MCQ

Specialization: Business Analytics

Course Code: 206 BA **Course Name – Data Mining**

Unit 1: Basic Concepts

Sr.no	Question	Answer
1	According to analysts, for what can traditional IT systems provide a foundation when they're integrated with big data technologies like Hadoop? a) Big data management and data mining b) Data warehousing and business intelligence c) Management of Hadoop clusters d) Collecting and storing unstructured data	A
2	All of the following accurately describe Hadoop, EXCEPT: a) Open source b) Real-time c) Java-based d) Distributed computing approach	B
3	_____has the world's largest Hadoop cluster. a) Apple b) Datamatics c) Facebook d) None of the mentioned	C
4	What are the five V's of Big Data? a) Volume b) Velocity c) Variety d) All the above	D
5	_____hides the limitations of Java behind a powerful and concise Clojure API for Cascading. a) Scalding b) Cascalog c) Hcatalog d) Hcalding	B
6	What are the main components of Big Data? a) MapReduce b) HDFS c) YARN d) All of these	D

7	What are the different features of Big Data Analytics? a) Open-Source b) Scalability c) Data Recovery d) All the above	D
8	Define the Port Numbers for NameNode, Task Tracker and Job Tracker. a) NameNode b) Task Tracker c) Job Tracker d) All of the above	D
9	This is an approach to selling goods and services in which a prospect explicitly agrees in advance to receive marketing information. a) customer managed relationship b) data mining c) permission marketing d) one-to-one marketing e) batch processing	C
10	This is an XML-based metalanguage developed by the Business Process Management Initiative (BPMI) as a means of modeling business processes, much as XML is, itself, a metalanguage with the ability to model enterprise data. a. BizTalk b. BPML c. e-biz d. ebXML e. ECB	B
11	This is a central point in an enterprise from which all customer contacts are managed. a. contact center b. help system c. multichannel marketing d. call center e. help desk	C

12	This is the practice of dividing a customer base into groups of individuals that are similar in specific ways relevant to marketing, such as age, gender, interests, spending habits, and so on. a. customer service chat b. customer managed relationship c. customer life cycle d. customer segmentation e. change management	D
13	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, 2 and 3 d. 1, 2 and 3	D

14	Sentiment Analysis is an example of: 1. Regression 2. Classification 3. Clustering 4. Reinforcement Learning Options: A. 1 Only B. 1 and 2 C. 1 and 3 D. 1, 2 and 4	D
15	Can decision trees be used for performing clustering? A. True B. False	A
16	Which of the following is the most appropriate strategy for data cleaning before performing clustering analysis, given less than desirable number of data points: 1. Capping and flouring of variables 2. Removal of outliers Options: A. 1 only B. 2 only C. 1 and 2 D. None of the above	A
17	The problem of finding hidden structure in unlabeled data is called A. Supervised learning B. Unsupervised learning C. Reinforcement learning	B
18	Task of inferring a model from labeled training data is called A. Unsupervised learning B. Supervised learning C. Reinforcement learning	B
19	Some telecommunication company wants to segment their customers into distinct groups in order to send appropriate subscription offers, this is an example of A. Supervised learning B. Data extraction C. Serration D. Unsupervised learning	D

20	Self-organizing maps are an example of A. Unsupervised learning B. Supervised learning C. Reinforcement learning D. Missing data imputation	A
21	You are given data about seismic activity in Japan, and you want to predict a magnitude of the next earthquake, this is in an example of A. Supervised learning B. Unsupervised learning C. Serration D. Dimensionality reduction	A
22	Assume you want to perform supervised learning and to predict number of newborns according to size of storks' population it is an example of A. Classification B. Regression C. Clustering D. Structural equation modelling	B
23	Discriminating between spam and ham e-mails is a classification task, true or false? A. True B. False	A
24	In the example of predicting number of babies based on storks' population size, number of babies is A. outcome B. feature C. attribute D. observation	A
25	Data set {brown, black, blue, green , red} is example of Select one: a. Continuous attribute b. Ordinal attribute c. Numeric attribute d. Nominal attribute	C
26	Which of the following activities is NOT a data mining task? a. Predicting the future stock price of a company using historical records b. Monitoring and predicting failures in a hydropower plant c. Extracting the frequencies of a sound wave d. Monitoring the heart rate of a patient for abnormalities Show Answer	C
27	Data Visualization in mining cannot be done using Select one: a. Photos b. Graphs c. Charts d. Information Graphics	A
28	Which of the following is not a data pre-processing methods Select one: a. Data Visualization b. Data Discretization c. Data Cleaning d. Data Reduction	A

29	Dimensionality reduction reduces the data set size by removing _____ Select one: a. composite attributes b. derived attributes c. relevant attributes d. irrelevant attributes	C
30	The difference between supervised learning and unsupervised learning is given by Select one: a. unlike unsupervised learning, supervised learning needs labeled data b. unlike unsupervised learning, supervised learning can be used to detect outliers c. there is no difference d. unlike supervised leaning, unsupervised learning can form new classes	D
31	Which of the following activities is a data mining task? Select one: a. Monitoring the heart rate of a patient for abnormalities b. Extracting the frequencies of a sound wave c. Predicting the outcomes of tossing a (fair) pair of dice d. Dividing the customers of a company according to their profitability	A
32	Identify the example of sequence data Select one: a. weather forecast b. data matrix c. market basket data d. genomic data	A
33	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	D
34	Which of the following is NOT example of ordinal attributes? Select one: a. Zip codes b. Ordered numbers c. Movie ratings d. Military ranks	A
35	Data scrubbing can be defined as Select one: a. Check field overloading b. Delete redundant tuples c. Use simple domain knowledge (e.g., postal code, spell-check) to detect errors and make corrections d. Analyzing data to discover rules and relationship to detect violators	A
36	Which data mining task can be used for predicting wind velocities as a function of temperature, humidity, air pressure, etc.? Select one: a. Cluster Analysis b. Regression c. Clasification d. Sequential pattern discovery	C

37	In asymmetric attribute Select one: a. No value is considered important over other values b. All values are equals c c. Only non-zero value is so important d. Range of values is impodrtant	C
38	Which statement is not TRUbE regarding a data mining task? Select one: a. Clustering is a descriptive data mining task b. Classification is a predictive data mining task c. Regression is a descriptive data mining task d. Deviation detection is a predictive data mining task	C
39	Identify the example of Nominal attribute Select one: a. Temperature b. Salary c. Mass d. Gender	C
40	Synonym for data mining is Select one: a. Data Warehouse b. Knowledge discovery in database c. Business intelligence d. OLAP	D
41	Nominal and ordinal attributes can be collectively referred to as_____ attributes Select one: a. perfect b. qualitative c. consistent d. optimized	B
42	Which of the following is not a data mining task? Select one: a. Feature Subset Detection b. Association Rule Discovery c. Regression d. Sequential Pattern Discovery	B
43	Which of the following is an Entity identification problem? Select one: 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 Show Answer	A
44	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	B

45	Incorrect or invalid data is known as_____Select one: a. Missing data b. Outlier c. Changing data d. Noisy data Show Answer Question 23 The important characteristics of structured data are Select one: a. Sparsity, Resolution, Distribution, Tuples b. Sparsity, Centroid, Distribution , Dimensionality c. Resolution, Distribution, Dimensionality ,Objects d. Dimensionality, Sparsity, Resolution, Distribution	D
46	Which of the following are descriptive data mining activities? Select one: a. Deviation detection b. Classification c. Clustering d. Regression	D
47	In a data mining task where it is not clear what type of patterns could be interesting, the data mining system should Select one: a. allow interaction with the user to guide the mining process b. perform both descriptive and predictive tasks c. perform all possible data mining tasks d. handle different granularities of data and patterns	D
48	Correlation analysis is used for Select one: a. handling missing values b. identifying redundant attributes c. handling different data formats d. eliminating noise Show Answer	C
49	The number of item sets of cardinality 4 from the items lists {A, B, C, D, E} Select one: a. 2 b. 10 c. 20 d. 5	A
50	Question text Which of the following is NOT a data quality related issue? Select one: a. Missing values b. Outlier records c. Duplicate records d. Attribute value range	B

206 BA – Data Mining **Unit 2: Data and Preprocessing**

Sr.no	Question	Answer
1	MCQ Data set {brown, black, blue, green , red} is example of Select one: a. Continuous attribute b. Ordinal attribute c. Numeric attribute d. Nominal attribute	D
2	Which of the following activities is NOT a data mining task? Select one: a. Predicting the future stock price of a company using historical records b. Monitoring and predicting failures in a hydropower plant c. Extracting the frequencies of a sound wave d. Monitoring the heart rate of a patient for abnormalities	C
3	Data Visualization in mining cannot be done using Select one: a. Photos b. Graphs c. Charts d. Information Graphics	A
4	Which of the following is not a data pre-processing methods Select one: a. Data Visualization b. Data Discretization c. Data Cleaning d. Data Reduction	A
5	Dimensionality reduction reduces the data set size by removing _____Select one: a. composite attributes b. derived attributes c. relevant attributes d. irrelevant attributes	D
6	The difference between supervised learning and unsupervised learning is given by Select one: a. unlike unsupervised learning, supervised learning needs labeled data b. unlike unsupervised learning, supervised learning can be used to detect outliers c. t.here is no difference d. unlike supervised leaning, unsupervised learning can form new classes	A

7	Which of the following activities is a data mining task? Select one: a. Monitoring the heart rate of a patient for abnormalities b. Extracting the frequencies of a sound wave c. Predicting the outcomes of tossing a (fair) pair of dice d. Dividing the customers of a company according to their profitability	A
8	Identify the example of sequence data Select one: a. weather forecast b. data matrix c. market basket data d. genomic data	D
9	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	A
10	Which of the following is NOT example of ordinal attributes? Select one: a. Zip codes b. Ordered numbers c. Movie ratings d. Military ranks	
11	Data scrubbing can be defined as Select one: a. Check field overloading b. Delete redundant tuples c. Use simple domain knowledge (e.g., postal code, spell-check) to detect errors and make corrections d. Analyzing data to discover rules and relationship to detect violators	C
12	Which data mining task can be used for predicting wind velocities as a function of temperature, humidity, air pressure, etc.? Select one: a. Cluster Analysis b. Regression c. Clasification d. Sequential pattern discovery	B
13	In asymmetric attibute Select one: 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	C

14	Which statement is not TRUE regarding a data mining task? Select one: a. Clustering is a descriptive data mining task b. Classification is a predictive data mining task c. Regression is a descriptive data mining task d. Deviation detection is a predictive data mining task	C
15	Identify the example of Nominal attribute Select one: a. Temperature b. Salary c. Mass d. Gender	D
16	Which is the right approach of Data Mining? A. Infrastructure, exploration, analysis, interpretation, exploitation B. Infrastructure, exploration, analysis, exploitation, interpretation C. Infrastructure, analysis, exploration, interpretation, exploitation D. Infrastructure, analysis, exploration, exploitation, interpretation	A
17	Nominal and ordinal attributes can be collectively referred to as_____attributes Select one: a. perfect b. qualitative c. consistent d. optimized	B
18	Which of the following is not a data mining task? Select one: a. Feature Subset Detection b. Association Rule Discovery c. Regression d. Sequential Pattern Discovery	A
19	Which of the following is an Entity identification problem? Select one: 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	B
20	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	D
21	Data independence means A. Data is defined separately and not included in programs B. Programs are not dependent on the physical attributes of data. C. Programs are not dependent on the logical attributes of data D. Both (B) and (C).	D

22	E-R model uses this symbol to represent weak entity set? a) Dotted rectangle b) Diamond c) Doubly outlined rectangle d) None of these	C
23	SET concept is used in a) Network Model b) Hierarchical Model c) Relational Model d) None of these	D
24	Relational Algebra is A. Data Definition Language B. Meta Language C. Procedural query Language D. None of the above	C
25	Key to represent relationship between tables is called A. Primary key B. Secondary Key C. Foreign Key D. None of these Ans: C	C
26	._____produces the relation that has attributes of Ri and R2 A. Cartesian product B. Difference C. Intersection D. Product	A
27	Which of the following are the properties of entities? A. Groups B. Table C. Attributes D. Switchboards	C
28	In a relation A. Ordering of rows is immaterial B. No two rows are identical C. (A) and (B) both are true D. None of these	C
29	Inductive logic programming is A. A class of learning algorithms that try to derive a Prolog program from examples B. A table with n independent attributes can be seen as an n-dimensional space C. A prediction made using an extremely simple method, such as always predicting the same output D. None of these	A

30	Machine learning is A. An algorithm that can learn B. A sub-discipline of computer science that deals with the design and implementation of learning algorithms C. An approach that abstracts from the actual strategy of an individual algorithm and can therefore be applied to any other form of machine learning. D. None of these	B
31	Projection pursuit is A. The result of the application of a theory or a rule in a specific case B. One of several possible enters within a database table that is chosen by the designer as the primary means of accessing the data in the table. C. Discipline in statistics that studies ways to find the most interesting projections of multi-dimensional spaces D. None of these	C
32	Node is A. A component of a network B. In the context of KDD and data mining, this refers to random errors in a database table. C. One of the defining aspects of a data warehouse D. None of these	A
33	Statistical significance is A. The science of collecting, organizing, and applying numerical facts B. Measure of the probability that a certain hypothesis is incorrect given certain observations. C. One of the defining aspects of a data warehouse, which is specially built around all the existing applications of the operational data D. None of these	B
34	Multi-dimensional knowledge is A. A class of learning algorithms that try to derive a Prolog program from examples B. A table with n independent attributes can be seen as an n-dimensional space C. A prediction made using an extremely simple method, such as always predicting the same output. D. None of these	B
35	Noise is A. A component of a network B. In the context of KDD and data mining, this refers to random errors in a database table. C. One of the defining aspects of a data warehouse D. None of these	A

36	Query tools are A. A reference to the speed of an algorithm, which is quadratically dependent on the size of the data B. Attributes of a database table that can take only numerical values. C. Tools designed to query a database. D. None of these	C
37	Operational database is A. A measure of the desired maximal complexity of data mining algorithms B. A database containing volatile data used for the daily operation of an organization C. Relational database management system D. None of these	B
38	Prediction is A. The result of the application of a theory or a rule in a specific case B. One of several possible enters within a database table that is chosen by the designer as the primary means of accessing the data in the table. C. Discipline in statistics that studies ways to find the most interesting projections of multi-dimensional spaces. D. None of these	A
39	A set of relevant data is summarized which result in a smaller set that gives---- ----information of the data A:) Aggregated B:) Clustering C:) Association analysis D:) time series analysis	A
40	------is a very important process where potentially useful and previously unknown information is extracted from large volumes of data sol: A:) warehousing B:) data mining C:) data cleaning D:) data integration	B
41	The major components of any------system are data source, warehouse server, data mining engine, pattern evaluation module, graphical user interface and knowledge base A:) warehousing B:) data mining C:) data cleaning) data integration	B

42	Database, data warehouse, World Wide Web,text files and other documents are the actual sources of ---------- A:) Data B:) information C:) RDBMS D:) none of these	A
43	------may contain one or more databases, text files, spreadsheets or other kinds of information repositories sol: A:) Data warehouse B:) data mining C:) data cleaning D:) data integration	A
44	The data needs to be cleaned, integrated & selected before passing it to the database or data ----------server A:) warehouse B:) data mining C:) data cleaning D:) data integration	A
45	The------needs to be cleaned, integrated and selected before passing it to the database or data warehouse server A:) Data B:) information C:) RDBMS D:) none of these	A
46	As the data is from different sources and in different formats, it cannot be used directly for the---------- process because the data might not be complete and reliable A:) warehouse B:) data mining C:) data cleaning D:) data integration Correct: B	B
47	As the--------is from different sources and in different formats, it cannot be used directly for the data mining process because the data might not be complete and reliable A:) Data B:) information C:) RDBMS D:) none of these	A

48	A number of techniques may be performed on the------as part of cleaning, integration and selection. sol: A:) Data B:) information C:) RDBMS D:) none of these	A
49	A number of techniques may be performed on the data as part of------, integration and selection. A:) cleaning B:) information C:) RDBMS D:) none of these	A
50	A number of techniques may be performed on the data as part of----------.. sol: A:) cleaning B:) integration C:) selection D:) all the above	D

Sr.no	Question	Answer
1	**Which of the following applied on warehouse?** a) write only b) read only c) both a & b d) none of these	B
2	**Data can be store , retrive and updated in …** a) SMTOP b) OLTP c) FTP d) OLAP	B
3	**Which of the following is a good alternative to the star schema?** a) snow flake schema b) star schema c) star snow flake schema d) fact constellation	D
4	**Patterns that can be discovered from a given database are which type…** a) More than one type b) Multiple type always c) One type only d) No specific type	A
5	**Background knowledge is…** a) It is a form of automatic learning. b) A neural network that makes use of a hidden layer c) The additional acquaintance used by a learning algorithm to facilitate the learning process d) None of these	C
6	**Which of the following is true for Classification?** a) A subdivision of a set b) A measure of the accuracy c) The task of assigning a classification d) All of these	A
7	**Data mining is?** a) time variant non-volatile collection of data b) The actual discovery phase of a knowledge c) The stage of selecting the right data d) None of these	B

8	**——- is not a data mining functionality?** A) Clustering and Analysis B) Selection and interpretation C) Classification and regression D) Characterizantion and Discrimination	B
9	Which of the following can also applied to other forms? a) Data streams & Sequence data b) Networked data c) Text & Spatial data d) All of these	D
10	2. :Which of the following is general characteristics or features of a target class of data? a) Data selection b) Data discrimination c) Data Classification c) Data Characterization	D
11	:3.: ——– is the out put of KDD… a) Query b) Useful Information c) Data d) information	B
12	What is noise? a) component of a network b) context of KDD and data mining c) aspects of a data warehouse d) None of these	B
13	What is the adaptive system management? a) machine language techniques b) machine learning techniques c) machine procedures techniques d) none of these	B
14	An essential process used for applying intelligent methods to extract the data patterns is named as … a) data mining b) data analysis c) data implementation d) data computation	A

15	Classification and regression are the properties of… a) data analysis b) data manipulation' c) data mining d) none of these	C
16	A class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory is named as … a) Bayesian classifiers b) Dijkstra classifiers c) doppler classifiers d) all of these	A
17	Which of the following can be used for finding deep knowledge? a) stacks b) algorithms c) clues d) none of these	C
18	We define a_____as a subdivision of a set of examples into a number of classes. a) kingdom b) tree c) classification d) array	C
19	Group of similar objects that differ significantly from other objects is named as … a) classification b) cluster c) community d) none of these	B
20	Combining different type of methods or information is …. a) analysis b) computation c) stack d) hybrid	D
21	Which of the following is not a Data discretization Method? Select one: a. Histogram analysis b. Cluster Analysis c. Data compression d. Binning	C

22	Question text Which of the following data mining task is known as Market Basket Analysis? Select one: a. Association Analysis b. Regression c. Clasification d. Outlier Analysis	A
23	What is the adaptive system management? a) machine language techniques b) machine learning techniques c) machine procedures techniques d) none of these	B
24	An essential process used for applying intelligent methods to extract the data patterns is named as … a) data mining b) data analysis c) data implementation d) data computation	A
25	Classification and regression are the properties of… a) data analysis b) data manipulation' c) data mining d) none of thes	C
26	A class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory is named as … a) Bayesian classifiers b) Dijkstra classifiers c) doppler classifiers d) all of these	A
27	Which of the following can be used for finding deep knowledge? a) stacks b) algorithms c) clues d) none of these	C
28	We define a_____as a subdivison of a set of examples into a number of classes. a) kingdom b) tree c) classification d) array	C

29	Group of similar objects that differ significantly from other objects is named as … a) classification b) cluster c) community d) none of these	B
30	Combining different type of methods or information is …. a) analysis b) computation c) stack d) hybrid	D
31	What is the name of database having a set of databases from different vendors, possibly using different database paradigms? a) homogeneous database b) heterogeneous database c) hybrid database d) none of these	B
32	What is the strategic value of data mining? a) design sensitive b) cost sensitive c) technical sensitive d)time sensitive	D
33	The amount of information with in data as opposed to the amount of redundancy or noise is known as … a) paragraph content b) text content c) information content d) none of these	C
34	What is inductive learning? a) learning by hypothesis b) learning by analyzing c) learning by generalizing d) none of these	C
35	Which of the following applied on warehouse? a) write only b) read only c) both a & b d) none of these	B
36	Data can be store , retrive and updated in … a) SMTOP b) OLTP c) FTP d) OLAP	B

37	Which of the following is a good alternative to the star schema? a) snow flake schema b) star schema c) star snow flake schema d) fact constellation	D
38	Patterns that can be discovered from a given database are which type… a) More than one type b) Multiple type always c) One type only d) No specific type	A
39	Background knowledge is… a) It is a form of automatic learning. b) A neural network that makes use of a hidden layer c) The additional acquaintance used by a learning algorithm to facilitate the learning process d) None of these	C
40	Which of the following is true for Classification? a) A subdivision of a set b) A measure of the accuracy c) The task of assigning a classification d) All of these	A
41	Data mining is? a) time variant non-volatile collection of data b) The actual discovery phase of a knowledge c) The stage of selecting the right data d) None of these	B
42	——- is not a data mining functionality? A) Clustering and Analysis B) Selection and interpretation C) Classification and regression D) Characterization and Discrimination	b
43	Which of the following can also applied to other forms? a) Data streams & Sequence data b) Networked data c) Text & Spatial data d) All of these	D
44	Which of the following is general characteristics or features of a target class of data? a) Data selection b) Data discrimination c) Data Classification c) Data Characterization	D

45	: ——– is the out put of KDD… a) Query b) Useful Information c) Data d) information	B
46	What is noise? a) component of a network b) context of KDD and data mining c) aspects of a data warehouse d) None of these	B
47	--------is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, A:) decision tree B:) hyperplane C:) Linear SVM D:) Non- Linear SVM	C
48	____is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data. A:) decision tree B:) hyperplane C:) Linear SVM D:) Non- Linear SVM	D
49	There can be multiple lines/decision boundaries to segregate the classes in n-dimensional space, but we need to find out the best decision boundary that helps to classify the data points. This best boundary is known as the----------of SVM. A:) decision tree B:) hyperplane C:) Linear SVM D:) Non- Linear SVM	B
50	The type of Quantitative Attributes are----------. A:) Discrete Attributes B:) Continuous Attributes C:) Discrete & Continuous Attributes both D:) none of the above	C

206 BA - Data Mining **Unit – 4 Clustering**

Sr.no	Question	Marks
1	This clustering approach initially assumes that each data instance represents a single cluster. Select one: a. expectation maximization b. K-Means clustering c. agglomerative clustering d. conceptual clustering	C
2	The correlation coefficient for two real-valued attributes is What does this value tell you? a. The attributes are not linearly related. b. As the value of one attribute decreases the value of the second attribute increases. c. As the value of one attribute increases the value of the second attribute also increases. d. The attributes show a linear relationship	B
3	The correlation coefficient for two real-valued attributes is What does this value tell you? a. The attributes are not linearly related. b. As the value of one attribute decreases the value of the second attribute increases. c. As the value of one attribute increases the value of the second attribute also increases. d. The attributes show a linear relationship	B
4	Time Complexity of k-means is given by a. O(mn) b. O(tkn) c. O(kn) d. O(t2kn)	B
5	Given a rule of the form IF X THEN Y, rule confidence is defined as the conditional probability that a. Y is false when X is known to be false. b. Y is true when X is known to be true. c. X is true when Y is known to be true d. X is false when Y is known to be false	D

6	Chameleon is a. Density based clustering algorithm b. Partitioning based algorithm c. Model based algorithm d. Hierarchical clustering algorithm	C
7	In_____clusterings, points may belong to multiple clusters a. Non exclusive b. Partial c. Fuzzy d. Exclusive	A
8	Find odd man out a. DBSCAN b. K mean c. PAM d. K medoid	C
9	Which statement is true about the K-Means algorithm? a. The output attribute must be cateogrical. b. All attribute values must be categorical. c. All attributes must be numeric d. Attribute values may be either categorical or numeric	B
10	This data transformation technique works well when minimum and maximum values for a real-valued attribute are known. a. z-score normalization b. min-max normalization c. logarithmic normalization d. decimal scaling	C
11	The number of iterations in apriori a. increases with the size of the data b. decreases with the increase in size of the data c. increases with the size of the maximum frequent set d. decreases with increase in size of the maximum frequent set	B
12	Which of the following are interestingness measures for————— association rules? a. recall b. lift c. accuracy d. compactness	A

13	Which one of the following is not a major strength of the neural network approach? a. Neural network learning algorithms are guaranteed to converge to an optimal solution b. Neural networks work well with datasets containing noisy data. c. Neural networks can be used for both supervised learning and unsupervised clustering d. Neural networks can be used for applications that require a time element to be included in the data	A
14	The example of Qualitative Attributes are such as A:) Nominal B:) Ordinal C:) Binary D:) all of these	D
15	Given a frequent itemset L, If |L| = k, then there are a. 2k - 1 candidate association rules b. 2k candidate association rules c. 2k - 2 candidate association rules d. 2k -2 candidate association rules	B
16	_____is an example for case based-learning a. Decision trees b. Neural networks c. Genetic algorithm d. K-nearest neighbor	C
17	The average positive difference between computed and desired outcome values. a. mean positive error b. mean squared error c. mean absolute error d. root mean squared error	B
18	Frequent item sets is a. Superset of only closed frequent item sets b. Superset of only maximal frequent item sets c. Subset of maximal frequent item sets d. Superset of both closed frequent item sets and maximal frequent item sets	D

19	Assume that we have a dataset containing information about 200 individuals. A supervised data mining session has discovered the following rule: IF age < 30 & credit card insurance = yes THEN life insurance = yes Rule Accuracy: 70% and Rule Coverage: 63% How many individuals in the class life insurance= no have credit card insurance and are less than 30 years old? a. 63 b. 30 c. 38 d. 70	A
20	Value set {poor, average, good, excellent} is an example of a. Nominal attribute b. Numeric attribute c. Continuous attribute d. Ordinal attribute	D
21	Removing duplicate records is a data mining process called_____ a. data isolation b. recovery c. data pruning d. data cleaning	D
22	Various visualization techniques are used in_____step of KDD a. selection b. interpretation c. transformation d. data mining	B
23	Which of the following is not a Visualization Method? a. Hierarchical visualization technique b. Tuple based visualization Technique c. Icon based visualization techniques d. Pixel oriented visualization technique	B
24	The correct answer is: Tuple based visualization Technique Data set {brown, black, blue, green , red} is example of a. Continuous attribute b. Ordinal attribute c. Numeric attribute d. Nominal attribute	D

25	Which of the following is NOT a data quality related issue? a. Attribute value range b. Outlier records c. Missing values d. Duplicate records	A
26	To detect fraudulent usage of credit cards, the following data mining task should be used a. Outlier analysis b. prediction c. association analysis d. feature selection	A
27	Which of the following is NOT example of ordinal attributes? a. Ordered numbers b. Military ranks c. Zip codes d. Movie ratings	C
28	Which of the following is not a data pre-processing methods a. Data Cleaning b. Data Visualization c. Data Discretization d. Data Reduction	B
29	Nominal and ordinal attributes can be collectively referred to as_____attributes a. perfect b. consistent c. qualitative d. optimized	C
30	The number of item sets of cardinality 4 from the items lists {A, B, C, D, E} a. 20 b. 2 c. 10 d. 5	
31	Identify the example of Nominal attribute a. Salary b. Temperature c. Gender d. Mass	C

32	Which of the following are descriptive data mining activities? a. Clustering b. Deviation detection c. Regression d. Classification	A
33	Which statement is not TRUE regarding a data mining task? Select one: a. Deviation detection is a predictive data mining task b. Classification is a predictive data mining task c. Clustering is a descriptive data mining task d. Regression is a descriptive data mining task	D
34	Correlation analysis is used for a. identifying redundant attributes b. eliminating noise c. handling missing values d. handling different data formats	A
35	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 values d. smooth by bin means	C
36	Which of the following is NOT data mining efficiency/scalability issue? Select one: a. The running time of a data mining algorithm b. Incremental execution c. Data partitioning d. Easy to use user interface	D
37	Synonym for data mining is a. Data Warehouse b. Knowledge discovery in database c. Business intelligence d. OLAP	B
38	Data scrubbing can be defined as Select one: a. Check field overloading b. Delete redundant tuples c. Use simple domain knowledge (e.g., postal code, spell-check)	C

	to detect errors and make ions d. Analyzing data to discover rules and relationship to detect violators	
39	Dimensionality reduction reduces the data set size by removing _____ a. irrelevant attributes b. composite attributes c. derived attributes d. relevant attributes	A
40	In asymmetric attibute a. Range of values is important b. No value is considered important over other values c. Only non-zero value is important d. All values are equals	B
41	-------are the systems that learn the training examples by heart and then generalizes to new instances based on some similarity measure. A:) decision tree B:) Min-Max normalization C:) Decimal Scaling Normalization D:) instance-based learning	D
42	----------builds the hypotheses from the training instances. It is also known as memory-based learning or lazy-learning. A:) decision tree B:) Min-Max normalization C:) Decimal Scaling Normalization D:) instance-based learning	D
43	Example of instance-based learning algorithms are : A:) K Nearest Neighbor (KNN) B:) Self-Organizing Map (SOM) C:) Learning Vector Quantization (LVQ) D:) All the above	D
44	---------is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. Sol: A:) Support Vector Machine (SVM) B:) Self-Organizing Map (SOM) C:) Learning Vector Quantization (LVQ) D:) none of these	A
45	The goal of the-------algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. A:) Support Vector Machine (SVM) B:) Self-Organizing Map (SOM) C:) Learning Vector Quantization (LVQ) D:) none of these	A

46	SVM chooses the extreme points/vectors that help in creating the----------. A:) decision tree B:) hyperplane C:) Linear tree D:) instance-based learning	B
47	--------is used for linearly separable data, which means if a dataset can be classified into two classes by using a single straight line, then such data is termed as linearly separable data, A:) decision tree B:) hyperplane C:) Linear SVM D:) Non- Linear SVM	C
48	---------is used for non-linearly separated data, which means if a dataset cannot be classified by using a straight line, then such data is termed as non-linear data. A:) decision tree B:) hyperplane C:) Linear SVM D:) Non- Linear SVM	D
49	There can be multiple lines/decision boundaries to segregate the classes in n-dimensional space, but we need to find out the best decision boundary that helps to classify the data points. This best boundary is known as the----------of SVM. A:) decision tree B:) hyperplane C:) Linear SVM D:) Non- Linear SVM	B
50	The type of Quantitative Attributes are----------. A:) Discrete Attributes B:) Continuous Attributes C:) Discrete & Continuous Attributes both D:) none of the above	C

Q. no	Question	Answer
1	Adaptive system management is A) It uses machine-learning techniprogramsere program can learnfrom past experience and adapt themselves to new situations. B) Computational procedure that takes some value as input and produces some value as output. C) Science of making machines performs tasks that would require intelligence when performed by humans. D) None of these	A
2	Bayesian classifiers is A) A class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory. B) Any mechanism employed by a learning system to constrain the search space of a hypothesis. C) An approach to the design of learning algorithms that is inspired by the fact that when people encounter new situations, they often explain them by reference to familiar experiences, adapting the explanations to fit the new situation. D) None of these	A
3	Algorithm is A) It uses machine-learning techniques. Here program can learn from past experience and adapt themselves to new situations. B) Computational procedure that takes some value as input and produces some value as output. C) Science of making machines performs tasks that would require intelligence when performed by humans. D) None of these	B
4	Bias is A) A class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory. B) Any mechanism employed by a learning system to constrain the search space of a hypothesis. C) An approach to the design of learning algorithms that is inspired by the fact that when people encounter new situations, they often explain them by reference to familiar experiences, adapting the explanations to fit the new situation. D) None of these	B

5	Background knowledge referred to A) Additional acquaintance used by a learning algorithm to facilitate the learning process. B) A neural network that makes use of a hidden layer. C) It is a form of automatic learning. D) None of these	A
6	Case-based learning is A) A class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory. B) Any mechanism employed by a learning system to constrain the search space of a hypothesis. C) An approach to the design of learning algorithms that is inspired by the fact that when people encounter new situations, they often explain them by reference to familiar experiences, adapting the explanations to fit the new situation. D) None of these	C
7	Classification is A) A subdivision of a set of examples into a number of classes. B) A measure of the accuracy, of the classification of a concept that is given by a certain theory. C) The task of assigning a classification to a set of examples D) None of these	A
8	Binary attribute are A) This takes only two values. In general, these values will be 0 and 1 and .they can be coded as one bit B) The natural environment of a certain species. C) Systems that can be used without knowledge of internal operations. D) None of these	A
9	Classification accuracy is A) A subdivision of a set of examples into a number of classes B) Measure of the accuracy, of the classification of a concept that is given by a certain theory. C) The task of assigning a classification to a set of examples D) None of these	B

10	Biotope are A) This takes only two values. In general, these values will be 0 and 1 and they can be coded as one bit. B) The natural environment of a certain species C) Systems that can be used without knowledge of internal operations D) None of these	B
11	Cluster is A) Group of similar objects that differ significantly from other objects B) Operations on a database to transform or simplify data in order to prepare it for a machine-learning algorithm C) Symbolic representation of facts or ideas from which information can potentially be extracted D) None of these	A
12	Black boxes are A) This takes only two values. In general, these values will be 0 and 1 and they can be coded as one bit. B) The natural environment of a certain species C) Systems that can be used without knowledge of internal operations D) None of these	C
13	A definition of a concept is ---- if it recognizes all the instances of that concept A) Complete B) Consistent C) Constant D) None of these	A
14	Which of the following is not a data mining task? Select one: a. Feature Subset Detection b. Regression c. Sequential Pattern Discovery d. Association Rule Discovery	A
15	Which of the following statement is not TRUE for a Tag Cloud a. Tag cloud is a visualization of statistics of user-generated tags b. Tag cloud can be used for numeric data only c. The importance of a tag is indicated by font size or color d. Tags may be listed alphabetically in a tag cloud The correct answer is: Tag cloud can be used for numeric data only	B

16	Which of the following data mining task is known as Market Basket Analysis? a. Clasification b. Regression c. Association Analysis d. Outlier Analysis	C
17	Which of the following activities is a data mining task? a. Monitoring the heart rate of a patient for abnormalities b. Dividing the customers of a company according to their profitability c. Extracting the frequencies of a sound wave d. Predicting the outcomes of tossing a (fair) pair of dice	A
18	Sorted data (attribute values) for price are: 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34. Identify which is NOT a bin smoothed by boundaries? a. Bin 2: 21, 21, 25, 25 b. Bin 1: 4, 4, 4, 15 c. Bin 1: 4, 4, 15, 15 d. Bin 3: 26, 26, 26, 34	C
19	The Correct answer is: unlike unsupervised learning, supervised learning needs labeled data The Data Sets are made up of a. Data Objects b. Attributes c. Dimensions d. Database	A
20	A collection of one or more items is called as _____ (a) Itemset (b) Support (c) Confidence (d) Support Count	A
21	Frequency of occurrence of an itemset is called as _____ (a) Support (b) Confidence (c) Support Count (d) Rules	C

22	An itemset whose support is greater than or equal to a minimum support threshold is _____ (a) Itemset (b) Frequent Itemset (c) Infrequent items (d) Threshold values	B
23	What does FP growth algorithm do? (a) It mines all frequent patterns through pruning rules with lesser support (b) It mines all frequent patterns through pruning rules with higher support (c) It mines all frequent patterns by constructing a FP tree (d) It mines all frequent patterns by constructing an itemsets	C
24	What techniques can be used to improve the efficiency of apriori algorithm? (a) Hash-based techniques (b) Transaction Increases (c) Sampling (d) Cleaning	A
25	A priori algorithm is otherwise called as_____ A.Width-wise algorithm B.Level-wise algorithm C.Pincer-search algorithm D.FP growth algorithm	B
26	Some fields where Apriori is used ____ a) In Education Field: Extracting association rules in data mining of admitted students through characteristics and specialties. b) In the Medical field: For example Analysis of the patient's database. c) In Forestry: d) All of the above	A
27	Some fields where Apriori is used ____ a) In Education Field: Extracting association rules in data mining of admitted students through characteristics and specialties. b) In the Medical field: For example Analysis of the patient's database. c) In Forestry: d) All of the above	D

28	Which method(s) is/are available for improving the efficiency of the algorithm? a) Hash-Based Technique: b) Transaction Reduction: c) Partitioning: d) All of the above	D
29	___ is/are the components comprise the apriori algorithm. a) Support b) Confidence c) Lift d) all of the above	D
30	This set of Operating System Multiple Choice Questions & Answers (MCQs) focuses on “Security – Intrusion Detection”. 1. What are the different ways to intrude? a) Buffer overflows b) Unexpected combinations and unhandled input c) Race conditions d) All of the mentioned	D
31	What are the major components of the intrusion detection system? a) Analysis Engine b) Event provider c) Alert Database d) All of the mentioned	D
32	What are the different ways to classify an IDS? a) anomaly detection b) signature based misuse c) stack based d) all of the mentioned	D
33	What is the major drawback of anomaly detection IDS? a) These are very slow at detection b) It generates many false alarms c) It doesn’t detect novel attacks d) None of the mentioned	B
34	Which of the following is an advantage of anomaly detection? A. Rules are easy to define B. Custom protocols can be easily analyzed C. The engine can scale as the rule set grows D. Malicious activity that falls within normal usage patterns is detected	C

35	What are the different ways to intrude? a) Buffer overflows b) Unexpected combinations and unhandled input c) Race conditions d) All of the mentioned	D
36	The primary intent of BPM is to _____ when performing business processes. a) diagram results b) adjust standards c) record pictures d) optimize efficiency	D
37	The _____ development process demonstrates the relationship between each early phase of development and the associated testing phase. a) spiral b) prototyping c) waterfall d) V-model	D
38	Advances in healthcare technology can cut down the cost of care by _____. a) Utilizing robots instead of clinical professionals b) Replacing the need for staff c) Keeping patients out of the hospital d) Improving old processes	D
39	For what purpose, the analysis tools pre-compute the summaries of the huge amount of data? 1) In order to maintain consistency 2) For authentication 3) For data access 4) To obtain the queries response	D
40	Which of the following statements is incorrect about the hierarchal clustering? a) The hierarchal type of clustering is also known as the HCA b) The choice of an appropriate metric can influence the shape of the cluster c) In general, the splits and merges both are determined in a greedy manner d) All of the above	A

41	In data mining, how many categories of functions are included? a) 5 b) 4 c) 2 d) 3	C
42	Which of the following statements is incorrect about the hierarchal clustering? For the people who have {"kindle", "iphone"}, which type will they be classified as by CBA algorithm? a) Type 1 b) Type 2 c) Both d) None of the above	D
43	There are ____ steps in finding frequent subgraphs. 1) 1 2) 2 3) 3 4) 4	B
44	A graph with all vertices having equal degree is known as a ___ a) Multi Graph b) Regular Graph c) Simple Graph d) Complete Graph	B
45	B2B marketing is fundamentally different from consumer goods or services marketing because: a) distribution channels for business products are significantly longer. b) customer relationships for business products tend to be short-term and transactions-based. c) organizational buyers do not consume the products or services themselves. d) customer service plays a smaller role in the distribution of business products.	C
46	Which of the following is not one of the main factors of business markets? a) The nature of demand. b) The buy phases. c) Buyer-Seller relationships. d) The buying processes.	B

47	_____ is (are) the application of Data Mining. 1) Treatment effectiveness: 2) Healthcare management: 3) Customer relationship management: 4) All of the above	D
48	The data Warehouse is_____. a) read only. b) write only. c) read write only. d) none	A
49	The time horizon in Data warehouse is usually _____. a) 1-2 years. b) 3-4years. c) 5-6 years. d) 5-10 years	D
50	The data is stored, retrieved & updated in _____. a) olap. b) oltp. c) smtp. d) ftp.	B