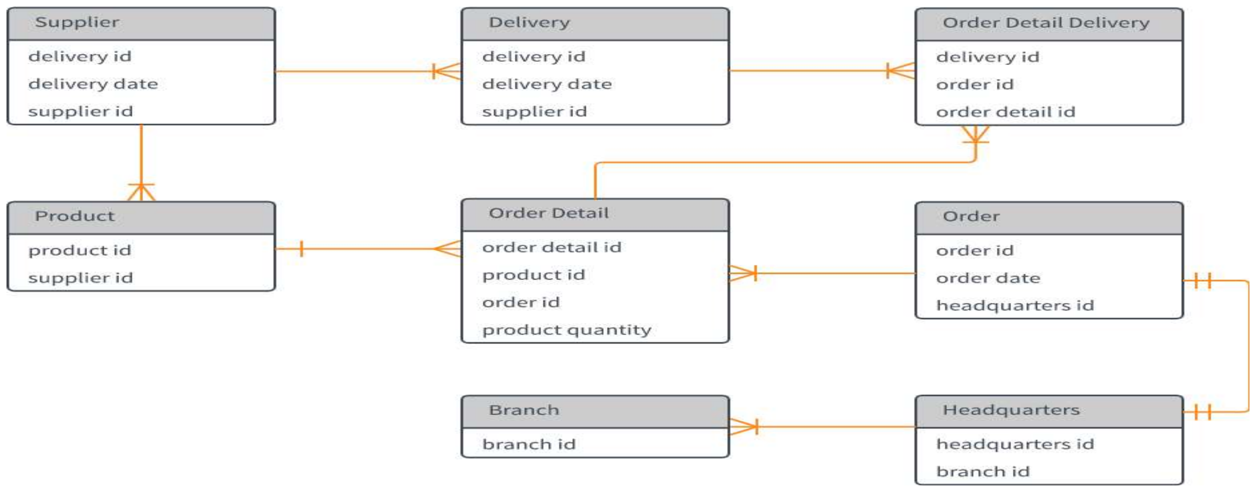
Q4. Which of the following statements are true regarding Unique Key?

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:-

Q5. Which of the following is/are example of referential constraint?

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

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



Q6. How many foreign keys are there in the Supplier table?

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

ANSWER: - D) 1

Q7. The type of relationship between Supplier table and Product table is:

A) One too many
B) Many to one
C) one to one
D) Many too many

ANSWER: - C) one to one

Q8. 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 too many

ANSWER: - One to many

Q9. 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

Q10. The number of foreign keys in order details is:

A) 0
B) 1
C) 3
D) 2
ANSWER:-

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

A) One too many B) Many to one

C) one to one D) Many too many
ANSWER:-

Q12. 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:-

Q13. 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:-

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

Q14. Which of the following is/are entity constraints in SQL?

A) Duplicate
B) Unique
C) Primary Key
D) Null
ANSWER:-2, 3 and 4

Q15. 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: - 1 and 2