MCQs Unit-I Solution

1. Dispersion of the data can be analyzed with
a. mean
b. median
c. mode
d. Standard deviation
Ans:d
Explanation: variance is used to compute the data dispertion. Variance is square of standard deviation

2. Most frequency data item is computed in terms of
a. mean
b. median
c. mode
d. Standard deviation
Ans: c
Explanation: Mode shows most frequent item in dataset

3. Calculate median of 100, 300, 450, 650, 779
a. 450
b. 455.8
c. 650.6
d. 400.8
ans: a
Explanation: Median for odd data is middle value

4. Calculate median of 100, 300, 450, 650, 800, 900
a. 450
b. 455.8
c. 650
d. 550
ans: d
Explanation: Median for even data is average of two middle value

5. Calculate median of 450, 300, 100, 650, 800
a. 450
b. 455.8
c. 650
d. 550
ans: a
Explanation: Median is middle value of sorted data

6. Calculate five number summary of 100, 200, 350, 650, 800, first quartiles Q1
a. 100

b. 200
c. 450
d. 650
Ans: b

Explanation: Q1 (25th) median of first(lower) partition

7. Calculate five number summary of 100, 200, 350, 650, 800, first quartiles Q1
a. 100
b. 200
c. 450
d. 650
Ans: b

Explanation: Q3 (75th) median of last(upper) partition

8. Five number summary is
a. min, Q1, Q3, mid, max
b. min, Q1, median,Q3, max
c. min, Q1, Q2 ,Q3, max
d. Q1, Q2, min, max, median
Ans: b

Explanation: A **summary** consists of **five** values: the most extreme values in the data set (the maximum and minimum values), the lower and upper quartiles, and the median.

9. Calculate standard deviation for 100, 300, 450, 650, 779
a. 242.1
b. 455.8
c. 650.6
d. 58614.56

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}.$$

A ns: a
Explanation: As per the formula

10. Calculate mean of 100, 300, 450, 650, 779
a. 450
b. 455.8
c. 650.6
d. 400.8
ans: b
Explanation: Mean is average of elements

11. Normalize the given data 100, 300, 450, 650, 779 with z-score . Value for 100 is.. [4 mks]
a. -1.469
b. 0.1
c. -0.643
d. 0.3
ans: a
Explanation: as per the formula v-mean/standard deviation

12. Normalize the given data 100, 300, 450, 650, 779 with z-score . Value for 300 is.. [4 mks]
a. -1.469
b. 0.1
c. -0.643
d. 0.3
ans: c
Explanation: as per the formula v-mean/standard deviation

13. Normalize the given data 100, 300, 450, 650, 779 with decimal scaling . Value for 300 is..
a. -1.469
b. 0.1
c. -0.643
d. 0.3
ans: d
Explanation: as per the formula v/10j

14. Normalize the give data 100, 300, 450, 650, 779 with decimal scaling . Value for 100 is..
a. -1.469
b. 0.1
c. -0.643
d. 0.3
ans: b
Explanation: as per the formula v/10j

15. Normalize the given data 100, 300, 450, 650, 800 with min_max into range [0,1]. Value for 100 is..
a. -1.469
b. 0.1
c. 0
d. 0.3
ans: c
Explanation: as per the formula v-old_min/old_max-old_min (n_max-n_min)+n_mean

16. Normalize the given data 100, 300, 450, 650, 800 with min_max into range [0,1]. Value for 800 is..

a. -1.469
b. 0.1
c. 0
d. 1
ans: d
Explanation: as per the formula v-old_min/old_max-old_min (n_max-n_min)+n_mean

17. Normalize the given data 100, 300, 450, 650, 800 with min_max into range [0,1]. Value for 300 is..
a. -1.469
b. 0.1
c. 0.285
d. 1
ans: c
Explanation: as per the formula v-old_min/old_max-old_min (n_max-n_min)+n_mean

18. Normalize the give data 100, 300, 450, 650, 800 with min_max into range [0,1]. Value for 300 is..
a. -1.469
b. 0.5
c. 0.285
d. 1
ans: b
Explanation: as per the formula v-old_min/old_max-old_min (n_max-n_min)+n_mean

19. In------- binning, bins have equal frequency
a. equal width
b. equal depth
Ans: b
Explanation: In equal Frequency/depth Binning : bins have equal number of data items

20. In------- binning, bins have equal intervals
a. equal width
b. equal depth
Ans: a
Explanation: **Equal Width Binning :** bins have equal width with a range of each bin are defined as [min + w], [min + 2w] …. [min + nw] where **w = (max – min) / (no of bins).**

21. The process of handling missing value, smoothing noise is called-------
a. data cleaning
b. data integration
c. data reduction
d data transformation
Ans: a

Explanation: Data cleaning (or data cleansing) routines attempt to fill in missing values, smooth out noise while identify-ing outliers, and correct inconsistencies in the data

22. If suppose student age is missing then fill it with
a. global constant
b. average
c. ignore tuple
d. none of above
Ans: b
Explanation: Most appropriate is fill the missing value with mean value

23. For bin1: 4 ,8, 15 smoothing by bin mean produce the result
a. 4,8,15
b. 4,4,15
c. 9,9,9
d. 8,8,8
Ans: c
Explain: In smoothing by bean mean values are replaced with bin mean

24. For bin1: 4 ,8, 15 smoothing by bin median produce the result
a. 4,8,15
b. 4,4,15
c. 9,9,9
d. 8,8,8
Ans:d
Explain: In smoothing by bean median values are replaced with bin median

25. For bin1: 4 ,8, 15 smoothing by bin boundary produce the result
a. 4,8,15
b. 4,4,15
c. 9,9,9
d. 8,8,8
Ans:b
Explain: In smoothing by bin boundary, values are replaced with closer boundary

26. The process of reducing volume of data is called-------
a. data cleaning
b. data integration
c. data reduction
d data transformation
Ans: c
Explanation: Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original

data.

27. The process of transforming data from one form to other is called-------
a. data cleaning
b. data integration
c. data reduction
d data transformation
Ans: d
Explanation: In data transformation, the data are transformed or consolidated into forms appropriate for mining

28. The process of combining volume of data is called-------
a. data cleaning
b. data integration
c. data reduction
d data transformation
Ans: b

Explanation: which combines data from multiple sources into a coherent data store, as in data warehousing. These sources may include multiple databases, data cubes, or flat files

29. Redundancy of data can be identified using---
a. chi-square test
b. normalization
c. Binning
d. None of above
Ans: a
Explanation: Chi-square test is useful for correlation analysis

30. Data can be transformed into required range by-----
a. chi-square test
b. normalization
c. Binning
d. None of above
Ans: b
Explanation: Normalization is technique to transform data from one format to other

31. ---- is noise smoothing technique
a. chi-square test
b. normalization
c. Binning
d. None of above
Ans: c
Explanation:Noise is smoothen by binning

32. Following are type of noise smoothing
a. Binning
b. regression
c. clustering
d. All are above
Ans: d
Explanation: All are noise handling techniques

33. Mean and standard deviation are required for---- normalization
a. z-score
b min-max
ans: a
Explanation: As per formula v-mean/SD

34. Data transformation includes _____.
A. a process to change data from a detailed level to a summary level.
B. A process to change data from a summary level to a detailed level.
C. joining data from one source into various sources of data.
D. separating data from one source into various sources of data.
ANSWER: A
Explanation: It is a transformation of data in required format

35. Dimensionality reduction reduces the data set size by removing _____.
A. relevant attributes.
B. irrelevant attributes.
C. derived attributes.
D. composite attribute
ANSWER: B
Explanation: It is a technique for volume reduction by removing unwanted information

36.The term that is not associated with data cleaning process is _____.
A. domain consistency.
B. deduplication.
C. disambiguation.
D. segmentation.
ANSWER: D
Explanation: Removing noise and redundancy, maintaining consistency are main tasks of data cleaning.

37. KDD is
a. Knowledge data discovery

b. Knowledge discovery in data
c. Knowledge diverse discovery
d. Knowledge diverse data
Answer: b
Explanation: It is the process of extracting knowledge from data
38) Data mining is
A) The actual discovery phase of a knowledge discovery process
B) The stage of selecting the right data for a KDD process
C) A subject-oriented integrated time variant non-volatile collection of data in support of management
D) None of these
Answer: A
Explanation: It is knowledge discovery process

39. Data selection is
A) The actual discovery phase of a knowledge discovery process
B) The stage of selecting the right data for a KDD process
C) A subject-oriented integrated time variant non-volatile collection of data in support of management
D) None of these
Answer: B
Explanation: As per the definition

40) Heterogeneous databases referred to
A) A set of databases from different vendors, possibly using different database paradigms
B) An approach to a problem that is not guaranteed to work but performs well in most cases.
C) Information that is hidden in a database and that cannot be recovered by a simple SQL query.
D) None of these
Answer: A
Explanation: collection of different databases

42) KDD (Knowledge Discovery in Databases) is referred to
A) Non-trivial extraction of implicit previously unknown and potentially useful information from data
B) Set of columns in a database table that can be used to identify each record within this table uniquely.
C) Collection of interesting and useful patterns in a database
D) none of these
Answer: A

Explanation:As per the definition

43) Which of the following is/are the Data mining tasks?
(a) Regression
(b) Classification
(c) Clustering
(d) All above
Answer: d
Explanation: Regression, Classification and Clustering are the data mining tasks.

44..... is a summarization of the general characteristics or features of a target class of data.
A) Data Characterization
B) Data Classification
C) Data discrimination
D) Data selection
Answer: A
Explanation: Data Characterization is the process of deriving characters from data

45) ..... is a comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes.
A) Data Characterization
B) Data Classification
C) Data discrimination
D) Data selection
Answer: C
Explanation: Data discrimination is comparison between two classes
46. The full form of KDD is .....
A) Knowledge Database
B) Knowledge Discovery in Database
C) Knowledge Data House
D) Knowledge Data Definition
Answer: B

47) The out put of KDD is .....
A) Data
B) Information
C) Query
D) Useful information

Answer: B

Explanation: extraction of knowledge/Useful information

48..... is an essential process where intelligent methods are applied to extract data patterns.

A) Data warehousing

B) Data mining

C) Text mining

D) Data selection

Answer: B

Explanation: It is the process of of KDD

49.In Binning, we first sort data and partition into (equal-frequency) bins and then which of the following is not a valid step Select one:

a. smooth by bin boundaries

b. smooth by bin median

c. smooth by bin means

d. smooth by bin values

Ans: d

Explanation: smooth by bin values is not the smoothing technique

50.Correlation analysis is used for :

a. handling missing values

b. identifying redundant attributes

c. handling different data formats

d. eliminating noise

Ans: b

Explanation: Correlation analysis is used for identifying redundant attributes

51.Which of the following is an Entity identification problem?

a. One person with different email address

b. One person’s name written in different way

c. Title for person

d. One person with multiple phone numbers

Ans: b

Explanation: Person can have different emails, phone numbers but name remains same and can be written in different ways.

MCQs on Unit-II

1. The data Warehouse is_____.
A. read only.
B. write only.
C. read write only.
D. none.
ANSWER: A
Explanation: Data is used for retrieval purpose only

2. DSS in DW is_____.
A. Data Support system.
B. Decision Single System.
C. Data Storable System.
D. Decision Support System.
ANSWER: D
Explanation: Decision support systems are mainly used to take business decisions

3. The data found within the data warehouse is_____.
A. subject-oriented.
B. time-variant.
C. integrated.
D. All of the above.
ANSWER: D
Explanation: As per the definition of data warehouse

4. The time horizon in the Data warehouse is usually _____.
A. 1-2 years.
B. 3-4years.
C. 5-6 years.
D. 5-10 years.
ANSWER: D
Explanation: Data warehouse stores historical data

5. The data is stored, retrieved & updated in _____.
A. OLTP.
B. OLAP.
C. SMTP.
D. FTP.
ANSWER: A
Explanation: Online transaction processing is about day today transactions

6. The star schema is composed of _____ fact table.
A. two
B. one
C. three.
D. four.
ANSWER: B
Explanation: Topological architecture of Star schema has centrally located fact table surrounded by Dimension tables

7. Data warehouse contains_____data that is never found in the operational environment.
A. normalized.
B.Summary .
C. informational.
D. denormalized.
ANSWER: B
Explanation: Summaries are stored in the DW

8. _____ is a good alternative to the star schema.
A. Star schema.
B. Snowflake schema.
C. Fact constellation.
D. Star-snowflake schema.
ANSWER: C
Explanation: Data is normalised in fact constellation

9. Fact tables are _____.
A. completely normalized.
B. partially demoralized.
C. completely denormalized.
D. partially normalized.
ANSWER: A
Explanation: No redundancy is present in fact data

10. OLAP stands for
a) Online analytical processing
b) Online analysis processing
c) Online transaction processing
d) Online aggregate processing
ANSWER: A
Explanation: As per the definition

11. Data that can be modeled as dimension attributes and measure attributes are called _____ data.
a) dimensional
b) Single Dimensional
c) Measured
d) Multidimensional
ANSWER: D
Explanation: OLAP supports multidimensional data models.

12. The process of viewing the cross-tab (Single dimensional) with a fixed value of one attribute is
a) Slicing
b) Dicing
c) Pivoting
d) Both Slicing and Dicing
ANSWER: A
Explanation: The operation performed on single attribute

13. The operation of moving from finer-granularity data to a coarser granularity is called a _____
a) Rollup
b) Drill down
c) Dicing
d) Pivoting
ANSWER: a
Explanation: The opposite operation—that of moving from coarser-granularity data to finer-granularity data—is called a drill down.

14. ..... schema supports multiple fact tables
A. Star schema.
B. Snowflake schema.
C. Fact constellation.
D. Star-snowflake schema.
ANSWER: C
Explanation: Multiple fact tables can be shared among Dts are present in fact constellation

15 Starting with the base cuboid [day, doctor, patient], what specific OLAP operations should be performed in order to list the total fee collected by each doctor in 2004?
a Roll-up on time from day to year.
b. Drill-down on time from day to year.
c. Roll-up on time from year to day.
d. Drill-down on time from year to day.
ANSWER: a

Explanation: Different steps are

1. Roll-up on time from day to year.

2. Slice for time=2004.

3. Roll-up on patient from individual patient to all.

16 Starting with the base cuboid [day, doctor, patient], what specific OLAP operations should be performed in order to list the total fee collected by each doctor for all patients?
a Roll-up on patients from all to individual patient.
b. Drill-down on patients from individual patient to all.
c. Roll-up on patients from individual patient to all.
d. Drill-down on patients from all to individual patient.
ANSWER: c
Explanation:

Different steps are

1. Roll-up on patient from individual to all.

2. Slice on doctor

17 Starting with the base cuboid [day, doctor, patient], what specific OLAP operations should be performed in order to list the total fee collected by doctor Mohan?
a Roll-up on doctor from individual doctor to all.
b. Drill-down on doctor from individual doctor to all.
c. Roll-up on doctor from all to Individual doctor.
d. Drill-down on doctor from all to Individual doctor.
ANSWER: d
Explanation:

Different steps are

1. Slice for Doctor=Mohan.

2. Drill_down on doctor from all to Dr. Mohan

18. _____databases are owned by particular departments or business groups.

A. Informational.
B. Operational.
C. Both informational and operational.
D. Flat.
ANSWER: B
Explanation: Operational data is generated from particular departments

19.The key used in an operational environment may not have an element of_____.

A. time.
B. cost.
C. frequency.
D. quality.

ANSWER: A
Explanation: Time is important attribute in data warehouse may not in operational

20. Data can be updated in _____environment
A. data warehouse.
B. data mining.
C. operational.
D. informational.
ANSWER:C
Explanation: Operational data gets frequent updates.

21. Data can not be updated in _____environment.
A. data warehouse.
B. data mining.
C. operational.
D. informational.
ANSWER:A
Explanation: Historical data is stored in a Data warehouse. The main purpose is read only

22. The extract process is _____.
A. capturing all of the data contained in various operational systems.
B. capturing a subset of the data contained in various operational systems.
C. capturing all of the data contained in various decision support system
D. capturing a subset of the data contained in various decision support systems
ANSWER: B
Explanation: It is a subset of operational data

23. The load and index is _____.
A. a process to reject data from the data warehouse and to create the necessary indexes.
B. a process to load the data in the data warehouse and to create the necessary indexes.
C. a process to upgrade the quality of data after it is moved into a data warehouse.
D. a process to upgrade the quality of data before it is moved into a data warehouse.
ANSWER: B
Explanation: It is loading and index creation

24. The type of relationship in star schema is _____.
A. many-to-many.
B. one-to-one.
C. one-to-many.
D. many-to-one.
ANSWER: C
Explanation: One fact and many Dimension tables

25. Fact tables are _____.

A. completely demoralized.
B. partially demoralized.
C. completely normalized.
D. partially normalized.
ANSWER: C
Explanation: Data is normalized in fact and may be redundant in Dts
26. Data warehouse architecture is based on _____.
A. DBMS.
B. RDBMS.
C. Sybase.
D. SQL Server.
ANSWER: B
Explanation: Schema designing is based on RDBMS

27. _____ is data about data.
A. Metadata.
B. Microdata.
C. Minidata.
D. Multidata.
ANSWER: A
Explanation: Metadat is data about data

28. MDDB stands for _____.
A. multiple data doubling.
B. multidimensional databases.
C. multiple double dimension.
D. multi-dimension doubling.
ANSWER: B
Explanation: As per the long form

29. _____ is the heart of the warehouse.
A. Data mining database servers.
B. Data warehouse database servers.
C. Data mart database servers.
D. Relational database servers.
ANSWER: B
Explanation: Data Warehouse servers comes at middle layer of architecture and actual store data and responsible for retrieval

30. Records cannot be updated in _____.

A. OLTP
B. files
C. RDBMS
D. data warehouse
ANSWER: D
Explanation: Data warehouse is used for read only purpose

31. Data transformation includes _____.
A. a process to change data from a detailed level to a summary level.
B. A process to change data from a summary level to a detailed level.
C. joining data from one source into various sources of data.
D. separating data from one source into various sources of data.
ANSWER: A
Explanation: Transformation is required for better query results and present in summarized view

32. { (item name, color, clothes size), (item name, color), (item name, clothes size), (color, clothes size), (item name), (color), (clothes size), () }
This can be achieved by using which of the following ?
a) group by rollup
b) group by cubic
c) group by
d) none of the mentioned
Answer: d
Explanation: 'Group by cube' is used .

33. Decision support systems (DSS) is
a. A family of relational database management systems marketed by IBM
b.Interactive systems that enable decision makers to use databases and models on a computer in order to solve ill-structured problems
c. It consists of nodes and branches starting from a single root node. Each node represents a test, or decision
d. none of above
Ans:b
Explanation:It supports decision making process

34.Data warehouse is__
a. The actual discovery phase of a knowledge discovery process
b. The stage of selecting the right data for a KDD process
c.A subject-oriented integrated time variant non-volatile collection of data in support of management
d. none of above
Ans: c
Explanation: As per the definition

35. ETL stands for
a. Expand translate load
b. Extend transfer load
c. Extract translate load
d. Extract transform load
Answer: d
explanation: It is the process of extract the data from multiple resources, transform in required format and load to data warehouse

36.In a data warehouse, if D1 and D2 are two conformed dimensions, then
(a) D1 may be an exact replica of D2
(b) D1 may be at a rolled up level of granularity compared to D2
(c) Columns of D1 may be a subset of D2 and vice versa
(d) Rows of D1 may be a subset of D2 and vice versa
Answer: A
Explanation: In a data warehouse, if D1 and D2 are two conformed dimensions, then D1 may be an exact replica of D2.

37. Which of the following is not an ETL tool?
(a) Informatica
(b) Oracle warehouse builder
(c) Datastage
(d) Visual studio
Answer: D
Explanation: Visual Studio is not an ETL tool.

38. The data is stored, retrieved and updated in .....
A) OLAP
B) OLTP
C) SMTP
D) FTP
Answer: B
39. An ..... system is market-oriented and is used for data analysis by knowledge workers, including managers, executives, and analysts.
A) OLAP
B) OLTP
C) Both of the above
D) None of the above
Answer: A

40. The core of the multidimensional model is the ..... , which consists of a large set of facts and a number of dimensions.
A) Multidimensional cube
B) Dimensions cube
C) Data cube
D) Data model
Answer: C
Explanation: Data cube is represented with fact and dimensions

41) The data from the operational environment enter ..... of data warehouse.
A) Current detail data
B) Older detail data
C) Lightly Summarized data
D) Highly summarized data
Answer: A
Explanation: current data generated due to day today transactions

42) A data warehouse is .....
A) updated by end users.
B) contains numerous naming conventions and formats
C) organized around important subject areas
D) contain only current data
Answer: C
Explanation: Stores subject specific data

43) Business Intelligence and data warehousing is used for .....
A) Forecasting
B) Data Mining
C) Analysis of large volumes of product sales data
D) All of the above
Answer: D) All of the above
Explanation: All are applications of DM

44..... are designed to overcome any limitations placed on the warehouse by the nature of the relational data model.
A) Operational database
B) Relational database
C) Multidimensional database
D) Data repository

Answer: C
Explanation: As Multidimensional database provides summarized data easy for retrieval
45. Data modeling technique used for data marts is
(a) Dimensional modeling
(b) ER – model
(c) Extended ER – model
(d) Physical model
(e) Logical model.
Answer: A
Explanation: Data modeling technique used for data marts is Dimensional modeling.

46. A warehouse architect is trying to determine what data must be included in the warehouse. A meeting has been arranged with a business analyst to understand the data requirements, which of the following should be included in the agenda?
(a) Number of users
(b) Corporate objectives
(c) Database design
(d) Routine reporting
(e) Budget.
Answer: D
Explanation: Routine reporting should be included in the agenda.

47. An OLAP tool provides for
(a) Multidimensional analysis
(b) Roll-up and drill-down
(c) Slicing and dicing
(d) Rotation
(e) Setting up only relations.
Answer: C
Explanation: An OLAP tool provides for Slicing and dicing.

48 Which of the following statements is true?
(a) A fact table describes the transactions stored in a DWH
(b) A fact table describes the granularity of data held in a DWH
(c) The fact table of a data warehouse is the main store of descriptions of the transactions stored in a DWH
(d) The fact table of a data warehouse is the main store of all of the recorded transactions over time
(e) A fact table maintains the old records of the database.
Answer: D

Explanation: The fact table of a data warehouse is the main store of all of the recorded transactions over time is the correct statement.

49. Concept description is the basic form of the

(a) Predictive data mining

(b) Descriptive data mining

(c) Data warehouse

(d) Relational database

(e) Proactive data mining.

Answer: B

Explanation: Concept description is the basis form of the descriptive data mining.

50. A Business Intelligence system requires data from:

(a) Data warehouse

(b) Operational systems

(c) All possible sources within the organization and possibly from external sources

(d) Web servers

(e) Database servers.

Answer: A

Explanation: A business Intelligence system requires data from Data warehouse

51. Which of the following projects is building a data mart for a business process/department that is very critical for your organization?

(a) High risk high reward

(b) High risk low reward

(c) Low risk low reward

(d) Low risk high reward

(e) Involves high risks.

Answer: A

Explanation: High risk high reward project is a building a data mart for a business process/department that is very critical for your organization

MCQs Unit-3

1. If in this formula if p=1 then it is..... $d(i, j) = (|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \cdots + |x_{in} - x_{jn}|^p)^{1/p},$
a. Manhattan distance
b. Minkowski distance
c. Euclidean distance
d. None of above

Ans: a
Explanation: As per the formula if p is set to 1 , it becomes formula of Manhattan, if p=2 then formula of Euclidean distance

2. A binary variable is _____ if both of its states are equally valuable and carry the same weight
a. Symmetric
b. Asymmetric

Ans: a
Explanation: As per the definition of symmetric and asymmetric variables

3. Color={red, white, blue} attribute is of type-----
a. Binary
b. Categorical
c. Ordered
d. Numerical

Ans: b
Explanation: It is distinct valued attribute.

4. Post={assistant, associate, professor} attribute is of type-----
a. Binary
b. Categorical
c. Ordinal
d. Numerical

Ans: c In
Explanation: ordinal variable maintains order of data elements.

5.Gender={male, female} attribute is of type-----
a. Asymmetric Binary
b. Symmetric Binary
c. Ordinal
d. Numerical

Ans: b
Explanation: Male and female have equal importance hence it is symmetric variable

6. Test={positive, negative} attribute is of type-----
a. Asymmetric Binary
b. Symmetric Binary
c. Ordinal
d. Numerical

Ans: a
Explanation: Positive results has high weigthage than negative value hence it is asymmetric binary variable.

6. Find Euclidean distance between A(23,12), B(10,34)----
a. 23
b. 25.55
c. 29
d. 0

Ans: b
Explanation: As per following formula

$$d(i, j) = (|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \cdots + |x_{in} - x_{jn}|^p)^{1/p},$$

7. Find Co-sine similarity between x=(1,1,0,0) y= (0,1,1,0) -----

a. 0
b. 1
c. 0.5
d. none of above
Ans: c
Explanation: as per the formula

8. Binary attribute are
A) This takes only two values. In general, these values will be 0 and 1
B) The natural environment of a certain species.
C) Systems that can be used without knowledge of internal operations.
D) None of these
Answer: A
Explanation: Two valued attribute

9. Euclidean distance measure is

A) A stage of the KDD process in which new data is added to the existing selection.
B) The process of finding a solution for a problem simply by enumerating all possible solutions according to some pre-defined order and then testing them
C) The distance between two points as calculated using the Pythagoras theorem.
D) None of these

Ans: c
Explanation: As per the formula

10. Which of the following is true about Manhattan distance?

A) It can be used for continuous variables
B) It can be used for categorical variables
C) It can be used for categorical as well as continuous
D) None of these

Ans: A

Explanation: Manhattan Distance is designed for calculating the distance between real valued features.

11. Which of the following will be Euclidean Distance between the two data point A(1,3) and B(2,3)?

A) 1
B) 2
C) 4
D) 8

Ans:A

Explanation: sqrt((1-2)^2 + (3-3)^2) = sqrt(1^2 + 0^2) = 1

12. Which of the following will be Manhattan Distance between the two data point A(1,3) and B(2,3)?

A) 1
B) 2
C) 4
D) 8

Ans:A

Explanation: sqrt(mod((1-2)) + mod((3-3))) = sqrt(1 + 0) = 1

13. Which of the following is true about Manhattan distance?

A) It can be used for continuous variables
B) It can be used for categorical variables
C) It can be used for categorical as well as continuous
D) None of these

Ans: A

Explanation: Manhattan Distance is designed for calculating the distance between real valued features.

14. Data set {brown, black, blue, green , red} is example of Select one:
a. Continuous attribute
b. Ordinal attribute
c. Numeric attribute
d. Nominal attribute
Ans: d

Explanation: Nominal or categorical attributes

15. Identify the example of sequence data Select one:
a. weather forecast
b. data matrix
c. market basket data
d. genomic data
Ans: d
Explanation:It is the genetic information of Individual

16. To detect fraudulent usage of credit cards, the following data mining task should be used Select one:
a. Outlier analysis
b. prediction
c. association analysis
d. feature selection
Ans: a
Explanation: Fraud detection is the application of Outlier analysis

17. Which of the following is not an example of ordinal attributes?
a. Zip codes
b. Ordered numbers
c. Movie ratings
d. Military ranks
Ans: a
Explanation: Ordinal follows order in data values

18. In asymmetric attribute
a. No value is considered important over other values
b. All values are equals
c. Only non-zero value is important
d. Range of values is important
Ans: c
Explanation: both output does not have equal weightage

19. Identify the example of Nominal attribute:
a. Temperature
b. Salary
c. Mass
d. Gender
Ans: d
Explanation: Gender has distinct values like Female and Male

20. Nominal and ordinal attributes can be collectively referred to as_____ attributes
a. perfect
b. qualitative
c. consistent
d. optimized
Ans: b
Explanation: It has distinct values

21. The dissimilarity of _____is computed with this formula d(i,j)=p-m/p
a. Continuous attribute
b. Ordinal attribute
c. Numeric attribute
d. Nominal attribute
Ans: d
Explanation: Nominal or categorical attributes are computed with these measures

22. If d(i,j)=0 then object i and j are more similar
a. true
b. false
Ans: true

Explanation: d(i,j) shows dissimilarity lower the value more similar , 1 shows non-similar

23. If d(i,j)=1 then object i and j are more similar
a. true
b. false
Ans: true

Explanation: d(i,j) shows dissimilarity lower the value more similar , 1 shows non-similar

24. Similarity between two objects is computed as sim (i,j)=1- d(i,j), dissimilarity
a. true
b. false
Ans: true
Explanation: 1-d(i,j) is the shows similarity

25. In the _____similarity sim(i,j) is also called as Jaccard coefficient
a symmetric binary
b asymmetric binary
c. Numeric
d. none of above
Ans:b
Explanation: The coefficient sim(i, j) is called the Jaccard coefficient in asymmetric binary

26. k-NN is based on ____distance measure
a. Manhattan distance
b. Minkowski distance
c. Euclidean distance
d. Supremum distance
Ans:c
Explanation: Eclidean distance is popular distance measure

27. Clustring is based on ____distance measure
a. Manhattan distance
b. Euclidean distance
c. Minkowski distance
d. Supremum distance

Ans:b
Explanation: Eclidean distance is popular distance measure

28. If in this formula if p=2 then it is..... $d(i, j) = (|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \cdots + |x_{in} - x_{jn}|^p)^{1/p},$
a. Manhattan distance
b. Minkowski distance
c. Euclidean distance
d. Supremum distance

Ans: c
Explanation: As per the formula if p is set to 1 , it becomes formula of Manhattan, if p=2 then formula of Euclidean distance and for p=1/2 it becames Minkowski distance

29. If in this formula if p is any finite number $d(i, j) = (|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \cdots + |x_{in} - x_{jn}|^p)^{1/p},$
then it is..... a. Manhattan distance
b. Minkowski distance
c. Euclidean distance
d. None of above

Ans: b
Explanation: Explanation: As per the formula if p is set to 1 , it becomes formula of Manhattan, if p=2 then formula of Euclidean distance and for p=any numer it becames Minkowski distance

30. Dissimilarity matrix is used in ____machine learning algortithm
a.K-Means
b. K-NN
c. NN
d. Decision tree
Ans: a
Explanation: Dissimilarity matrix represents distance between data points. It is used in clustering algorithms.

31. Dissimilarity matrix is used in ____machine learning algortithm
a.K-Medoid
b. K-NN
c. NN
d. Decision tree
Ans: a
Explanation: Dissimilarity matrix represents distance between data points. It is used in clustering algorithms.

32. Calculate the disimilarity(p,p1) for p(-7,-4) and p1(17,6)
a. 25
b. 26
c. 27
d. 28
Ans: b
Explanation: by default distance measure is Euclidean distance measure

33. In Euclidean distance , value is always ___
a. positive integer
b. negative integer
Ans: a
Explanation: Due to square the value is always positive

34. In Manhattan distance , value is always ___
a. positive integer
b. negative integer
Ans: a
Explanation: Due to mod operator value is always positive

35. The standardization form of Euclidean distance is called as___
a. Manhattan distance
b. Minkowski distance
c. Euclidean distance
d. Supremum distance
Ans: b
Explanation: In the above formuls p can set to 1

36. Most commonly used measure of the similarity in between two objects is called___
a. Manhattan distance
b. Minkowski distance
c. Euclidean distance
d. Both c and b
Ans:c
Explanation: Euclidean is popular and by default similarity measure

37. Calculate the dissimilarity(p,p1) for p(1,1) and p1(1,1)
a. 0
b. 1
Ans: a
Explanation: In dissimilarity 0 represents similarity and 1 represents dissimilarity

39. Calculate the dissimilarity(p,p1) for p(2,2) and p1(-2,-2)
a. non-zero
b. zero
Ans: b
Explanation: Two points are not similar (one is positive and other is negative) so distance is non-zero value

40. The city block distance is also known as
a. Manhattan distance
b. Minkowski distance
c. Euclidean distance
d. Supremum distance

Ans: a
Explanation: None

41. Manhattan distance between x1(1,2) and x2 (3,5) is __
a. 5
b. 3.61
c. 4
d. 1
Ans: a
Explanation: |1-3|+|2-5|=2+3=5

42. Euclidean distance between x1(1,2) and x2 (3,5) is __
a. 5
b. 3.61
c. 4
d. 1
Ans: b
Explanation: (|1-3|^2+|2-5|^2)^1/2=3.61

43. City block distance between x1(1,2) and x2 (3,5) is __
a. 5
b. 3.61
c. 3
d. 1
Ans: a
Explanation: |1-3|+|2-5|=2+3=5. City block is also known as Manhattan.

44. The generalized form of Minkowski distance is called as___
a. Manhattan distance
b. Supremum distance
c. Euclidean distance
d. None of above
Ans: b
Explanation: It uses maximum difference in values between the two objects.

45. Compute Supremum distance x1(1,2) and x2 (3,5) is __
a. 5
b. 3.61
c. 3
d. 1
Ans: c
Explanation: The second attribute gives greatest difference between values for the object, which is 5-3=2.

46. Given two objects represnted by tuple (22,1) and (20,0) then Eucliden distance is
a. 3
b. 2.23
c. 2
d. 2.08
Ans: b
Explanation: ((22-20)^2+(1-0)^2)^1/2

47. Given two objects represnted by tuple (22,1) and (20,0) then Supremum distance is
a. 3
b. 2.23
c. 2
d. 2.08
Ans: c
Explanation: 22-20=2

48. Given two objects represnted by tuple (22,1) and (20,0) then Minkowski distance (p=3)is
a. 3
b. 2.23
c. 2
d. 2.08
Ans: d
Explanation: ((22-20)^3+(1-0)^3)^1/3

49. Given two objects represnted by tuple (22,1) and (20,0) then Manhattan distance is
a. 3
b. 2.23
c. 2
d. 2.08
Ans: a
Explanation: |22-20|+|1-0|=3

50. Interval -scaled and raio-scaled are type of
a. a. Continuous attribute
b. Ordinal attribute
c. Numeric attribute
d. Nominal attribute
Ans: c
Explanation: Interval scaled attributes measured in fixed and equal units. Ratio-scaled are numeric attributes with an inherent zero scale.