Reg. No. : | | | | | | | | | | | | |

Question Paper Code : J1512

M.Sc. DEGREE EXAMINATION, FEBRUARY/MARCH 2018.

Third Semester

Computer Science

DCS 7302 — DATA WAREHOUSING AND MINING

(Regulations 2013)

Time : Three hours Maximum : 100 marks

Answer ALL questions.

PART A — (10 × 2 = 20 marks)

1. Write down the difference between operational database systems and data warehouses.
2. Define OLAP.
3. What is the need of data preprocessing?
4. What is meta data?
5. List out the two measures of an association rule.
6. What is correlation analysis?
7. Write down the difference between Lazy learners and Eager learners.
8. How the prediction is differed from classification?
9. Define the term outlier analysis.
10. What is Wave cluster?

PART B — (5 × 13 = 65 marks)

11. (a) Discuss about multidimensional database, data mart and data cube? Explain schemas for multi-dimensional database. (13)

Or

(b) (i) Define data warehouse with five features. With neat sketch explain the architecture of data warehouse with suitable block diagram. (7)

(ii) Explain the various types of OLAP in detail. (6)

12. (a) Explain with diagrammatic illustration about data mining as a step in the process of knowledge discovery. (13)

Or

(b) List and discuss the steps for integrating a data mining system with a data. (13)

13. (a) What is called constrain-based mining? Discuss the types of constraints. Explain Meta rule-Guided Mining of Association Rules and Constraint-Based Pattern Generation: Pruning Pattern Space and Pruning Data Space. (13)

Or

(b) Write and explain the algorithm for mining frequent item sets without candidate generation. Consider given database with nine transactions let min_sup = 30%. Find all frequent item sets. (13)

TID	List of items_IDs
1	a,b,e
2	b,d
3	b,c
4	a,b,d
5	a,c
6	b,c
7	a,c
8	a,b,c,e
9	a,b,c

14. (a) Describe various attribution selection measures in classification with examples. (13)

Or

(b) The following table consists of training data from an employee database. The data have been generalized. For example, "31 .... 35" for age represents the age range of 31 to 35. For a given row entry, count represents the number of data tuples having the values for department, status, age, and salary given in that row. (13)

Department	Status	Age	Salary	Count
Sales	Senior	31.....35	46K.....50K	30
Sales	Junior	62.....30	26K.....30K	40
Sales	Junior	31.....35	31K.....35K	40
Systems	Junior	21.....25	46K.....50K	20

Systems	Senior	31.....35	66K.....70K	5
Systems	Junior	26.....30	46K.....50K	3
Systems	Senior	41.....45	66K.....70K	3
Marketing	Senior	36.....40	46K.....50K	10
Marketing	Junior	31.....35	41K.....45K	4
Secretary	Senior	46.....50	36K.....40K	4
Secretary	Junior	26.....30	26K.....30K	6

Let status be the class label attribute.

(i) How would you modify the basic decision tree algorithm to take into consideration the count of each generalized data tuple (i.e., of each row entry)?

(ii) Use your algorithm to construct a decision tree from the given data.

(iii) Given a data tuple having the values "systems", "26...30" and "46–50K" for the attributes department, age, and salary, respectively, what would a naive Bayesian classification of the status for the tuple be? (13)

15. (a) Consider five points {X_1, X_2, X_3, X_4, X_5} with the following coordinates as a two dimensional sample for clustering. X_1 = (0.5, 2.5); X_2 = (0, 0); X_3 = (1.5, 1); X_4 = (5, 1); X_5 = (6, 2). Illustrate the K-means partitioning algorithms using the above data set. (13)

Or

(b) (i) Write the difference between CLARA and CLARANS. (7)

(ii) Explain the different types of data used in cluster analysis. (6)

PART C — (1 × 15 = 15 marks)

16. (a) Explain K-means algorithm with an example.

Or

(b) How you evaluate the accuracy of a classifier (or) predictor?

———————

Reg. No. : | | | | | | | | | | | | |

Question Paper Code : BS2512

M.Sc. DEGREE EXAMINATION, AUGUST/SEPTEMBER 2017.

Third Semester

Computer Science

DCS 7302 — DATA WAREHOUSING AND MINING

(Regulations 2013)

Time : Three hours Maximum : 100 marks

Answer ALL questions.

PART A — (10 × 2 = 20 marks)

1. What is data mining?
2. List two applications of data mining.
3. Differentiate database from a data warehouse.
4. Give an example of where OLTP is suitable and OLAP is not.
5. Explain the purpose of data cleaning.
6. What is Concept Hierarchy?
7. List two advantages of classification using a decision tree.
8. What is Support Vector Machine (SVM)?
9. What is clustering?
10. How does identification of outliers aid in mining?

PART B — (5 × 13 = 65 marks)

11. (a) Suppose a data warehouse consists of three dimensions time, doctor and patient and 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 that are popularly used for modelling data warehouses.

    (ii) Draw a schema diagram for above data warehouse using schema classes listed in (i). (13)

Or

(b) (i) Explain the architecture of data warehouse. (6)

(ii) List differences between OLAP and OLTP. (3)

(iii) Explain any two OLAP operations with examples. (4)

12. (a) (i) List three major issues in mining data. Explain. (6)

(ii) Explain the method used to find proximity measures for attributes with binary values. (7)

Or

(b) (i) The minimum and maximum values for the attribute 'income' are 12,000 and 98,000. With min-max normalization, map a value of 73,600 for 'income' to the range [0.0, 1.0]. (4)

(ii) The mean and standard deviation of the values for the attribute 'income' are 54,000 and 16,000. With z-score normalization, transform a value of 73,600 for income. (3)

(iii) The attribute age has the values :

33, 40, 21, 25, 13, 22, 15, 30, 19, 35, 16, 20, 16, 22, 25, 25, 33, 36. Use smoothing by bin means to smooth these data, using a bin depth of 3. (4)

(iv) What is the need for data reduction? (2)

13. (a) (i) What is Apriori method? (2)

(ii) Using the following table, generate Association rules using Apriori method. Minimum Support = 33%, Minimum confidence = 68%. (11)

TID	List of Items
T001	{Calendar, Stickers, CDs}
T002	(Stickers, Books, Magazines}
T003	{Calendar, Books, Magazines)
T004	{Stickers, Books, Magazines, CDs}
T005	{Magazines, CDs}
T006	{Calendar, CDs}
T007	(Stickers, Books}
T008	{Calendar, Books, CDs)
T009	{Calendar, Stickers, Books, CDs}
T0010	{Books, CDs}

Or

(b) (i) What is constraint based association mining? Explain its working for mining frequent item sets from Transactional databases. (6)

(ii) List advantages and disadvantages of the candidate generation. (4)

(iii) How can the efficiency of the association rule mining be enhanced? (3)

14. (a) (i) Differentiate classification and prediction methods. (4)

(ii) What is meant by 'False Positives' and 'True Negatives' in accuracy calculation? (4)

(iii) List the different types of outcome based on splitting criterion. Give examples for each. (5)

Or

(b) Given the following table of students' grade : (13)

Midterm :	72	50	81	74	94	86	59	83	65	33	88	81
Final :	84	63	77	78	90	75	49	79	77	52	74	90

Use the method of least squares to find an equation for the prediction of a student's final exam grade based on the student's midterm grade in the course. Also predict the final exam grade of a student who received an 86 in the midsem exam.

15. (a) (i) Differentiate Agglomerative and Divisive clustering. (2)

(ii) How is the k-Medoids clustering method different from Agglomerative and Divisive clustering. Explain the method and develop an algorithm. (11)

Or

(b) (i) Develop an algorithm for k-means method. Group the following data set into two groups using k-means algorithm. (9)

Medicine	Attribute 1(X) : weight	Attribute 2(Y) : pH index
Medicine A	1	1
Medicine B	2	1
Medicine C	4	3
Medicine D	5	4

(ii) Compare the grid-based and density-based methods of clustering. (4)

PART C — (1 × 15 = 15 marks)

16. (a) Consider the following training set for a binary classification problem. (15)

Person-ID	Qualification	Player	Community	Class
1	PG	Yes	BC	E1
2	UG	Yes	OC	E1
3	PG	No	MBC	E1
4	UG	Yes	BC	E2
5	UG	No	BC	E2

Person-ID	Qualification	Player	Community	Class
6	UG	No	MBC	E2
7	PG	Yes	OC	E1
8	PG	No	BC	E1
9	UG	No	MBC	E1
10	UG	Yes	MBC	E2
11	PG	No	BC	E2
12	UG	No	BC	E2
13	PG	Yes	OC	E2
14	PG	Yes	MBC	E1
15	UG	No	BC	E1

Using the table,

(i) Order the attributes using Information Gain method.

(ii) Order the attributes using Gain Ratio method.

Or

(b) What are "Lazy learners"? How are they used for data mining? Give example.

———————

Reg. No. : | | | | | | | | | | | | |

Question Paper Code : KJ1512

M.Sc. DEGREE EXAMINATION, FEBRUARY/MARCH 2017.

Third Semester

Computer Science

DCS 7302 — DATA WAREHOUSING AND MINING

(Regulations 2013)

Time : Three hours Maximum : 100 marks

Answer ALL questions.

PART A — (10 × 2 = 20 marks)

1. Define data mart and Virtual Warehouse.
2. Compare OLTP & OLAP.
3. Why data preprocessing is required?
4. List the issues during data integration.
5. Define support and confidence in Association rule mining.
6. Give few techniques to improve the efficiency of Apriori algorithm.
7. Define classification and prediction.
8. What is confusion matrix?
9. What are the different types of data used for clustering?
10. What is CURE?

PART B — (5 × 16 = 80 marks)

11. (a) With a neat diagram, explain the architecture of data warehouse.

Or

(b) (i) With an example explain about OLAP Operations. (8)

(ii) Discuss the various schemas for multidimensional database. (8)

12. (a) Explain about the architecture of typical data mining system and steps involved in KDD process.

Or

(b) Briefly explain about the various data reduction techniques with example. (16)

13. (a) (i) With an example explain about data mining functionalities. (8)

(ii) Explain about constraint based association rule mining. (8)

Or

(b) Write an algorithm to find the frequent item set using candidate generation and also explain it with example. (16)

14. (a) What is classification? Explain how decision trees are used for classification. (16)

Or

(b) (i) Explain in detail about the Back Propagation technique. (10)

(ii) Explain the issues regarding classification and prediction. (6)

15. (a) What is clustering? Explain the K-Means and K-Mediods clustering algorithm with example. (16)

Or

(b) Explain about :

(i) Clustering high dimensional data (8)

(ii) Constraint based cluster analysis. (8)

Reg. No. : | | | | | | | | | | | | | |

Question Paper Code : S1512

M.Sc. DEGREE EXAMINATION, FEBRUARY/MARCH 2016.

Third Semester

Computer Science

DCS 7302 — DATA WAREHOUSING AND MINING

(Regulations 2013)

Time : Three hours Maximum : 100 marks

Answer ALL questions.

PART A — (10 × 2 = 20 marks)

1. Differentiate OLTP and OLAP.
2. Mention the advantages of bitmap indexing.
3. List the five primitives for specifying data mining tasks.
4. Name the methods used for dimensionality reduction.
5. What is meant by frequent item set in pattern mining?
6. Define a null-invariant measure.
7. What are the advantages of tree pruning?
8. How do case-based reasoning classifiers work?
9. What is meant by Tanimoto coefficient?
10. Name the grid based methods for clustering.

PART B — (5 × 16 = 80 marks)

11. (a) Explain the following

(i) Data cube (8)

(ii) Fact Constellation (4)

(iii) Starnet Query Model (4)

Or

(b) Explain the following

(i) 3-tier data warehousing architecture (8)

(ii) OLAP query processing (8)

12. (a) Explain the process of knowledge discovery from various databases with appropriate examples. (16)

Or

(b) (i) Assume that the values of the income attribute are 2000, 3000, 4000, 6000 and 10000. The income has to be mapped to the range [0.0, 1.0]. Do min-max normalization, z-score normalization and decimal scaling for income attribute. (8)

(ii) Describe the concept hierarchy generation for numerical data. (8)

13. (a) (i) Explain the Apriori algorithm. (8)

(ii) Consider the following transactions. What association rules can be found in the set, if the minimum support is 6% and the confidence is 80%?

T1 {f, a, d, b}

T2 {d, a, c, e, b}

T3{c,a, b,e}

T4{b, a, d} (8)

Or

(b) (i) Explain the techniques for mining multilevel association rules and quantitative association rules. (8)

(ii) Describe constraint based association mining with appropriate example. (8)

14. (a) (i) Explain how Bayesian classification works with a suitable example. (8)

(ii) Explain how back-propagation networks are used in classification. (8)

Or

(b) (i) Explain any two regression models. (8)

(ii) Write the Ada boost algorithm. Explain how it is used to increase the classifier and predictor accuracy. (8)

15. (a) (i) Consider the points A1(2,10), A2(2, 5), A3(8, 4), B1(5,8), B2(7, 5), B3(6, 4), C1(1, 2) and C2(4, 9). Assume that Euclidean distance is used and the initial centers of clusters are A1, B 1 and C2. Show the final clusters along with their centers using k -means algorithm. (8)

(ii) Explain any two density based methods for clustering. (8)

Or

(b) (i) Describe how high-dimensional data can be clustered. (8)

(ii) Explain constraint based cluster analysis. (8)

Reg. No. : | | | | | | | | | | | |

Question Paper Code : 80512

M.Sc. DEGREE EXAMINATION, AUGUST 2015.

Third Semester

Computer Science

DCS 7302 — DATA WAREHOUSING AND MINING

(Regulations 2013)

Time : Three hours Maximum : 100 marks

Answer ALL questions.

PART A — (10 × 2 = 20 marks)

1. Mention some of the benefits of Data Warehousing.

2. State any two differences between technical and business meta data.

3. Compare Multi-dimensional with Multi-relational OLAP.

4. What is data discretization?

5. What are the possible ways of integrating Data Mining system with a Data warehouse?

6. Define Pattern. Mention the ways to identify the interestingness of patterns.

7. What is lazy learner? Give example.

8. What is correlation analysis?

9. What is an Outlier? Mention its applications.

10. Mention the difference between agglomerative and divisive hierarchical clustering.

PART B — (5 × 16 = 80 marks)

11. (a) Explain the process of Knowledge Discovery in Databases. (16)

Or

(b) (i) Explain the need for Data Preprocessing. (8)

(ii) Explain the concept of Data Discretization and Concept Hierarchy generation. (8)

12. (a) How does a Clustering and Nearest Neighbor Prediction Work? Differentiate. (16)

Or

(b) (i) Differentiate the following :

(1) Discovery and Prediction. (4)

(2) Characterisation and Clustering. (4)

(ii) Explain Bayesian classification in detail. (8)

13. (a) (i) What is the algorithm used to find frequent itemsets using Candidate Generation. Explain. (12)

(ii) Mention its methods for improving the efficiency. (4)

Or

(b) (i) Explain the process of classification using Decision Tree Induction. (12)

(ii) What are some enhancements to basic decision tree induction? (4)

14. (a) (i) Explain the concept of Data warehouse and its impact on their enterprise initiatives. (8)

(ii) How does a Data warehouse affect the existing system? (8)

Or

(b) (i) Differentiate OLAP and OLTP with reference to its characteristics. (8)

(ii) With example mention how tables and spreadsheets are converted to Data Cubes. (8)

15. (a) (i) Write the Procedure for Principal Component Analysis. (12)

(ii) Explain the techniques used for Numerosity Reduction. (4)

Or

(b) (i) What are the best ways to measure a data mining tool (8)

(ii) What future holds for embedded data mining? (8)

———————

Reg. No. : | | | | | | | | | | | | |

Question Paper Code : 22463

M.Sc. DEGREE EXAMINATION, FEBRUARY/MARCH 2015.

Third Semester

Computer Science

DCS 7302 — DATA WAREHOUSING AND MINING

(Regulations 2013)

Time : Three hours Maximum : 100 marks

Answer ALL questions.

PART A — ($10 \times 2 = 20$ marks)

1. Define data warehouse and its characteristics.

2. Compare and contrast OLAP and OLTP.

3. Are KDD and data mining same? Justify your answer.

4. Mention any two techniques that are commonly used for handling noise data values.

5. Define support and confidence in association rule mining.

6. What are closed itemsets? Why are they mostly preferred than frequent itemsets in association rule mining?

7. Distinguish supervised and unsupervised learning techniques with an example for each.

8. What are the two approaches in decision tree pruning? Also specify the reason for pruning.

9. Calculate the distance between two objects, $A = (10, 4, 2, 10)$ and $B = (4, 2, 1, 4)$ using the Manhattan and Euclidean distance measures.

10. What is an outlier? Mention the methods of detecting outliers.

PART B — (5 × 16 = 80 marks)

11. (a) With a neat diagram, give a brief explanation of the various components of three-tier data warehouse architecture.

Or

(b) (i) Suppose that a college data warehouse consists of four dimensions course, student, staff, department and two measures count and charge, where charge is the fee that a department charges a student for offering a course. Draw a star schema diagram for the data ware house and also give necessary details. (8)

(ii) State the different classes of OLAP tools. Illustrate the characteristics of each class of OLAP tool. Also, Specify the similarities and differences between them. (2 + 3 + 3)

12. (a) Illustrate the knowledge discovery process, list the phases, and indicate the activities in each phase.

Or

(b) Explain the need of data preprocessing and steps of data preprocessing in detail.

13. (a) Illustrate with an example, explain each of the following data mining functionalities: Association and Correlation analysis, Classification, Prediction, Clustering and Evolution analysis.

Or

(b) Given the following database with 5 transactions and a minimum support threshold of 60% and a minimum confidence threshold of 80%, compute all frequent itemsets using Apriori algorithm by generating candidate itemsets. List all strong association rules obtained from frequent itemsets. Specify, how FP-Growth is better than Apriori?

Trans Id	Itemset
1	M,O,N,K,E,Y
2	D,O,N,K,E,Y
3	M,A,K,E
4	M,U,C,K,Y
5	C,O,L,I,E

14. (a) What are the three phases of decision tree construction? Illustrate each of these phases in detail.

Or

(b) "Naive Bayes is a lazy classifier". What does it mean? Mention the advantages and disadvantages of Naïve Bayes classifier. Briefly outline the major ideas of Naïve Bayes classification. (2 + 8 + 6)

15. (a) Explain the steps in k-means algorithm. Cluster the following data set of 10 objects into 3 clusters using k-means and Euclidean distance function: x1 (4,8), x2 (5,6), x3 (5,10), x4 (6,9), x5 (8,4), x6 (8,6), x7 (9,5), x8 (9,6), x9 (10,7), x10 (9,8). Draw a 10 by 10 space with all the 10 points and show the clusters after each iteration.

Or

(b) Why similarity metric is so important in clustering? List the difficulties in handling categorical data for clustering. Briefly outline the computation of dissimilarity between objects described by the following types of variables :

(i) Interval–scaled variables

(ii) Asymmetric binary variables

(iii) Categorical variables

(iv) Ratio-scaled variables.
