DATA MINING--REVISION

Time Allowed: 03 Hour(s) Total Marks: 60 Questions in 4 Pages

Answer All the following Questions Model C

Q 1.[35 Marks]: Choose the Correct Answer. *(Use the Answer Sheet and Write the Model Name)*

1. Data warehouse contains ..... data that is never found in the operational environment.
a. normalized b. informational
c. summary d. denormalized

2. Various visualization techniques are used in ..... Step of KDD
a. data mining (b) interpretation
c. transformation d. selection

3. In ..... the groups are not predefined
a. association rules b. prediction
(c) clustering d. summarization

4. Which of the following is not a component of a data warehouse?
a. Data extraction/cleaning/preparation programs
b. Data warehouse data c. Data metadata

5. Reducing the number of attributes to solve the high dimensionality problem is called as .....
(a) Dimensionality reduction b. Cleaning
c. Overfitting d. Dimensionality

6. ..... is the process of finding a model that describes and distinguishes data classes
a. Data Characterization (b) Data Classification
c. Data discrimination d. Data selection

7. The full form of KDD is .....
a. Knowledge Database Design
(b) Knowledge Discovery in Database
c. Knowledge Data Definition

8. The patterns that can be discovered from a given database can be of .....
a. one type only b. no specific type
(c) more than one type d. multiple type always

9. ..... supports basic OLAP operations, including slice and dice, drill-down, roll-up and pivoting.
a. Information processing b. Analytical processing
c. Data mining d. Transaction processing

10. A data warehouse is .....
a. contains numerous naming conventions and formats b. updated by end users.
c. organized around important subject areas
d. contain only current data

11. Nonparametric data reduction strategies include all the following except:
a. Histograms (b) Regression
c. Clustering d. Sampling

12. If you want to give all attributes an equal weight, which preprocess task will you use
a. Cleaning b. integration
(c) Transformation d. Reduction

13. Task of inferring a model from labeled training data is called
a. Cluster analysis b. Association rule
c. Transformation (d) Classification

14. The KDD process consists of ..... steps
a. three b. four (c) five d. six

15. The bottleneck of the Apriori algorithm is caused by all the following except
(a) the number of association rules
b. the number of scans required
c. the computation of support for candidates
d. the number of generated candidates

16. The process of grouping a set of objects into classes of similar objects is called .....
(a) clustering b. classification
c. Association d. all of them

17. Which of the following is the process of detecting and correcting the wrong data
a. data selection (b) data cleaning
c. Data integration d. all of them

18. Which of the following is the process of combining data from different resources
a. Data selection b. Data cleaning
(c) Data integration d. all of them

19. IF the lift measure of the items bred and rice is equal to 0.5. This means that …
a. if clients buy bred they are more likely to buy rice
(b) if clients buy bred they are less likely to buy rice
c. if clients buy bred they can buy rice or not with the same probability d. None of them

20. Correlation analysis is used to:
a. extract association rules
b. define support and confidence values
c. eliminate misleading rules

21. What classifiers are normally considered to be easy to interpret?
a. SVM b. Naïve Bayes
c. Decision trees d. k-Nearest Neighbor

22. If the mean is larger than the median then this might be an indication that the data is .....
a. positively skewed b. negatively skewed
c. symmetric d. correlated

23. ..... is the result of a tuple firing more than one rule with different class predictions
a. Interested rule b. Strong rule
c. Rule conflict d. Association rule

24. Regression is a method of all the following except:
a. Cleaning b. Integration
c. Transformation d. Reduction

25. The terminating conditions in decision tree include all the following except
a. No tuples for a given branch b. No Noise
c. No remaining attributes
d. All tuples belong to the same class

26. Clustering algorithms which can find clusters of arbitrary shape
a. K-means b. DBSCAN
c. Agglomerative d. Divisive

27. If the object deviates significantly from the rest of the dataset, it will be
a. Global Outlier b. local outlier
c. Contextual outlier d. Collective outlier

28. Which of the following are interestingness measures for association rules?
a. Lift b. Compactness c. Recall d. Accuracy

29. For the following association rule:
Computer → Webcam (60%, 100%): Which of the following is true?
I. 100% of customers bought both a computer and a webcam
II. 60% of customers bought both a computer and a webcam
III. 100% of customers who bought a computer bought also a webcam
IV. 60% of customers who bought a computer bought also a webcam
a. II only b. III Only c. I and IV d. II and III

30. In K-nearest neighbor algorithm K stands for .....while in K-means algorithm K stand for .....
a. no. of cluster, no. of neighbors'
b. no. of neighbors, no. of cluster
c. no. of rules, no. of classes

31. ..... is the correlation analysis measure for nominal attribute
a. Covariance b. Chi-square
c. Lift d. Correlation Coefficient

32. Clustering algorithm which can find clusters of spherical shape are .....
a. K-mean b. DBSCAN
c. Agglomerative d. KNN

33. Which of the following clustering requires merging approach?
a. Partitional b. Hierarchical
c. Density-based d. Grid-based

34. Which of the following is required by K-means clustering?
a) defined distance metric b) number of clusters
c) initial guess as to cluster centroids d) all of them

35. What does the term 'outlier' mean?
a. A score that is left out of the analysis because of missing data b. The arithmetic mean
c. A type of variable that cannot be quantified
d. An extreme value at either end of a distribution

36. Which of the following is finally produced by Hierarchical Clustering?
a. final estimate of cluster centroids
b. tree showing how close things are to each other
c. assignment of each point to clusters
d. all of them

37. Point out the wrong statement:
a. k-means clustering is density based method
b. k-means clustering aims to partition n observations into k clusters
c. k-nearest neighbor is same as k-means
d. None of them

38. We have Market Basket data for 1,000 rental transactions at a Video Store. There are four videos for rent -- Video A, Video B, Video C and Video D. The probability that both Video C and Video D are rented at the same time is known as _____.
a. Correlation b. support c. lift d. confidence

39. A core point in DBSCAN is an Object whose ϵ-neighborhood contains ..... objects
a. at most MinPts b. at least MinPts c. MinPts

From the given Confusion Matrix what the value of the following:

Confusion Matrix				
		Predicted		Total
		Yes	No	
Actual	Yes	6954	46	7000
	No	412	2588	3000
Total		7366	2634	10000

40. Accuracy is
a. 0.99 b. 0.95
c. 0.86 d. 0.05

41. Error rate is
a. 0.99 b. 0.95
c. 0.86 d. 0.05

42. Sensitivity is
a. 0.99 b. 0.95 c. 0.86 d. 0.05

43. Specificity is
a. 0.99 b. 0.95 c. 0.86 d. 0.05

44. Which of the following lists all parts of the five number summary?
a. Mean, Median, Mode, Range, and Total
b. Minimum, Quartile1, Median, Quartile3, and Maximum
c. Smallest, Q1, Q2, Q3, and Q4
d. Minimum, Maximum, Range, Mean, and Median

45. If the information gain of age, income and sex attributes are 0.42, 0.24 and 0.024 respectively which one you will chose as the splitting attribute.
a. age b. income c. gender d. all of them

46. Base on the Apriori property, All nonempty subsets of a frequent itemset.....
a. must also be frequent b. can't be frequent
c. may be frequent d. all of them

for the given transaction database: Suppose that minimum support is 40% and minimum confidence 70%.

TID	Items
T100	A, B, C, D
T200	A, B, C, E
T300	A, B, E, F, H
T400	A, C, H

47. The support of the item set A, B, E is.....
a. 50% b.40% c. 70% d. 66%

48. Based on the given minimum support, the item set A,B,E is.....
a. frequent b. not frequent c. strong d. not strong

49. The confidence of the rule A, B → E is
a. 50% b.40% c. 100% d. 66%

50. Based on the given minimum confidence the rule A, B → E is.....
a. frequent b. not frequent c. strong d. not strong

51. The lift of the rule A, B → E is.....
a. 1.33 b.1 c. 0.89 d. 0.66

52. The value of the lift in the previous question means that items are.....
a. positive correlated b. negative correlated
c. independent d. strong

53. For the given data {33, 25, 42, 25, 31, 37, 46, 29, 38} the five numbers summery will be ...
a. 25, 27, 32, 35, 46 b. 25, 27, 33, 35, 46
c. 14, 27, 33, 35, 47 d. 19, 29, 32, 38, 43

54. If you use min-max normalization to transform the value 33 onto the range [1.0, 2.0] the new value
a. 0.38 b. 1.38 c. 0.038 d. 1.038

55. Identify the outlier for the given data?
23, 34, 27, 7, 30, 26, 28, 31, 34
a. 7 b. 23 c. 31 d. 34

Assume, you want to cluster 7 observations into 3 clusters using K-Means clustering algorithm. After first iteration clusters, C1, C2, C3 has following observations: C1: {(2, 2), (4, 4), (6, 6)}
C2: {(0, 4), (4, 0)} C3: {(5, 5), (9, 9)}

56. What will be the cluster centroids if you want to proceed for second iteration?
a. C1: (4, 4), C2: (2, 2), C3: (7, 7)
b. C1: (6, 6), C2: (4, 4), C3: (9, 9)
c. C1: (2, 2), C2: (0, 0), C3: (5, 5)
d. None of them

57. What will be the Manhattan distance for observation (9, 9) from cluster centroid C1 in second iteration?
a. 10 b. $5\sqrt{2}$ c. $13\sqrt{2}$ d. None of them

58. Consider the given data: {3, 4, 5, 10, 21, 32, 43, 44, 46, 52, 59, 67}, Using equal-width partitioning and four bins, how many values are there in the first bin?
a. 3 b. 4 c. 5 d. 6

59. If smooth by median is applied to the previous bins, what is the new value of the data in the first bin
a. 4 b. 4.5 c. 5 d. 7.5

60. What is the Naive Bayes Algorithm used for?
a. Generate mining models
b. Estimating the probability of a class value during classification and prediction
c. To make decisions for reporting. d. Both a and b

61. Data used to build a data mining model.
a. validation data b. training data
c. test data d. hidden data

62. Which statement is true about the K-Means algorithm?
a. All attribute values must be categorical.
b. The output attribute must be cateogrical.
c. Attribute values may be either categorical or numeric.
d. All attributes must be numeric.

63. Supervised learning differs from unsupervised clustering in that supervised learning requires
a. at least one input attribute.
b. input attributes to be categorical.
c. at least one output attribute.
d. output attributes to be categorical.

64. The attribute data type of the number of telephones in your house is .....
(a) Nominal b. ordinal c. interval d. ratio

65. This approach is best when we are interested in finding all possible interactions among a set of attributes.
a. decision tree (b) association rules
c. K-Means algorithm d. genetic learning

66. This technique uses mean and standard deviation scores to transform real-valued attributes.
(a) z-score normalization b. min-max normalization
c. decimal scaling d. logarithmic normalization

67. Point out the correct statement:
a. Combining classifiers improves interpretability
b. Combining classifiers reduces accuracy
(c) Combining classifiers improves accuracy
d. All of them

68. Selecting data so as to assure that each class is properly represented in both the training and test set.
a. cross validation (b) stratification
c. verification d. bootstrapping

69. This clustering algorithm initially assumes that each data instance represents a single cluster.
(a) agglomerative clustering b. conceptual clustering
c. K-Means clustering d. expectation maximization

70. Which statement about outliers is true?
(a) It should be identified and removed from a dataset.
b. It should be part of the training dataset but should not be present in the test data.
c. It should be part of the test dataset but should not be present in the training data.
(d) The nature of the problem determines how outliers are used.

Q2 [25 Marks]: True or False; correct the error

X 1. Binning is a method of reduction
X 2. Data mining is extraction of interesting (trivial, implicit, previously known and potentially useful) patterns or knowledge from huge amount of data
X 3. In lazy learner we interest in the largest distance.
X 4. Sorting a student database based on student identification numbers. Is a data mining task
✓ 5. Association rules provide information in the form of "if-then" statements.
X 6. Data matrix stores a collection of proximities for all pairs of n objects as an n-by-n matrix
T 7. K-Nearest Neighbor Classifiers do classification when new test data is available
T 8. In decision tree algorithms, attribute selection measures are used to rank attributes
T 9. Intrinsic methods measure how well the clusters are separated
10. The silhouette coefficient is a method to determine the natural number of clusters for hierarchical algorithms — F
F 11. Multimedia Mining is the application of data mining techniques to discover patterns from the Web
F 12. If all the proper subsets of an itemset are frequent, then the itemset itself must also be frequent.
F 13. For an association rule, if we move one item from the right-hand-side to the left-hand-side of the rule, then the confidence will never change.
F 14. The Pruning make the decision tree more complex
F 15. An object is an outlier if its density is equal to the density of its neighbors.
F 16. the object is local outlier if it is deviate significantly from the rest of the dataset
F 17. Binary variables are sometimes continuous
18. An OLAP system focuses mainly on the current data within an enterprise
19. A data cube allows data to be modeled and viewed in single dimension
20. A data warehouse is a subject-oriented, isolated, time-variant, and volatile collection of data
21. An OLAP system is customer-oriented
22. OLTP is used for decision support
23. In Star schema, a fact table is located in the middle connected to a set of dimension tables
24. The access patterns of an OLAP system consist mainly of short and atomic transactions
F 25. Euclidean distance is used to measure discimilarity of Nominal attributes

End of Exam
My Best Wishes

Cleaning	Integration	Reduction	Transformation/Discretization
Binning			Binning
Regression		Regression	Regression
	Correlation analysis		Correlation
		Histograms	Histogram analysis
		Clustering	Clustering
		Attribute construction	Attribute construction
			Aggregation
			Normalization
Outlier analysis			
		Wavelet transforms	
		PCA	
		Attribute subset selection	
		Sampling	
			Concept hierarchy

Choose the Correct Answer.

1. For the following association rule: Computer → Webcam (60%, 100%): Which of the following is true?

I. 100% of costumers bought both a computer and a webcam

II. **60% of costumers bought both a computer and a webcam**

III. **100% of costumers who bought a computer bought also a webcam**

IV. 60% of costumers who bought a computer bought also a webcam

a. II only b III Only c. I and IV **d. II and III**

2. We have Market Basket data for 1,000 rental transactions at a Video Store. There are four videos for rent -- Video A, Video B, Video C and Video D. The probability that both Video C and Video D are rented at the same time is known as _____ .

a. Correlation **b. support** c. lift d. confidence

Consider the following transaction database: Suppose that minsup is set to 40% and minconf. to 70%.

TransID	Items
T100	A, B, C, D
T200	A, B, C, E
T300	A, B, E, F, H
T400	A, C, H

3. The support of the item set A, B, E is……

a. 50% b.40% c. 70% d. 66%

4. Based on the given minimum support the item set A,B,E is……..

a. frequent b. not frequent c. strong d. not strong

5. The confidence of the rule A, B→E is

a. 50% b.40% c. 100% **d. 66%**

6. Based on the given minimum confidence the rule A, B→E is……..

a. frequent b. not frequent c. strong **d. not strong**

7. The lift of the rule A, B→E is…….

a. 1.33 b.1 c. 0.89 d. 0.66

8. The value of the lift in the previous question means that items are…….

a. positive correlated b. negative correlated

c. independent d. strong

9. Identify the outlier for the given data? 23, 34, 27, 7, 30, 26, 28, 31, 34

a. 7 b. 23 c. 31 d. 34

From the given Confusion Matrix

Confusion Matrix				
		Predicted		Total
		Yes	No	
Actual	Yes	6954	46	7000
	No	412	2588	3000
Total		7366	2634	10000

10. Accuracy is…….

a.0.99 **b.0.95** c.0.86 d.0.05

11. Error rate is…….

a.0.99 b.0.95 c.0.86 **d.0.05**

12. Sensitivity is…….

a.0.99 b.0.95 c.0.86 d.0.05

13. Specificity is …….

a.0.99 b.0.95 **c.0.86** d.0.05

Assume, you want to cluster 7 observations into 3 clusters using K-Means clustering algorithm. After first iteration clusters, C1, C2, C3 has following observations: **C1: {(2, 2), (4, 4), (6, 6)} C2: {(0, 4), (4, 0)} C3: {(5, 5), (9, 9)}**

14. What will be the cluster centroids if you want to proceed for second iteration?

a. C1: (4, 4), C2: (2, 2), C3: (7, 7) b. C1: (6, 6), C2: (4, 4), C3: (9, 9)

c. C1: (2, 2), C2: (0, 0), C3: (5, 5) d. None of these

15. What will be the Manhattan distance for observation (9, 9) from cluster centroid C1 in second iteration?

a. 10 b. $5\sqrt{2}$ c. $13\sqrt{2}$ d. None of these

16. Consider the given data: {3, 4, 5, 10, 21, 32, 43, 44, 46, 52, 59, 67}, Using equal-width partitioning and four bins, how many values are there in the first bin?

a. 3 **b. 4** c. 5 d. 6

17. If smooth by median is applied to the previous bins, what is the new value of the data in the first bin?

a. 4 **b. 4.5** c. 5 d. 7.5

18. The correlation between the number of years an employee has worked for a company and the salary of the employee is 0.75. What can be said about employee salary and years worked?

a. There is no relationship between salary and years worked.

b. **Individuals that have worked for the company the longest have higher salaries.**

c. Individuals that have worked for the company the longest have lower salaries.

d. The majority of employees have been with the company a long time.

e. The majority of employees have been with the company a short period of time.

19. The correlation coefficient for two real-valued attributes is –0.85. What does this value tell you?

A. The attributes are not linearly related.

B. As the value of one attribute increases the value of the second attribute also increases.

C. **As the value of one attribute decreases the value of the second attribute increases.**

D. The attributes show a curvilinear relationship

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

A. Data warehousing

C. Text mining

B. **Data mining**

D. Data selection

21. Data mining is best described as the process of

a. identifying patterns in data.

b. deducing relationships in data.

c. representing data.

d. simulating trends in data.

22. Unlike traditional production rules, association rules

a. allow the same variable to be an input attribute in one rule and an output attribute in another rule.

b. allow more than one input attribute in a single rule.

c. require input attributes to take on numeric values.

d. require each rule to have exactly one categorical output attribute.

23. Given desired class C and population P, lift is defined as

a. the probability of class *C* given population *P* divided by the probability of *C* given a sample taken from the population.

b. the probability of population *P* given a sample taken from *P.*

c. the probability of class *C* given a sample taken from population *P*.

d. **the probability of class *C* given a sample taken from population *P* divided by the probability of *C* within the entire population *P*.**

24. Association rule support is defined as

a. the percentage of instances that contain the antecendent conditional items listed in the association rule.

b. the percentage of instances that contain the consequent conditions listed in the association rule.

c. **the percentage of instances that contain all items listed in the association rule.**

d. the percentage of instances in the database that contain at least one of the antecendent conditional items listed in the association rule.

25. The full form of KDD is .....

A. Knowledge Database

B. **Knowledge Discovery Database**

C. Knowledge Data House

D. Knowledge Data Definition

26. This approach is best when we are interested in finding all possible interactions among a set of attributes.

a. decision tree

b. **association rules**

c. K-Means algorithm

d. genetic learning

27. If the information gain of age, income and gender attributes are 0.42, 0.24 and 0.024 which one will you choose as splitting attribute

a. **age**

b. income

c. gender

d. all of them

28. Based on Apriori property all nonempty subsets of frequent itemset:

a. **must also be frequent**

b. may be frequent

c. can't be frequent

d. all of them

29. Reducing the number of attributes to solve the high dimensionality problem is called

a. cleaning

b. over fitting

c. **dimensionality reduction**

d. Dimensionality

30. The bottleneck of the Apriori algorithm is caused by all the following except

a. the number of association rules

b. the number of scans required

c. **the computations of support for candidates**

d. the number of generated candidates

31. Which of the following is the process of detecting and correcting wrong data:

a. **data cleaning**

b. data selection

c. data integration

d. all of them

32. Which of the following is the process of combining data from different sources:

a. data cleaning

b. data selection

c. **data integration**

d. all of them

33. Which of the following are interesting measures for association rules:

a. lift

b. Recall

c. Accuracy

d. Compactness

34. If the lift measure of items bred and rice if equal 0.5 this means that:

a. if client buy bred they are more likely to but rice

b. **if client buy bred they are less likely to but rice**

c. if client buy bred they can buy rice or not with the same probability

d. none of them

35. Nonparametric data reduction strategies include all the following except:

a. Histograms

b. Clustering

c. Sampling

d. **Regression**

36. If you want to give all attributes equal weight, which preprocess task you will use:

a. Cleaning

b. Transformation

c. Integration

d. Reduction

37. Task of inferring a model from labeled training data is called:

a. Transformation

b. Cluster analysis

c. **Classification**

d. Association rues

38. Regression is a method of the following except:

a. Cleaning

b. Transformation

c. Reduction

d. **Integration**

39. The termination condition of the decision tree include the following except:

a. No tuples for a given branch

b. **No noise**

c. No remaining attributes

d. All tuples belong to the same class

40. Correlation analysis is used to:

a. extract association rules

b. define support and confidence values

c. **eliminate misleading rules**

41. If the mean is larger than the median then this might be an indication that the data is

a. negatively skewed

b. <u>positively skewed</u>

c. symmetric

d. correlated

42. _____is the result of tuple firing more than one rule with different class prediction.

a. Association rule

b. Strong rule

c. **<u>Rule conflict</u>**

43. _____ is the correlation analysis measure for nominal attributes.

a. Covariance

b. **<u>Chi-square</u>**

c. Lift

d. Correlation co-efficient

44. We have Market Basket data for 1,000 rental transactions at a Video Store. There are four videos for rent -- Video A, Video B, Video C and Video D. The probability that Video D will be rented given that Video C has been rented is known as _____ .

<u>A.</u> the basic probability

<u>B.</u> support

<u>C.</u> lift

<u>D.</u> **<u>confidence</u>**

45. Data used to build data mining model:

a. Validation data

d. Hidden data

b. Test data

c. **<u>Training data</u>**

46. This technique uses mean and standard deviation scores to transform real-valued attributes.
    a. decimal scaling
    b. min-max normalization
    c. **z-score normalization**
    d. logarithmic normalization

47. Point out the correct statement:
    a) Combining classifiers improves interpretability
    b) Combining classifiers reduces accuracy
    c) **Combining classifiers improves accuracy**
    d) All of the Mentioned

48. The attribute data type of Number of telephones in your house is
    a. **Nominal** **c.** Interval
    b. Ordinal **d.** Ratio

49. Which of the following best describes the process of finding the interquartile range for a set of data?
    a. ADD the biggest and smallest numbers
    b. Place the number in order from least to greatest then find the middle.
    c. Find the difference between the Maximum and the Minimum.
    d. **Subtract Q3 from Q1**

50. What is the term for the median of the lower half of the data?
    a. **Lower Quartile**
    b. Upper Quartile
    c. Median
    d. Maximum

51. What is the term that means the middle data value?
    a. Mean
    b. **Median**
    c. Mode
    d. Range

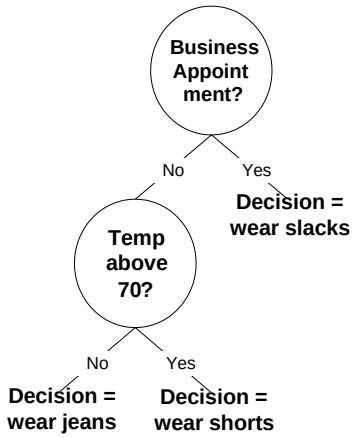
52. What is the mode?
    a. **# happening the most**
    b. the average
    c. biggest - smallest
    d. the middle #
53. The heights of some students are given.

    158cm 172cm 164cm

    164cm 167cm 159cm

    What is the range of the heights?

    a. 13cm
    b. **14cm**
    c. 164cm
    d. 330cm
54. Supervised learning and unsupervised clustering both require at least one
    a. hidden attribute.
    b. output attribute.
    c. **input attribute.**
    d. categorical attribute.
55. Supervised learning differs from unsupervised clustering in that supervised learning requires
    a. at least one input attribute.
    b. input attributes to be categorical.
    c. **at least one output attribute.**
    d. ouput attriubutes to be categorical.
56. Which of the following is a valid production rule for the decision tree below?



    a. IF Business Appointment = No & Temp above 70 = No

THEN Decision = wear slacks

b. IF Business Appointment = Yes & Temp above 70 = Yes

THEN Decision = wear shorts

c. IF Temp above 70 = No

THEN Decision = wear shorts

d. ***IF Business Appointment= No & Temp above 70 = No*** THEN Decision = wear jeans

57. A nearest neighbor approach is best used

a. with large-sized datasets.

b. **when irrelevant attributes have been removed from the data.**

c. when a generalized model of the data is desireable.

d. when an explanation of what has been found is of primary importance.

58. If a customer is spending more than expected, the customer's intrinsic value is _____ their actual value.

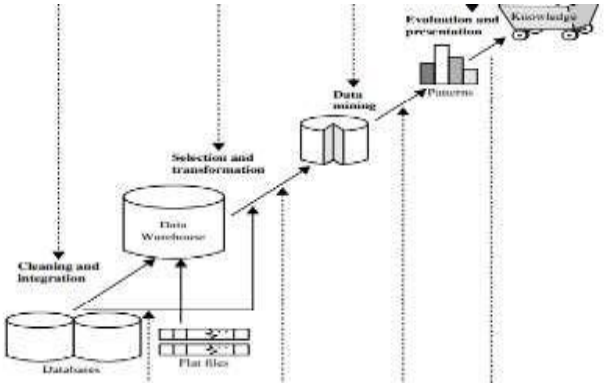
a. greater than

b. **less than**

c. less than or equal to

d. equal to

59. . ..... 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

60. Data mining can also applied to other forms such as ..... i) Data streams ii) Sequence data iii) Networked data iv) Text data v) Spatial data
    a. i, ii, iii and v only
    b. ii, iii, iv and v only
    c. i, iii, iv and v only
    d. ***All i, ii, iii, iv and v***

61. Which of the following is not a data mining functionality?
    a. Characterization and Discrimination
    b. Classification and regression
    c. **Selection and interpretation**
    d. Clustering and Analysis

62. ..... 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

63. ..... 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

64. Strategic value of data mining is .....
    a. cost-sensitive
    b. work-sensitive
    c. **time-sensitive**
    d. technical-sensitive

65. ..... is the process of finding a model that describes and distinguishes data classes or concepts.
    a. Data Characterization
    b. **Data Classification**
    c. Data discrimination
    d. Data selection

66. The various aspects of data mining methodologies is/are .....

i) Mining various and new kinds of knowledge
ii) Mining knowledge in multidimensional space
iii) Pattern evaluation and pattern or constraint-guided mining.
iv) Handling uncertainty, noise, or incompleteness of data

a. i, ii and iv only
b. ii, iii and iv only
c. i, ii and iii only
d. **All i, ii, iii and iv**

67. The out put of KDD is .....

a. Data
b. Information
c. Query
d. **Useful information**

68. 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

69. 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

70. If the mean, median and mode of a distribution are 5, 6, 7 respectively, then the distribution is:

a. ***skewed negatively***
b. not skewed
c. skewed positively
d. symmetrical
e. bimodal.

71. Which of the following measures of central tendency tends to be most influenced by an extreme score?

a. median
c. ***mean***
b. mode

72. Which of the following is not a measure of central tendency?

a. mean
d. ***standard deviation***
b. median
e. none of these
c. mode

73. In a group of 12 scores, the largest score is increased by 36 points.What effect will this have on the mean of the scores?

a. it will be increased by 12 points
b. it will remain unchanged
c. ***it will be increased by 3 points***
d. it will increase by 36 points
e. there is no way of knowing exactly how many points the mean will be increased.

74. Non-parametric data reduction strategies includes all the following except

a-Histogram b- **regression** c- clustering d- sampling

75. If you want to give all attributes an equal weight which preprocess task you will use

a-Cleaning b-integration c-**transformation** d-reduction

76. Regression is a method of all of the following except

a-Cleaning b- ***integration*** c-transformation d-reduction

77. Which of the following lists all parts of the five-number summary?

a. Mean, Median, Mode, Range, and Total
b. **Minimum, Quartile1, Median, Quartile3, and Maximum**
c. Smallest, Q1, Q2, Q3, and Q4
d. Minimum, Maximum, Range, Mean, and Median

True or False

1. All continuous variables are ratio

    a. True (all numbers)

    b. **False**

2. Attributes are sometimes called variables and objects are sometimes called observations

    a. True (the opposite)

    b. False

3. Computing the total sales of a company. Is a data mining task?

    a. True (accounting)

    b. **False**

4. Dissimilarity matrix stores n data objects that have p attributes as an n-by-p matrix

    a. True (n-by-n)

    b. False

5. Dividing the customers of a company according to their profitability. is a data mining task?

    a. True (accounting)

    This is an accounting calculation, followed by the application of a threshold. However, predicting the profitability of a new customer would be data mining.

    b. **False**

6. In decision tree algorithms, attribute selection measures are used to reduce the dimensionality

    a. **True**

    b. False

7. Strategies for data transformation include chi-square test

    a. True (integration)

    b. **False**

8. Accuracy is interestingness measures for association rules

    a. True (correlation, lift)

    b. **False**

9. Correlation is a method of cleaning

    a. True (integration)

    b. **False**

10. Extracting the frequencies of a sound wave. Is a data mining task?

    a. True (signal processing)

    b. **False**

11. Median is a value that occurs most frequently in the attribute values

    a. True (mode)

    b. **False**

12. Euclidean distance is used to measure dissimilarity of nominal attributes:

    a. True (numeric)

    b. False

13. Binning is a method of reduction

    a. True (cleaning)

    b. False

14. Data mining is an extraction of interesting (trivial, implicit, previously known and useful) patterns or knowledge from huge amount of data.

    a. True (the opposite)
    b. **False**

15. Sorting a student database based on student identification number is a data mining task

    a. True (database query)

    b. <u>False</u>

16. Association rules provide information in the form of "if-then" statements

    <u>a. True</u>

    b. False

17. Data matrix stores a collection of proximities for all pairs of n objects as an n-by-n matrix

    a. True (Dissimilarity)

    b. <u>False</u>

18. If all proper subsets of an itemset are frequent, then the itemset itself must also be frequent.

    a. True (may)
    b. **<u>False</u>**

19. For an association rule, if we move one item from right hand side to the left hand side of the rule, then the confidence will never change

    a. True (support)
    b. **<u>False</u>**

20. Binary variables are sometimes continuous

    a. True (numeric with two values)
    b. **<u>False</u>**

21. You must find the 5 number summary in order to make a box and whisker plot.

    a. True (boxplot)

    b. <u>False</u>

22. Correlation analysis can be used to eliminate misleading rules

    a. True

    b. False

23. The bottleneck of apriori algorithm caused by the number of association rules

    a. True (number of candidate and scan)

    b. False

24. Strong Rules Are Not Necessarily Interesting

    a. True

    b. False

25. There is no difference between noise and outlier

    a. True (noise and outlier are different)

    b. False

26. Boxplot used for data smoothing

    a. True (Description)

    b. False

27. Incomplete data problem can be solved by binning

    a. True (binning is a smoothing technique)

    b. False

28. Sampling used for smoothing

    a. True (reduction)

    b. False

29. Cluster is the process of finding a model that describes and distinguishes data classes or concepts.

    a. True (classification)

    b. False

30. Correlation analysis divides data into groups that are meaningful, useful, or both.

    a. True (Cluster Analysis)

    b. False

31. Database mining refers to the process of deriving high-quality information from text.

    a. True (Text Mining)

    b. False

32. In decision tree algorithms, attribute selection measures are used to rank attributes

    a. True

    b. False

33. In lazy learner we interest in the largest distance.

    a. True (minimum)

    b. False

34. Intrinsic methods measure how well the clusters are separated

    a. True

    b. False

35. Multimedia Mining is the application of data mining techniques to discover patterns from the Web.

    a. True (Web data mining)

    b. False

36. Regression is a method of integration

    a. True (Correlation)

    b. False

37. The Pruning make the decision tree more complex

    a. True (reliable)

    b. False

38. An object is an outlier if its density is equal to the density of its neighbors.

    a. True (far away)

    b. False

39. A common weakness of association rule mining is that it is not produce enough interesting rules

    a. True (Too many rules)

    b. False

40. Core object is an object whose ϵ -neighborhood contains objects less than MinPts (F)

    a. True (more than)

    b. False

41. K-Nearest Neighbor Classifiers do classification when new test data is available

    a. True

    b. False

42. Mode is a middle value in set of ordered values

    a. True (median)

    b. False

43. One strength of a Bayesian Classifier is that it can be easily trained

    a. True

    b. False

44. Outlier analysis is a method of transformation

    a. True (Cleaning)

    b. False

45. Predicting the outcomes of tossing a (fair) pair of dice. Is a data mining task?

    a. True

    Answer: No. Since the die is fair, this is a probability calculation. If the die were not fair, and we needed to estimate the probabilities of each outcome from the data, then this is more like the

problems considered by data mining. However, in this specific case, solutions to this problem were developed by mathematicians a long time ago, and thus, we wouldn˜t consider it to be data mining.

b. False

46. Recall is interestingness measures for association rules

    a. True (lift)

    b. False

47. Redundancy is an important issue in data cleaning

    a. True (integration)

    b. False

48. Sampling methods smooth noisy data

    a. True (reduction)

    b. False

49. The goal of clustering analysis is to maximize the number of clusters

    a. True

    Maximize intra-cluster similarity and minimize inter-cluster similarity

    b. False

50. the object is local outlier if it is deviate significantly from the rest of the dataset

    a. True (global)

    b. False

51. The silhouette coefficient is a method to determine the natural number of clusters for hierarchical algorithms

    a. True (partitioning algorithms)

    b. False