MACHINE LEARNING

Assignment-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

Ans :- a) 2 Only - ii) Clustering

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

Ans :- d) 1, 2 and 4 - i) Regression ,ii) Classification & iv) Reinforcement

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

Ans :- 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:
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

Ans :- a) 1 only - i) 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

Ans :- b) 1

6. For two runs of K-Mean clustering is it expected to get same clustering results?
a) Yes
b) No

Ans :- 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

Ans :- 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 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

Ans :- 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

Ans :- 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

Ans :- d) All of the above

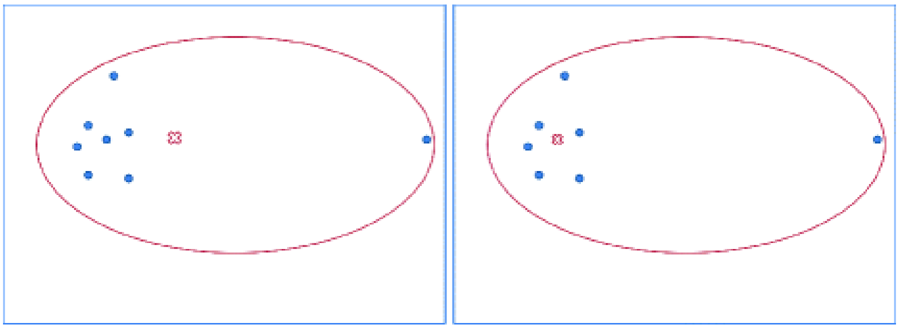
11. What could be the possible reason(s) for producing two different dendrograms using agglomerative clustering algorithms for the same dataset?
a) Proximity function used
b) of data points used
c) of variables used
d) All of the above

Ans :- 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?

Ans :- Yes. The *K*-means clustering algorithm is sensitive to outliers, because a mean is easily influenced by extreme values. *K*-medoids clustering is a variant of *K*-means that is more robust to noises and outliers. Instead of using the mean point as the center of a cluster, *K*-medoids uses an actual point in the cluster to represent it. Medoid is the most centrally located object of the cluster, with minimum sum of distances to other points. Figure 1 shows the difference between mean and medoid in a 2-D example. The group of points in the right form a cluster, while the rightmost point is an outlier. 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.



(a) Mean (b) Medoid

13. Why is K means better?

Ans :- K-means clustering algorithm can be significantly improved by using a better initialization technique, and by repeating (re-starting) the algorithm.

When the data has overlapping clusters, k-means can improve the results of the initialization technique.

When the data has well separated clusters, the performance of k-means depends completely on the goodness of the initialization.

Initialization using simple furthest point heuristic (Maxmin) reduces the clustering error of k-means from 15% to 6%, on average.

14. Is K means a deterministic algorithm?

Ans :- No. K-Means is a non-deterministic algorithm. This means that a compiler cannot solve the problem in polynomial time and doesn’t clearly know the next step. This is because some problems have a great degree of randomness to them. These algorithms usually have 2 steps — 1) Guessing step 2) Assignment step. On similar lines is the K-means algorithm. The K-Means algorithm divides the data space into K clusters such that the total variance of all data points with respect to the cluster mean is minimized.

K-means Clustering is undoubtedly one of the most popular unsupervised learning algorithms. The reason behind it being used so frequently is the strong yet simple statistical backbone. This story would first explain the logical approach behind K-Means clustering, then it would bring forth a practical drawback and a few suggestions to avoid it.

WORKSHEET 2 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

Ans :- 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

Ans :- 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.

Ans :- A) Each entry in the primary key uniquely identifies each entry or row in the table

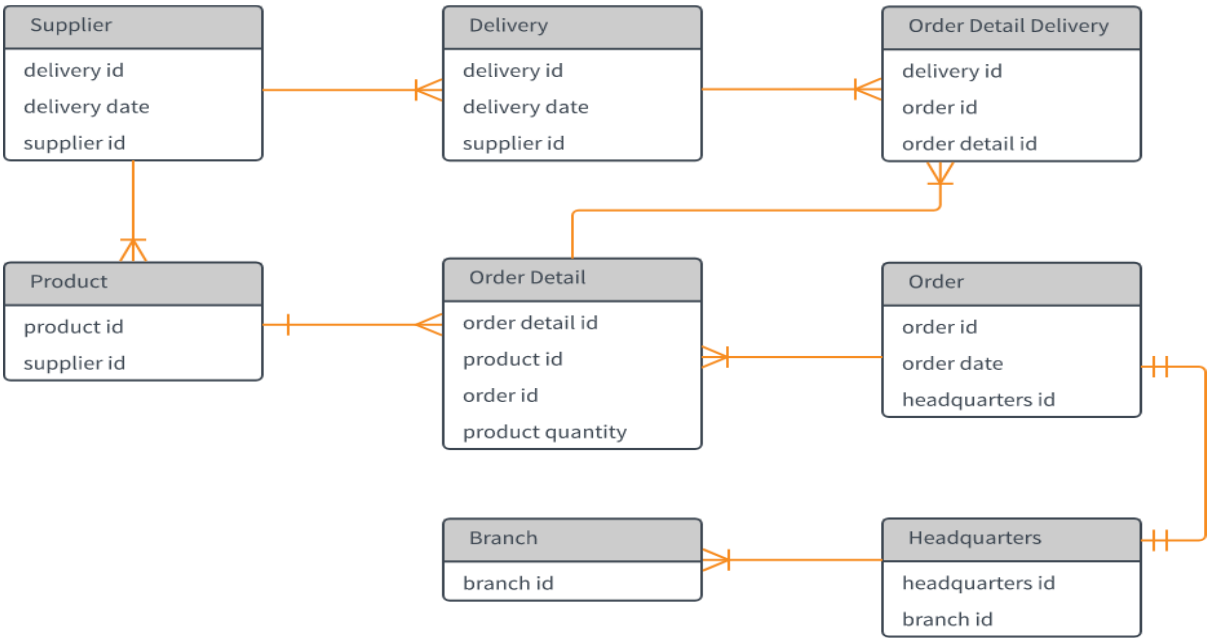
4. Which of the following statements are true regarding Unique Key?
A) There should not be any duplicate entries
B) Null values are not allowed
C) Multiple columns can make a single unique key together
D) All of the above

Ans :- C) Multiple columns can make a single unique key together

5. Which of the following is/are example of referential constraint?
A) Not Null B) Foreign Key
C) Referential key D) All of them

Ans :- 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

Ans :- B) 3

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

Ans :- A) 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

Ans :- D) many to many

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

Ans :- B) supplier id

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

Ans :- 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

Ans :- A) one to many

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

Ans :- 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

Ans :- B) Update

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

Ans :- B) Unique & 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

Ans :- A) A blood group can contain one of the following values - A, B, AB and O.
& B) A blood group can only contain characters

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

Ans :- C) both [A) SD & B) mean]

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

Ans :- 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

Ans :- 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

Ans :- C) Both of these

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

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

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

Ans :- B) Data set

7. In multiple regression,_______ independent variables are used
A) 2 or more
B) 2
C) 1
D) 1 or more

Ans :- B) 2

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

Ans :- 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

Ans :- 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

Ans :- A) Z-score

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

Ans :- 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

Ans :- 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

Ans :- 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

Ans :- A) Descriptive and inferences

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

Ans :- D) H-L