PG DEPARTMENT OF COMPUTER APPLICATIONS

DATA MINING AND WAREHOUSING – 18PMC533

K1 LEVEL QUESTIONS

Unit – I

1. Expand OLTP

   a) Online Transaction Processing b) Online Transform Processing

   c) Open Transaction Processing d) None of the above

2. Multiple data sources may be combined

   a) Data Integration b) Data Cleaning

   c) Data Transformation d) None of the above

3. A _____ is a collection of tables, each of which is assigned a unique name.

   a) Relational database b) Spatial Database

   c) Time series d) None of the above

4. A _____ is a department subset of a data warehouse

   a) Data Mart b) Data Cube

   c) Data Design d) None of the above

5. A _____ typically stores relational data that include time-related attributes.

   a) Sequence **b) Temporal**

   c) Time series d) None of the above

6. _____store image, audio, and video data.

   a) Multimedia databases b) Combine Database

   c) Heterogeneous Database d) None of the above

7. _____is a summarization of the general characteristics or features of a target class of data

   a) Data characterization b) Data Summarization

   b) Data Clustering d) None of the above

8. Regression analysis is a _____ methodology

   a) Statistical b) Numerical

   b) Categorical methodology d) None of the above

9. _____ represents the percentage of transactions from a transaction database that the given rule satisfies

   a) Confidence **b) Support**

   b) Modelling d) None of the above

10. A _____ is a measure that must be computed on the entire data set as a whole.

a) Holistic measure b) Distributive measure

c) Algebraic measure d) None of the above

Unit – II

11. _____ allows data to be modeled and viewed in multiple dimensions.

a) Data Mart **b) Date Cube**

c) Algebraic measure d) None of the above

12. The cuboid that holds the lowest level of summarization is called the _____

a) Base cuboid b) Cuboid

c) Cube d) None of the above

13. Expand DMQL

a) Data Mining Query Language

b) DataMining Quick Language

c) DataMining Query List

d) None of the above

14. _____ is a visualization operation that rotates the data axes in view in order to provide an alternative presentation of the data.

a) **Pivot** b) Roll Down

c) Scroll d) None of the above

15. A statistical database is a database system that is designed to support

a) Numerical Database **b) Statistical Database**

c)Temporal d) None of the above

16. The data are extracted using application program interfaces known as

a) gateways b) interface

c) intermediate d) None of the above

17. A virtual warehouse is a set of views over _____ databases.

a) Time Series **b) Operational**

c) Spatial d) None of the above

18. _____ are data about data

a) MetaData b) Self Data

c) Inner Data d) None of the above

19. Expand MOLAP

a) Multidimensioanl OLAP b) Multiple OLAP

c) Mutual OLAP d) None of the above

20. Expand OLAM

a) OnLine Analytical Mining b) OnLine Analysis Mining

c) OnLine Analytical Metadata d) None of the above

Unit – III

21. Class label of each training tuple is not known

a) Supervised **b) Unsupervised**

c) Clustering d) None of the above

22. Learning of decision trees from class labelled training tuples

a)**Decision tree induction** b)Association

c)Classification d) None of the above

23. Expand CART

a)**Classification and Regression Trees**

b)Classification and Retrieval Trees

c)Clustering and Regression Trees

d) None of the above

24. ID3 uses information gain assets

a)Numeric selectiom measure

b)Attribute selevtion measure

c)None of the above

25. C4.5 is

a)Successor of ID3

b)Predecessor of ID3

c)Gain ratio

d)None of the above

26. Write an example of post pruning approach

A)Time complexity

B)Cost complexity

C)Space complexity

D)None of the above

27. Write the decision tree algorithm that takes completely different approach to scale ability

a)BOAT

b)BAOT

c)BTOA

d)None of the above

28. Regression was developed by

a)Sir Francis Galton

b)Charles Darwin

c)James Gosling

d)None of the above

29. C4.5 adopts

a)Class based ordering scheme

b)Rule based ordering scheme

c)False based ordering scheme

d)None of the above

30. Neral network learning is also referred as

a)Non connectionist learning

b)Connectionist learning

c)Divert learning

d)None of the above

Unit – IV

31. Clustering is the process of grouping a set of data objects into multiple groups

a) Analysing

b) Clustering

c) Partitioning

d) Processing

32. Clustering is also called data segmentation

a) OutLier

b) data

c) data segmentation

d) Detection

33. In data mining, efforts have focused on finding methods for efficient and effective cluster analysis in large databases.
    a) Small database
    b) large database
    c) distance based
    d) database
34. Traditional partitioning methods can be extended for subspace clustering, rather than searching the full data space.
    a) Advanced Partitioning
    b) hierarchical Partitioning
    c) Partitioning
    d) Traditional Partitioning
35. Hierarchical methods suffer from the fact that once a step (merge or split) is done, it can never be undone.
    a) Density-based
    b) Hierarchical
    c) Partition
    d) Grid-based
36. To deal with larger data sets, a sampling-based method called CLARA
    a) CLARA
    b) PAM
    c) Random
    d) Sequential
37. A tree structure called a dendrogram is commonly used to represent the process of hierarchical clustering.
    a) Dendrogram
    b) DIANA
    c) Single-Linkage
    d) None of the above
38. An agglomerative hierarchical clustering method uses a bottom-up strategy
    a) Bottom-up
    b) bottom-down
    c) top-down
    d) None of the above

39. A divisive hierarchical clustering method employs a top-down strategy.
    a) Bottom-up
    b)bottom-down
    c) top-down
    d) None of the above
40. BIRCH is designed for clustering a large amount of numeric data by integrating hierarchical clustering and other clustering methods such as iterative partitioning
    a) BIRCH
    b) ROCK
    c) OPTICS
    d) Chameleaon

Unit – V

41. One important direction toward improving the overall efficiency of the mining process while increasing user interaction is **constraint-based mining.**
    a) Constraint b) Constraint-based
    c) STING Based. D) None of the above
42. The exploration of data mining for businesses continues to expand as e-commerce and e-marketing have become main stream in the retail industry.
    a) Exploration b) Constraint
    c) Isolation D) None of the above.
43. A systematic development of such techniques will facilitate the promotion of human participation for effective and efficient data analysis.
    a) Systematic b) Exploration
    c) Traditional D) None of the above
44. Traditional data mining methods, designed to work at a centralized location, do not work well in many of the distributed computing environments present today.
    a) Traditional b) Real-time
    c) Distributed. D) None of the above
45. Classification and clustering methods can be used for customer group identification and targeted marketing.
    a) Classification b) Cluster
    c) Distributed. D) None of the above

46. The retail industry is a well-fit application area for data mining, since it collects huge amounts of data on sales, customer shopping history, goods transportation, consumption, and service.

a) Retail b) Financial

c) both D) None of the above

47. Which is the IBM datamining Product

a) Intelligent Miner b) Mineset

c)Clementine d)None of the above

48. _____is developed at University of Waikato

a) SPSS **b)Weka**

c)R d) None of the above

49. _____ is the building blocks of proteins

a) Amino Acids b)Cell

c) Plasma d)None of the above

50. DNA sequences are comprised of _____ blocks

a) 4 b)3

c)2 d)1

PG DEPARTMENT OF COMPUTER APPLICATIONS

DATA MINING AND WAREHOUSING – 18PMC533

K2 LEVEL QUESTIONS

Unit- I

1. **Define Datawarehouse**

   Data warehouse a repository of multiple heterogeneous data sources organized under a unified schema at a single site in order to facilitate management decision making

2. **Define Data Mining**

   Extracting the hidden data from a very large database

3. **What is Sequence Database**

   A sequence database stores sequences of ordered events, with or without a concrete notion of time.

4. **What is Spatiotemporal database**

   A spatial database that stores spatial objects that change with time is called a spatiotemporal database

5. **What do you mean by No Coupling**

   No coupling means that a DM system will not utilize any function of a DB or DW system

6. **Define Data Reduction**

   Data reduction obtains a reduced representation of the data set that is much smaller in volume, yet produces the same (or almost the same) analytical results.

7. **List out the measures of central tendency**

   Measures of central tendency include mean, median, mode, and midrange

8. **Write the measures of central tendency**

   Distributive measure, algebraic measure, and holistic measure.

9. **What is noise?**

   Noise is a random error or variance in a measured variable

10. **Define DataScrubbing**

    Datascrubbing tools use simple domainknowledge (e.g.,knowledge of postal addresses, and spell-checking) to detect errors and make corrections in the data.

Unit- II

11. Distinguish data warehouses from other data repository systems

The four keywords, subject-oriented, integrated, time-variant, and non volatile, distinguish data warehouses from other data repository systems

12. What is a data cube?

A data cube allows data to be modeled and viewed in multiple dimensions. It is defined by dimensions and facts.

13. List the multidimensional model

Star schema, a snowflake schema, or a fact constellation schema.

14. Define DataMart

A data mart is a department subset of the data warehouse that focuses on selected subjects.

15. What is a data cube measure

A data cube measure is a numerical function that can be evaluated at each point in the data cube space.

16. Write the examples of holistic functions

median(), mode(), and rank()

17. Name the four different views regarding the design of datawarehouse

The top-down view, the data source view, the data warehouse view, and the business query view.

18. Define waterfall model

The waterfall method performs a structured and systematic analysis at each step before proceeding to the next, which is like a waterfall, falling from one step to the next.

19. What are the types of OLAP Server

ROLAP , MOLAP and HOLAP

20. Name the three kinds of data warehouse applications

Information processing, analytical processing, and data mining:

Unit- III

21. Define Unsupervised Learning

Unsupervised learning in which the class label of each training tuple is not known

22. Define Data cleaning.

Data Cleaning refers to the preprocessing of data in order to remove or reduce noise (by applying smoothing techniques), and the treatment of missing values.

23. Define Accuracy

The accuracy of a classifier refers to the ability of a given classifier to correctly predict the class label of new or previously unseendata

24. Define Gini index

The Gini index is used in CART. Using the notation described above, the Gini index measures the impurity of D, a data partition or set of training tuples, as

Gini(D)=1−m ∑ i=1 p2 i ,

25. Define ROC Curves

ROC curves are a useful visual tool for comparing two classification models. The name ROC stands for Receiver Operating Characteristic. ROC curves come from signal detection theory that was developed during World War II for the analysis of radar images.

26 Who developed ID3

J. Ross Quinlan, a researcher in machine learning, developed a decision tree algorithm known as ID3

27. Name any two algorithm which can handle categorical and continuous valued attributes.

SLIQ and SPRINT

28. Expand BOAT

BOAT (Bootstrapped Optimistic Algorithm for Tree Construction)

29. Name the two common approaches to tree pruning

PrePruning and Post Pruning

30. What are the components of belief network

A belief network is defined by two components—a *directed acyclic graph* and a set of *conditional probability tables*

Unit- IV

31. Write down the several basic clustering techniques?

Partitioning methods, hierarchical methods, density-based methods, and grid-based methods

32. What is Clustering?

Cluster analysis or simply clustering is the process of partitioning a set of data objects (or observations) into subsets.

33. What is automatic classification?

A cluster is a collection of data objects that are similar to one another within the cluster and dissimilar to objects in other clusters, a cluster of data objects can be treated as an implicit class. In this sense, clustering is sometimes called automatic classification.

34. Write the two methods for clustering in large databases?

Mixed numerical and nominal data.

35. Write any 2 or 3 requirements of clustering in data mining?

Scalability

Ability to deal with different types of attributes

Discovery of clusters with arbitrary shape

36. What are the 3 density-based representative methods?

Ans: DBSCAN, OPTICS and DENCLUE

37. Expand OPTICS?

Ordering Points to Identify the Clustering Structure

38. What is DENCLUE?

Density estimation is a core issue in density-based clustering methods. **DENCLUE** (DENsity-based CLUstEring) is a clustering method based on a set of density distribution functions.

39. What are the two grid-based clustering?

STING, CLIQUE

40. What is STING?

STING is a grid-based multiresolution clustering technique in which the embedding spatial area of the input objects is divided into rectangular cells.

Unit- V

41. Write any 3 trends in data mining ?

Application exploration

Scalable and interactive data mining methods

Mining social and information networks

42. What is human discover data of huge amount?

Visual and audio data mining is an effective way to integrate with humans' visual and audio systems and discover knowledge from huge amounts of data.

43. What is traditional data mining ?

Traditional data mining methods, designed to work at a centralized location, do not work well in many of the distributed computing environments present today.

44. What are the real time data to be generated?

Many applications involving stream data require dynamic data mining models to be built in real time.

45. Write any two cases of data mining for financial Analysis?

Design and construction of data warehouses for multidimensional data analysis and data mining.

Loan payment prediction and customer credit policy analysis.

46. Write any two cases of data mining in retail ?

Multidimensional analysis of sales, customers, products, time, and region,

Analysis of the effectiveness of sales campaigns.

47. Expand CRM?

Customer Relationship management.

48. Name the four basic building blocks of DNA

DNA sequences are comprised of four basic building blocks, called nucleotides: adenine (A), cytosine (C), guanine (G), and thymine (T).

49. List any two tools for the systematic analysis of genomic and proteomic data

BLAST and FASTA

50 Expand ODBC

Open Database Connectivity

PG DEPARTMENT OF COMPUTER APPLICATIONS

DATA MINING AND WAREHOUSING – 18PMC533

K3 LEVEL QUESTIONS

Unit I

1. Explain the major components of architecture of data mining system
2. Comment on transactional database
3. Differentiate heterogeneous data base and Legacy Database
4. Justify Data mining as a confluence of multiple discipline
5. How is a data warehouse different from a database? How are they similar?
6. List and describe the five primitives for specifying a data mining task.
7. Data quality can be assessed in terms of accuracy,completeness,and consistency. Propose two other dimensions of data quality.
8. Suppose that the values for a given set of data are grouped into intervals. The intervals and corresponding frequencies are as follows.

age	frequency
1–5	200
5–15	450
15–20	300
20–50	1500
50–80	700
80–110	44

   Compute an approximate median value for the data.
9. Discuss issues to consider during data integration.
10. Explain noisy data in detail

Unit II

11. Explain the concept of datawarehouse
12. How are organizations using the information from data warehouses?
13. Compare OLTP and OLAP systems.
14. Write the necessity of having a separate datawarehouse
15. How are measures computed
16. Sketch the design of a data warehouse
17. List and explain Data Warehouse Back-End Tools and Utilities

18. Discuss the types of OLAP servers
19. How do you convert OLAP to OLAM
20. Discuss the architecture for On-Line Analytical Mining

Unit III

21. How does classification work?
22. How is prediction different from classification?
23. How do you prepare the data from classification and prediction?
24. Compare classification and prediction method
25. Which attribute selection method is best? Explain
26. How does Tree Pruning works?
27. How does the Bayesian belief network learn?
28. Explain linear regression
29. Explain information gain with example
30. Explain IF_THEN Rules for classification

Unit IV

31. Show the typical requirement of clustering in data mining
32. Analyze the K means clustering algorithm
33. How efficient is the K- mediod algorithm on large dataset?
34. How do you improve the quality and scalability of CLARA?
35. Explain BARCH
36. Discuss ROCK
37. Explain sting
38. Compare OPTICS and DENCLUE
39. Differentiate ROCK and chameleon
40. How can we make K means algorithm more scalable?

Unit V

41. Apply data mining concept for financial data analysis
42. Apply data mining concept for retail industry
43. Apply data mining concept for telecom industry
44. Show the challenges of data mining for intrusion detection

45. Show the examples of commercial data mining system from database system and graphic system vendor
46. List out the commercial data mining system from the vendors of statistical analysis
47. Show the commercial data mining system that are originating from machine learning community
48. Brief on data mining in mining the complex data types
49. List out the aspects where data mining may contribute to biological data analysis
50. Discuss any 5 features to choose a data mining system

PG DEPARTMENT OF COMPUTER APPLICATIONS

DATA MINING AND WAREHOUSING – 18PMC533

K4/K5 QUESTIONS

Unit 1

1. Discuss the evolution of Database Technology
2. Elaborate on Data Mining Functionalities
3. Write in detail the classification of datamining system
4. Explain Descriptive Data Summarization
5. Explain Data Integration and Transformation

Unit 2

6. Explain Stars, Snowflakes, and Fact Constellations
7. Discuss OLAP Operations in the Multidimensional Data Model
8. Sketch the Three-Tier Data Warehouse Architecture
9. Transform Data Warehouse to Data Mining
10. Explain Starnet Query Model for Querying Multidimensional Database

Unit 3

11. Explain the basic algorithm fro inducing a decision tree from training tuples
12. Explain Gini index with example
13. Elaborate Naïve Bayesian Classification
14. Illustrate the rule induction using a sequential covering algorithm
15. Give an example for rule extraction from a decision tree

Unit 4

16. analyze the types of data in cluster analysis
17. Explain Agglomerative and Divisive Hierarchical clustering
18. Show and explain any one Density Based Methods
19. Explain the concept of Wave cluster
20. Illustrate the concept of DB Scan

Unit 5

21. Apply data mining for biological data analysis
22. How to choose a data mining system?
23. Explain ubiquitous and invisible data mining
24. Explain Trends in data mining
25. Elaborate on data mining privacy and data security.