INSTITUTE OF AERONAUTICAL ENGINEERING

(Autonomous)

Dundigal, Hyderabad - 500 043

Department of Computer Science and Engineering

TUTORIAL QUESTION BANK

Course Title	**DATA WAREHOUSING AND DATA MINING**
Course Code	**A70520 - JNTUH - R15**
Class	**IV B.Tech I Semester**
Branch	**CSE**
Year	**2018 – 2019**
Course Faculty	Dr. K Suvarchala, Professor, CSE Dr. M. Madhubala, Professor, CSE Ms. B. Padmaja, Associate Professor, CSE Mr. P Anjaiah , Assistant Professor, CSE

OBJECTIVES

To meet the challenge of ensuring excellence in engineering education, the issue of quality needs to be addressed, debated and taken forward in a systematic manner. Accreditation is the principal means of quality assurance in higher education. The major emphasis of accreditation process is to measure the outcomes of the program that is being accredited.

In line with this, faculty of Institute of Aeronautical Engineering, Hyderabad has taken a lead in incorporating philosophy of outcome based education in the process of problem solving and career development. So, all students of the institute should understand the depth and approach of course to be taught through this question bank, which will enhance learner's learning process

S. No	QUESTION	Blooms Taxonomy Level	Course Outcome
UNIT – I DATA WAREHOUSE			
Part - A (Short Answer Questions)			
1	Define online analytical processing?	Remember	2
2	List the key features of data warehouse?	Understand	2
3	Define data mart?	Remember	1
4	Define enterprise warehouse?	Remember	1
5	What is virtual warehouse?	Remember	1
6	List the metadata repository?	Understand	1
7	List the various multidimensional models?	Understand	1
8	Explain about the star schema?	Understand	2
9	Explain the snowflake schema?	Understand	2
10	Define about the fact constellation model?	Remember	2
11	Name the OLAP operations?	Understand	2
12	Express what is slice and dice operation?	Understand	2
13	Define Pivot operation?	Remember	2
14	Distinguish between the OLAP Systems and Statistical databases?	Understand	2
15	State the various views of data warehouse design?	Understand	2

16	Define the Relational OLAP (ROLAP) server?	Remember	1
17	Explain about Multidimensional OLAP (MOLAP) server?	Understand	1
18	Discuss about Hybrid OLAP (HOLAP) server?	Understand	1
19	Define Data warehouse?	Remember	1
20	Define the use of concept hierarchy?	Remember	1
Part - B (Long Answer Questions)			
1	Differentiate operational database systems and data warehousing?	Understand	1
2	Discuss briefly about the multidimensional data models?	Understand	1
3	Explain with an example the different schemas for multidimensional databases?	Understand	1
4	Describe the three-tier data warehousing architecture?		1
5	Discuss the efficient processing of OLAP queries?	Remember	2
6	Explain the data warehouse applications?	Understand	2
7	Explain the architecture for on-line analytical mining?	Understand	2
8	Describe the common techniques are used in ROLAP and MOLAP?	Understand	1
9	Describe the complex aggregation at multiple granularities?	Remember	1
10	Explain about the concept description? And what are the differences between concept description in large databases and OLAP?	Remember	2
11	Discuss about Metadata Repository?	Understand	2
12	Compare the schemas for the multidimensional data models?		2
13	Explain about the data warehouse implementation with an example?	Understand	1
14	Discuss about types of OLAP Servers?	Understand	1
15	Explain OLAP operations in the Multidimensional Data Model?	Understand	2
16	Compare Enterprise warehouse, data mart and virtual warehouse?	Understand	1
17	Compare Data cleaning, data transformation?	Understand	1
18	Explain the differences between the three main types of data warehouse usage: information processing, analytical processing and data mining? Discuss the motivation behind OLAP mining (OLAM)?	Understand	1
19	Explain a data warehouse can be modeled by either a star schema or a Snowflake schema. Briefly describe the similarities and the differences of the two models, and then analyze their advantages and disadvantages with regard to one another?	Understand	2
20	Explain Indexing OLAP Data?	Understand	1
Part – C (Critical Thinking and Analytical Questions)1			
1	Analyze that a data warehouse consists of the three dimensions time, doctor and patient, and the two measures count and charge, where charge is the fee that a doctor charges a patient for a visit. i. Enumerate three classes of schemas classes of schemas that are popularly used for modeling data warehouses. ii. Draw a schema diagram for the above data warehouse using one of the schema classes listed in (a). iii. Starting with the base cuboids [day, doctor, patient], what specific OLAP operations should be performed in order to list the total fee collected by each doctor in 2004? iv. To obtain the same list, write an SQL query assuming the data is stored in a relational database with the schema fee (day, month, year, doctor, hospital, patient, count, charge).	Remember	1
2	State why, for the integration of multiple heterogeneous information Sources, many companies in industry prefer the update-driven approach (which constructs and uses data warehouses), rather than the query-driven approach (which applies wrappers and integrators). Describe situations where the query-driven approach is preferable over the update-driven approach.	Understand	2

3	Suppose that a data warehouse for Big University consists of the following four dimensions: student, course, semester, and instructor, and two measures count and average g grade. When at the lowest conceptual level (e.g., for a given student, course, semester, and instructor combination), the average grade measure stores the actual course grade of the student. At higher combination. Compute the number of cuboids i. Draw a snowflake schema diagram for the data warehouse. ii. Starting with the base cuboids [student, course, semester, instructor], what specific OLAP operations (e.g., roll-up from semester to year) should one perform in order to list the average grade of CS courses for each Big University student. iii. If each dimension has five levels (including all), such as “student < major < status < university < all”, how many cuboids will this cube contain (including the base and apex cuboids)?	Understand	1
4	Suppose that a data warehouse consists of the four dimensions, date, spectator location, and game, and the two measures, count and charge, where charge is the fare that a spectator pays when watching a game on a given date. Spectators may be students, adults, or seniors, with each category having its own charge rate. Write the following i. Draw a star schema diagram for the data warehouse. ii. Starting with the base cuboids [date, spectator, location, game], what specific OLAP operations should one perform in order to list the total charge paid by student spectators at GM Place in 2004? iii. Bitmap indexing is useful in data warehousing. Taking this cube as an example, briefly discuss advantages and problems of using a bitmap index structure.	Understand	2
5	Design a data warehouse for a regional weather bureau. The weather bureau has about 1,000 probes, which are scattered throughout various land and ocean locations in the region to collect basic weather data, including air pressure, temperature, and precipitation at each hour. All data are sent to the central station, which has collected such data for over 10 years. Your design should facilitate efficient querying and on-line analytical processing, and derive general weather patterns in multidimensional space.	Remember	1
6	Explain the computation of measures in a data cube: i. Enumerate three categories of measures, based on the kind of aggregate functions used in computing a data cube. ii. For a data cube with the three dimensions time, location, and item, which category does the function variance belong to? Describe how to compute it if the cube is partitioned into many chunks. $\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x}_i)^2$ Hint: The formula for computing variance is where x_i is the average of N x_i’s. iii. Suppose the function is “top 10 sales”. Discuss how to efficiently compute this measure in a data cube.	Remember	1,2
7	Suppose that we need to record three measures in a data cube: min, average, and median. Design an efficient computation and storage method for each measure given that the cube allows data to be deleted incrementally (i.e., in small portions at a time) from the cube.	Understand	
8	Observe that a data warehouse contains 20 dimensions, each with about five levels of granularity. i. Users are mainly interested in four particular dimensions, each having ii. Three frequently accessed levels for rolling up and drilling down. How would you design a data cube structure to efficiently support this preference? iii. At times, a user may want to drill through the cube, down to the raw data for one or two particular dimensions. How would you support this feature?	Understand	1,2

9	Observe A data cube, C, has n dimensions, and each dimension has exactly p distinct values in the base cuboid. Assume that there are no concept hierarchies associated with the dimensions. i. What is the maximum number of cells possible in the base cuboid? ii. What is the minimum number of cells possible in the base cuboid? iii. What is the maximum number of cells possible (including both base cells and aggregate cells) in the data cube, C? iv. What is the minimum number of cells possible in the data cube, C?	Understand	1,2
10	Observe A popular data warehouse implementation is to construct a multidimensional database, known as a data cube. Unfortunately, this may often generate a huge, yet very sparse multidimensional matrix. Present an example illustrating such a huge and sparse data cube.	Understand	1

UNIT – II
INTRODUCTION TO DATA MINING

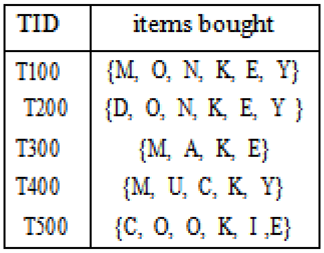
Part - A (Short Answer Questions)

1	Define data mining?	Remember	3
2	Explain the definition of data warehouse?	Understand	4
3	Distinguish between data mining and data warehouse?	Understand	3
4	Identify any three functionality of data mining?	Remember	3
5	Interpret major issues in data mining?	Understand	3
6	Name the steps in the process of Remember discovery?	Remember	3
7	Discuss relational databases?	Understand	4
8	State object –oriented Databases?	Understand	3
9	Explain the spatial databases?	Understand	4
10	Contrast heterogeneous databases and legacy databases?	Understand	5
11	Differentiate classification and Prediction?	Understand	3
12	Describe transactional data bases?	Remember	4
13	List the types of data that can be mined?	Remember	5
14	Define data cube?	Remember	3
15	Define multidimensional data mining?	Remember	4
16	Define data characterization?	Remember	5
17	Express what is a decision tree?	Understand	4
18	Explain the outlier analysis?	Understand	3
19	Name the steps involved in data preprocessing?	Understand	4
20	Interpret the dimensionality reduction?	Understand	4
1	Describe data mining? In your answer, address the following: i. Is it another hype? ii. Is it a simple transformation of Technology developed from databases, statistics, and machine learning? iii. Explain how the evolutions of database technology lead to data mining? iv. Describe the steps involved in data mining when viewed as a process of remember discovery.	Remember	5
2	Distinguish between the data warehouse and databases? How they are similar?	Remember	5
3	Explain the difference between discrimination and classification? Between characterization and clustering? Between classification and prediction? For each of these pairs of tasks, how are they similar?	Remember	5
4	Describe three challenges to data mining regarding data mining methodology and user interaction issues?	Remember	4
5	Distinguish between the data warehouses and data mining?	Understand	3
6	Discuss briefly about the data smoothing techniques?	Understand	3

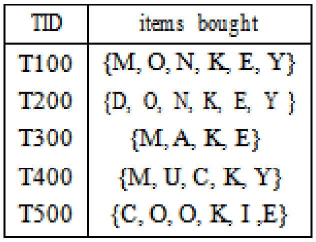
7	Explain Data Integration and Transformation?	Understand	4
8	Describe the various data reduction techniques?	Understand	4
9	Define data cleaning? Express the different techniques for handling missing values?	Remember	4
10	Differentiate between descriptive and predictive data mining?	Understand	5
11	Explain data mining as a step in the process of Remember discovery?	Understand	3
12	Describe briefly Discretization and concept hierarchy generation for numerical data?	Understand	5
13	Discuss about the concept hierarchy generation for categorical data?	Understand	3
14	List and describe the five primitives for specifying a data mining task?	Understand	4
15	Discuss issues to consider during data integration?	Remember	3
16	Describe the following advanced database systems and applications: object- relational databases, spatial databases, text databases, multimedia databases, stream data, the World Wide Web.	Understand	4
17	Describe why concept hierarchies are useful in data mining.	Remember	3
18	Describe the differences between the following approaches for the integration of a data mining system with a database or data warehouse system: no coupling, loose coupling, semi-tight coupling, and tight coupling. State which approach you think is the most popular, and why?	Remember	4
19	Explain Data quality can be assessed in terms of accuracy, completeness, and consistency. Propose two other dimensions of data quality.	Remember	5
20	Apply the two methods below to normalize the following group of data: 200, 300, 400, 600, 1000 i. min-max normalization by setting min = 0 and max = 1 ii. z-score normalization	Remember	5
Part – C (Critical Thinking and Analytical Questions)			
1	Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25,25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52,70. Compute the following: i. Mean of the data? Median? mode of the data? Comment on the data's modality ii. (i.e., bimodal, trimodal) iii. Midrange of the data?	Understand	5
2	Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19,20,20,21,22,22,25,2525, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52,70. Compute the following: i. Can you find (roughly) the first quartile (Q1) and the third quartile Q3 of the data? ii. Give the five-number summary of the data. iii. Show a box plot of the data. iv. How is a quantile-quantile plot different from a quantile plot?	Remember	4
3	Use the data for age given above answer the following. i. Use smoothing by bin means to smooth the above data, using a bin depth of 3. Illustrate your steps. Comment on the effect of this technique for the given data ii. How might you determine outliers in the data? iii. What other methods are there for data smoothing?	Remember	5
4	Suppose a hospital tested the age and body fat data for 18 randomly selected adults with the following result. Age 23 23 27 27 39 41 47 49 50 %fat 9.5 26.5 7.8 17.8 31.4 25.9 27.4 27.2 31.2 Age 52 54 54 56 57 58 58 60 61 %fat 34.6 42.5 28.8 33.4 30.2 34.1 32.9 41.2 35.7 Examine the following: i. The mean, median and standard deviation of age and %fat. ii. Draw the box plots for age and %fat. iii. Draw a scatter plot and a q-q plot based on these two variables.	Remember	4

5	Write an example where data mining is crucial to the success of a business. What data mining functions does this business need? Can they be performed alternatively by data query processing or simple statistical analysis?	Remember	4
6	Suppose your task as a software engineer at Big University is to design a data mining system to examine the university course database, which contains the following information: the name, address, and status (e.g., undergraduate or graduate) of each student, the courses taken, and the cumulative grade point average (GPA). Describe the architecture you would choose. What is the purpose of each component of this architecture?	Understand	5
7	Examine the following consider the following data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. i. Use min-max normalization to transform the value of 35 for age on to the range [0.0, 1.0]. ii. Use z-score normalization to transform the value 35 for age, where the standard deviation of age is 12.94 years. iii. Use normalization by decimal scaling to transform the value 35 for age.	Remember	4
8	Suppose a group of 12 sales price records has been sorted as follows: 5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215 Examine the following methods by partition them into three bins i. equal-frequency (equidepth) partitioning ii. equal-width partitioning iii. clustering	Remember	4
UNIT – III **ASSOCIATION RULES**			
Part - A (Short Answer Questions)			
1	Define frequent patterns?	Understand	5
2	Define closed itemset?	Remember	6
3	State maximal frequent itemset?	Remember	5
4	List the techniques of efficiency of Apriori algorithm?	Remember	6
5	Explain ECLAT algorithms usage?	Remember	5
6	Name the pruning strategies in mining closed frequent itemsets?	Remember	6
7	Define substructure of a structural pattern?	Understand	5
8	Interpret the rule of support for itemsets A and B?	Remember	6
9	Classify the confidence rule for itemsets A and B?	Remember	5
10	Define itemset?	Remember	5
11	Name the steps in association rule mining?	Remember	5
12	Explain the join step?	Remember	5
13	Describe the prune step?	Understand	5
14	State how can we mine closed frequent itemsets?	Remember	6
15	Name the pruning strategies of closed frequent itemsets?	Remember	6
16	Explain the two kinds of closure checking?	Understand	6
17	Summarize the constraint-based mining?	Understand	6
18	Describe the five categories of pattern mining constraints?	Remember	6
19	List the applications of pattern mining?	Remember	5
20	Define Support and Confidence?	Remember	5
Part - B (Long Answer Questions)			
1	Define the terms frequent itemsets, closed itemsets and association rules?	Understand	5
2	Discuss which algorithm is an influential algorithm for mining frequent itemsets for boolean association rules? Explain with an example?	Remember	5

3	Describe the different techniques to improve the efficiency of Apriori? Explain?	Understand	5
4	Discuss the FP-growth algorithm? Explain with an example?	Remember	6
5	Explain how to mine the frequent itemsets using vertical data format?	Remember	5
6	Discuss about mining multilevel association rules from transaction databases in detail?	Understand	6
7	Explain how to mine the multidimensional association rules from relational databases and data warehouses?	Understand	6
8	Describe briefly about the different correlation measures in association analysis?	Remember	6
9	Discuss about constraint-based association mining?	Understand	6
10	Explain the Apriori algorithm with example?	Remember	5
11	Discuss the generating association rules from frequent itemsets?	Remember	5
12	Discuss about mining multilevel association rules from transaction databases in detail?	Understand	5
13	Describe multidimensional association rules using static Discretization?	Remember	6
14	Explain what are additional rule constraints to guide mining?	Understand	6
15	Explain, how can we tell which strong association rules are really interesting? Explain with an example?	Remember	6
16	Describe about the correlation analysis using Chi-square?	Understand	5

17	Apply the following rules on a database has five transactions. Let min sup =60% and min conf = 80%.
	
	TID | items bought
	T100 | {M, O, N, K, E, Y}
	T200 | {D, O, N, K, E, Y }
	T300 | {M, A, K, E}
	T400 | {M, U, C, K, Y}
	T500 | {C, O, O, K, I ,E}
	i. Find all frequent itemsets using Apriori .
	ii. List all of the strong association rules (with support s and confidence matching the following metarule, where X is a variable representing customers, and item i denotes variables representing items (e.g., “A”, “B”, etc.): ∀x ∈ transaction, buys(X , item1) ∧ buys(X , item2) ⇒ buys(X , item3) [s, c]
		Remember	6

18	Describe about the Mining closed frequent itemset.	Understand	5
19	Write a short example to show that items in a strong association rule may actually be negatively correlated.	Remember	6
20	Explain Association rule mining often generates a large number of rules. Discuss effective methods that can be used to reduce the number of rules generated while still preserving most of the interesting rules.	Understand	5
	Part – C (Critical Thinking and Analytical Questions)		
1	The Apriori algorithm uses prior knowledge of subset support properties. Analyze i. That all nonempty subsets of a frequent itemset must also be frequent. ii. The support of any nonempty subset s 0 of itemset s must be at least as great as the support of s. iii. Given frequent itemset l and subset s of l, prove that the confidence of the rule “s 0 ⇒ (l − s 0)” cannot be more than the confidence of “s ⇒ (l –s)”, where s 0 is a subset of s. iv. A partitioning variation of Apriori subdivides the transactions of a database D into n nonoverlapping partitions. Prove that any itemset that is frequent in D must be frequent in at least one partition of D.	Understand	5

<table>
<tr><td>2</td><td>Implement three frequent itemset mining algorithms introduced in this chapter : Apriori [AS94], (2) FP-growth [HPY00], and (3) ECLAT [Zak00] (mining using vertical data format), using a programming language that you are familiar with, such as C++ or Java. Compare the performance of each algorithm with various kinds of large data set. Write a report to analyze the situations (such as data size, data distribution, minimal support threshold setting, and pattern density) where one algorithm may perform better than the others, and state why.</td><td>Remember</td><td>5</td></tr>
<tr><td>3</td><td>Suppose that a large store has a transaction database that is distributed among four locations. Transactions in each component database have the same format, namely Tj : {i1, . . . , im}, where Tj is a transaction identifier, and ik (1 ≤ k ≤m) is the identifier of an item purchased in the transaction. Construct an efficient algorithm to mine global association rules (without considering multilevel associations). You may present your algorithm in the form of an outline. Your algorithm should not require shipping all of the data to one site and should not cause excessive network communication overhead.</td><td>Understand</td><td>6</td></tr>
<tr><td>4</td><td>Suppose that frequent itemsets are saved for a large transaction database, DB. Illustrate how to efficiently mine the (global) association rules under the same minimum support threshold if a set of new transactions, denoted as ΔDB, is incrementally) added in?</td><td>Remember</td><td>6</td></tr>
<tr><td>5</td><td>Most frequent pattern mining algorithms consider only distinct items in a transaction However, multiple occurrences of an item in the same shopping basket, such as four cakes and three jugs of milk, can be important in transaction data analysis. Analyze how can one mine frequent itemsets efficiently considering multiple occurrences of items? Propose modifications to the well-known algorithms, such as Apriori and FP-growth, to adapt to such a situation.</td><td>Remember</td><td>6</td></tr>
<tr><td></td><td></td><td></td><td></td></tr>
<tr><td>6</td><td>A database has five transactions. Let min sup = 60% and min con f = 80%.

<table>
<tr><th>TID</th><th>items bought</th></tr>
<tr><td>T100</td><td>{M, O, N, K, E, Y}</td></tr>
<tr><td>T200</td><td>{D, O, N, K, E, Y }</td></tr>
<tr><td>T300</td><td>{M, A, K, E}</td></tr>
<tr><td>T400</td><td>{M, U, C, K, Y}</td></tr>
<tr><td>T500</td><td>{C, O, O, K, I ,E}</td></tr>
</table>
Examine the following

i. Find all frequent itemsets using FP-growth.
ii. List all of the strong association rules (with support s and confidence c) matching the following metarule, where X is a variable representing customers, and itemi denotes variables representing items (e.g., “A”, “B”, etc.): ∀x ∈ transaction, buys(X , item1) ∧ buys(X , item2) ⇒ buys(X , item3) [s,c].</td><td>Understand</td><td>5</td></tr>
<tr><td>7</td><td>The following contingency table summarizes supermarket transaction data, where hot dogs refers to the transactions containing hot dogs, hot dogs refers to the transactions that do not contain hot dogs, hamburgers refers to the transactions containing hamburgers, and hamburgers refers to the transactions that do not contain hamburgers.
<table>
<tr><th></th><th>Hot dogs</th><th>Hot dogs</th><th></th></tr>
<tr><td>Hamburgers</td><td>2000</td><td>500</td><td>2500</td></tr>
<tr><td>Hamburgers</td><td>1000</td><td>1500</td><td>2500</td></tr>
<tr><td></td><td>3000</td><td>2000</td><td>5000</td></tr>
</table>
Observe that the association rule “hot dogs ⇒ hamburgers” is mined. Given a minimum support threshold of 25% and a minimum confidence threshold of 50%, is this association rule strong? Based on the given data, is the purchase of hot dogs independent of the purchase of hamburgers? If not, what kind of correlation relationship exists between the two?</td><td>Understand</td><td>5</td></tr>
</table>

8	Sequential patterns can be mined in methods similar to the mining of association rules. Design an efficient algorithm to mine multilevel sequential patterns from a transaction database. An example of such a pattern is the following: “A customer who buys a PC will buy Microsoft software within three months,” on which one may drill down to find a more refined version of the pattern, such as “A customer who buys a Pentium PC will buy Microsoft Office within three months.”	Remember	5
9	The price of each item in a store is nonnegative. The store manager is only interested in rules of the form: “one free item may trigger $200 total purchases in the same transaction.” Describe how to mine such rules efficiently.	Remember	6
10	The price of each item in a store is nonnegative. For each of the following cases, identify the kinds of constraint they represent and briefly discuss how to mine such association rules efficiently. i. Containing at least one Nintendo game. ii. Containing items the sum of whose prices is less than $150. iii. Containing one free item and other items the sum of whose prices is at least $200 where the average price of all the items is between $100 and $500.	Remember	6

UNIT – IV
CLASSIFICATION

Part - A (Short Answer Questions)

1	State classification?	Understand	7
2	Define regression analysis?	Remember	7
3	Name the steps in data classification?	Understand	7
4	Define training tuple?	Remember	7
5	Describe accuracy of a classifier?	Remember	8
6	Differentiate supervised learning and unsupervised learning?	Understand	8
7	Define the decision tree?	Understand	8
8	Define information gain?	Remember	8
9	State gain ratio?	Remember	7
10	State Gini index?	Remember	8
11	Explain tree pruning?	Understand	7
12	Define the construction of naïve Bayesian classification?	Remember	8
13	Explain the IF-THEN rules for classification?	Understand	7
14	Explain Decision Tree Induction?	Remember	8
15	List the Attribute Selection Measures?	Remember	7
16	Define Bayes’ Theorem?	Understand	8
17	Define Naïve Bayesian Classification?	Understand	7
18	Explain K-Nearest-Neighbor Classifiers?	Remember	8

Part - B (Long Answer Questions)

1	Explain about the classification and prediction? Example with an example?	Understand	8
2	Discuss about basic decision tree induction algorithm?	Remember	8
3	Explain briefly various measures associated with attribute selection?	Understand	7,8
4	Summarize how does tree pruning work? What are some enhancements to basic decision tree induction?	Remember	7,8
5	Explain how scalable is decision tree induction? Explain?	Remember	8
6	Describe the working procedures of simple Bayesian classifier?		8
7	Explain Bayesian Belief Networks?		8
8	Discuss about k-nearest neighbor classifier and case-based reasoning?		7
9	Explain about classifier accuracy? Explain the process of measuring the accuracy of a classifier?	Remember	7
10	Describe any ideas can be applied to any association rule mining be applied to classification?	Remember	7
11	Explain briefly about the Naïve Bayesian Classification?	Understand	7

12	Explain about the major issues regarding classifications and predictions?	Remember	7
13	Differentiate classification and prediction methods?	Understand	8
14	Explain briefly various measures associated with attribute selection?	Remember	8
15	Explain how tree pruning useful in decision tree induction? What is the drawback of using a separate set of tuples to evaluate pruning?	Remember	8
16	Explain for a given a decision tree, you have the option of (a) converting the decision tree to rules and then pruning the resulting rules, or (b) pruning the decision tree and then converting the pruned tree to rules. What advantage does (a) have over (b)?	Remember	7
17	Compare the advantages and disadvantages of eager classification (e.g.decision tree, Bayesian, neural network) versus lazy classification (e.g., k-nearest neighbor, case-based reasoning).	Understand	7
18	Write an algorithm for k-nearest-neighbor classification given k and n, the number of attributes describing each tuple.	Remember	7
19	Describe each of the following clustering algorithms in terms of the following criteria: (i) shapes of clusters that can be determined; (ii) input parameters that must be specified and (iii) limitations. i. k-means ii. k-medoids	Remember	7
20	Explain training of Bayesian belief networks?		7
Part – C (Analytical Questions)			
1	Illustrate why is tree pruning useful in decision tree induction? Explain the drawback of using a separate set of tuples to evaluate pruning?	Understand	7,8
2	Given a decision tree, you have the option of (a) converting the decision tree to rules and then pruning the resulting rules, or (b) pruning the decision tree and then converting the pruned tree to rules. Explain advantage does (a) have over (b)?	Remember	7,8
3	Outline the major ideas of naive Bayesian classification. Explain why is naïve Bayesian classification called “naive”?	Understand	7,8
4	Design an efficient method that performs effective Naive Bayesian classification over an infinite data stream (i.e. you can scan the data stream only once). If we wanted to discover the evolution of such classification schemes (e.g., comparing the classification scheme at this moment with earlier schemes, such as one from a week ago), construct modified design would you suggest?	Remember	7,8
5	The support vector machine (SVM) is a highly accurate classification method. However, SVM classifiers suffer from slow processing when training with a large set of data tuples. Explain how to overcome this difficulty and develop a scalable SVM algorithm for efficient SVM classification in large datasets.	Remember	7,8
6	Given a 5 GB data set with 50 attributes (each containing 100 distinct values) and 512 MB of main memory in your laptop, outline an efficient method that constructs decision trees in such large data sets. Justify your answer by rough calculation of your main memory usage.	Remember	7,8
7	What is associative classification? Why is associative classification able to achieve higher classification accuracy than a classical decision tree method? Explain how associative classification can be used for text document classification.	Understand	7,8
8	It is difficult to assess classification accuracy when individual data objects may belong to more than one class at a time. In such cases, Explain on what criteria you would use to compare different classifiers modeled after the same data.	Remember	8
9	Describe each of the following clustering algorithms in terms of the following criteria: (i) shapes of clusters that can be determined; (ii) input parameters that must be specified; and (iii) limitations. i. k-means ii. k-medoids iii. CLARA iv. BIRCH v. ROCK vi. Chameleon vii. DBSCAN	Understand	8

UNIT – V
CLUSTERING

Part - A (Short Answer Questions)

1	Define Clustering?	Remember	9
2	Illustrate the meaning of cluster analysis?	Remember	9
3	Explain the fields in which clustering techniques are used?	Remember	9
4	List out the requirements of cluster analysis?	Remember	10
5	Express the different types of data used for cluster analysis?	Remember	10
6	State interval scaled variables?	Remember	10
7	Define Binary variables? And what are the two types of binary variables?	Remember	9
8	Define nominal, ordinal and ratio scaled variables?	Remember	10
9	Illustrate mean by partitioning method?	Understand	10
10	Define CLARA and CLARANS?	Understand	10
11	State hierarchical method?	Understand	10
12	Differentiate agglomerative and divisive hierarchical clustering?	Remember	10
13	State K-Means method?	Understand	10
14	Define Outlier Detection?	Remember	10
15	Define Chameleon method?	Remember	10

Part - B (Long Answer Questions)

1	Discuss the various types of data in cluster analysis?	Understand	9
2	Explain the categories of major clustering methods?	Remember	9
3	Write algorithms for k-means and k-medoids? Explain?	Remember	9
4	Describe the different types of hierarchical methods?	Remember	9
5	Demonstrate about the following hierarchical methods i. BIRCH ii. Chamelon	Remember	9
6	Explain about semi-supervised cluster analysis?	Remember	9
7	Explain about the outlier analysis?	Remember	9
8	Define the distance-based outlier? Illustrate the efficient algorithms for mining distance based algorithm?	Remember	9
9	Explain about the Statistical-based outlier detection?	Remember	9
10	Describe about the distance-based outlier detection?	Understand	9
11	Discuss about the density-based outlier detection?	Understand	9
12	Demonstrate about the BIRCH hierarchical methods?	Remember	10
13	Demonstrate about the ROCK(Robust Clustering using links) hierarchical methods?	Remember	10
14	Explain about the agglomerative and divisive hierarchical methods?	Remember	10
15	Demonstrate how to compute the dissimilarity between objects described by the following types of variables: i. Numerical (interval-scaled) variables ii. Asymmetric binary variables iii. Categorical variables iv. Ratio-scaled variables v. Non-metric vector objects	Remember	10
16	Apply the following measurements for the variable age:18, 22, 25, 42, 28, 43, 33, 35, 56, 28 standardize the variable by the following: i. Compute the mean absolute deviation of age. ii. Compute the z-score for the first four measurements.	Remember	10
17	Illustrate the strength and weakness of k-means in comparison with the k-medoids algorithm. Also, illustrate the strength and weakness of these schemes in comparison with a hierarchical clustering scheme (such as AGNES).	Remember	10

18	Explain why is outlier mining important? Briefly describe the different approaches behind statistical-based outlier detection, distanced-based outlier detection, density-based local out- lier detection, and deviation-based outlier detection.	Remember	10
19	Apply the given following measurements for the variable age:28, 32, 15, 42, 28, 43, 30, 32, 55, 26. Standardize the variable by the following: i. Compute the mean absolute deviation of age. ii. Compute the z-score for the first four measurements.	Remember	10
20	Discuss indetail about deviation-based outlier detection techniques?	Remember	10

Part – C (Analytical Questions)

1	Given the following measurements for the variable age: 48, 12, 25, 42,28,43,33,35,56, 28, standardize the variable by the following: Compute i. The mean absolute deviation of age. ii. The z-score for the first four measurements.	Understand	10
2	Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36,8): Compute i. The Euclidean distance between the two objects. ii. The Manhattan distance between the two objects. iii. The Minkowski distance between the two objects, using p = 3.	Remember	10
3	Suppose that the data mining task is to cluster the following eight points (with (x,y) representing location) into three clusters. A1(2, 10), A2(2, 5), A3(8, 4), B1(5, 8),B2(7, 5), B3(6, 4), C1(1, 2), C2(4, 9). The distance function is Euclidean distance. Suppose initially we assign A1, B1, and C1 as the center of each cluster respectively. Use the k-means algorithm to show only i. The three cluster centers after the first round of execution and ii. The final three clusters	Remember	10
4	Explain why is it that BIRCH encounters difficulties in finding clusters of arbitrary shape but OPTICS does not? Can you propose some modifications to BIRCH to help it find clusters of arbitrary shape?	Remember	10
5	Clustering has been popularly recognized as an important data mining task with broad applications. Show one application example for each of the following cases: i. An application that takes clustering as a major data mining function ii. An application that takes clustering as a preprocessing tool for data preparation for other data mining tasks	Remember	10
6	Clustering has been popularly recognized as an important data mining task with broad applications. Give example for each of the following cases: i. An application that takes clustering as a major data mining function. ii. An application that takes clustering as a preprocessing tool for data preparation for other data mining tasks	Understand	10
7	Data cubes and multidimensional databases contain categorical, ordinal, and numerical data in hierarchical or aggregate forms. Based on what you have learned about the clustering methods, Design a clustering method that finds clusters in large data cubes effectively and efficiently.	Remember	10
8	Human eyes are fast and effective at judging the quality of clustering methods for two-dimensional data. Design data visualization methods that may help humans visualize data clusters and judge the clustering quality for three-dimensional data? What about for even higher-dimensional data?	Remember	10
9	Given the following measurements for the variable age: 29, 31, 25, 41,27,43,33,35 56, 28, standardize the variable by the following: Compute i. The mean absolute deviation of age. ii. The z-score for the first three measurements.	Remember	10
10	Given two objects represented by the tuples (21, 2, 41, 11) and (21, 1, 32,6): Compute i. The Euclidean distance between the two objects. ii. The Manhattan distance between the two objects. iii. The Minkowski distance between the two objects, using p = 2.	Remember	10

Prepared By: Ms. K Suvarchala, Professor, CSE **HOD, CSE**