Computer Engineering Department, S V N I T, Surat.
Mid-Semester Examinations, March 2019
M Tech I (CO), 2nd Semester
Course: Data Mining and Data Warehousing (CO622)

Dated: 06th March 2019 Time: 14.00 hrs to 15.30 hrs Max Marks: 30

Instructions
- **Please start** the answer to **each question on new page ONLY** of your answer sheets.
- Please write your correct exam no without fail on the answer sheets as well as the question papers.
- Please be brief and to-the-point in the attempting the answers to the theory questions. Unnecessary and unjustified elaboration will not fetch more marks. The Hallmark of an answer to a short question indeed is its brevity.

Q.1 What are the major challenges of mining a huge amount of data (such as billions of tuples) in comparison with mining a small amount of data (such as a few hundred tuple data set)? [02]

Q.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, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. 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. [03]

Q.3 Suppose a hospital tested the age and body fat data for 18 randomly selected adults with the following result. [06]

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

a. Calculate the mean, median and standard deviation of *age* and *%fat*.
b. Calculate the *correlation coefficient* (Person's product moment coefficient). Are these two variables positively or negatively correlated? [15]

Q.4 Suppose model is trained to predict whether an email is Spam or Not Spam. After training the model, for test set of 500 new email messages (also labeled) and the model produces the contingency matrix below. [06]

		True Class	
		Spam	Not Spam
Predicted Class	Spam	70	10
	Not Spam	40	380

a. Compute the precision of this model with respect to the Spam class.
b. Compute the recall of this model with respect to the Spam class.
c. What is your opinion if model supports High-precision and low recall with respect to SPAM and High recall and low precision with respect to SPAM? Justify.

Q.5

TID	Items
11	C, D, H, I
12	C, E, F
13	B, C, D, F
14	A, B, C, D
15	C, H, I, J
16	A, D, E, E, F, H
17	F, G, H
18	A, D, H
19	D, E, F
20	B, C, D, E, H

a. Write property, bottlenecks with solutions for Apriori algorithm. [03]
b. For the given transactions, discover frequent itemsets using ECLAT approach for Supp.count=2. [03]

OR

a. Explain any one correlation measure which is not affected by null transaction.
b. For the given transactions, discover closed itemsets for Supp.count=1.
c. With the help of equation ONLY, calculate the possible total number of frequent items from 1 to K itemsets for Supp.count=1. [02]
d. For the given transactions, discover association rules {X} → {Y} and {Y} → {X} having the same confidence and Supp.count=2. [03]
e. What is the maximum size of K of frequent itemsets for Supp.count=1? Make the generalized statement for your answer. [02]

Computer Engineering Department, S V N I T, Surat.
Supplimentary Examinations, July 2019

M.Tech. I (CO) – 2nd Semester
Course : Software Engineering Methodologies-CO604

Dated: 9th July, 2019 Time: 10:00 hrs to 13:00 hrs Max Marks: 50

Q.1 (A) The controlling software for elevator system is to be modeled using petri nets. Identify atleast five rules for the elevator system and specify them using 1) the natural language specification and 2) petri nets. [7]

Q.1 (B) Identify the scenario where the tokens are required to assign some values in the petri net modelling. [3]

Q.2 (A) Draw the finite state machine model for producer consumer problem. Clearly specify the requirements. [5]

Q.2 (B) Show the importance of the use of life cycle model. For large and complex applications, which model do you suggest ? Explain with suitable diagram and application scenario. [10]

Q.2 (C) Discuss the problems associated with non functional requirement elicitation process. [5]

Q.3 (A) Specify an abstract data type Bag (i.e. Multiset) [7]

A bag is similar to a set, which can be thought of as a collection of elements taken from some universe of elements. The distinction between the two is that, with respect to a set, each element of the universe is simply either a member, or not a member. With a bag, we have the notion that an element may occur in it any number of times (i.e., zero or more). For example, the set given by the enumeration

{ 4, 2, 0, 2, 3, 4, 7 }

is exactly the same as the one given by

{ 2, 4, 7, 3, 0 }

However, viewed as bags, they differ because the first contains two occurrences of both 2 and 4, while the second contains only one of each.

Write the algebraic specification for Bags while considering the following operations:

Sort - Bag (Elem)

isEmpty→ checks whether a bag is empty or not

NumOcc →returns number of occurrences of an item in a bag

Size→ returns a total number of occurrences of all items

Empty→ yields a bag with no members

Insert→yields a bag obtained by inserting an elem

Delete→yields bag obtained by deleting an elem

Q.3 (B) Explain any three properties of Formal methods in software specification. [3]

Q.4 (A) Consider the software for library management system with the following requirements: [5]

- A library maintains records of all the books and members.
- To borrow a book, a member requests for a book and a librarian check the status of the book. If the book is available in the library, then a book is issued to the member.
- Upon successful issue of the book, a receipt is generated.

Design a sequence diagram for the above scenario.

Q.4 (B) Tic-Tac-Toe is a computer game in which a human player and the computer make alternate moves on a 3 x 3 square. A move consists of marking a previously unmarked square. The player who is first to place three consecutive marks along a straight line on the square wins. As soon as either the human player or the computer wins, a message congratulating the winner is displayed. If neither player manages to get three consecutive marks along a string line, and all the square on the board are filled up, then the game is drawn. The computer always tries to win a game. Draw Level 0 and Level 1 data flow diagram. [5]

Marks: 50

Sardar Vallabhbhai National Institute of Technology, Surat
Computer Engineering Department
End Semester Examination, Nov. – Dec. 2019
B.Tech. IV (CO) – 7th Semester
DATA WAREHOUSING AND MINING (CO415)

Date: 5th December 2019 Time: 15.30 hrs to 18.30 hrs Max Marks: 100

Instructions: 1. Please start the answer to each question on new page ONLY of your answer sheets.
2. Please write your correct exam no without fail on the answer sheets as well as the question papers.

Q. 1 Answer the following questions [Any Two]: [08]

1. Describe the data mining application in financial data analysis and intrusion detection.
2. Explain suitable Regression method for predicting categorical variable.
3. What are the difference between centroid and medoids? Are centroids always data instances in the dataset? Are medoids always data instances in the dataset? Explain.

Q. 2 Answer the following questions [Any Three]: [18]

1. What is SVM? Explain LSVM and NLSVM using diagram. Explain how NLSVM handle the misclassification of data.
2. What is Fourier Transform (FT)? Explain why Wavelet transform is used as a replacement of Fourier Transform?
3. What are underfitting and overfitting? How to detect and prevent overfitting problem in classification?
4. What is Posterior probability and Prior probability? Explain in detail how it is used to predict the class label.

Q. 3 Answer the following questions [Any Two]: [16]

1. Why is it important to consider density when clustering a dataset? How does DBSCAN use different points to cluster the dataset?
2. What are the limitations of Agglomerative clustering? Explain suitable one clustering method which is used to overcome those limitations?
3. What are the limitations of K-means clustering? Explain Clustering Using REpresentatives method to overcome above limitations.

Q. 4 Answer the following questions:

1. Calculate Principal Component Analysis for the following dataset: [18]

X	2	3	5	7	9
Y	1	4	0	6	2

2. Explain Graphic Displays of Basic Statistical Descriptions of Data. [12]

Q. 5[A] Answer the following questions: [06]

1. Explain different types of association rules based on the number of attributes usage together with their examples where one of the attribute is color of item. [20]
2. Explain the utilization of association rule for dynamic discretization.
3. Explain the OLAP server which is scalable in nature.
4. Explain the different parameters that can affect the data warehouse storage.
5. Explain the default indexing technique of data warehouse.

Q. 5[B] Answer the following questions: [20]

1. Define the usage of the closed item sets and discover them for the given transactions for absolute support>=2.

   T1: a1, a2, a3, …, a99, a100
   T2: a51, a52, a53, …, a99, a100
   T3: a1, a2, a3, …., a99, a100

2. Explain the technique MIS.
3. Justify the statement: The strong association rules are suitable for all applications.
4. Compare the query driven approach database integration over data warehouse.

OR

Q. 5[B] Answer the following questions: [20]

1. Define the usage of the maximal item sets and discover them for the given transactions for absolute support>=3.

   T1: a1, a2, a3, …, a99, a100
   T2: a51, a52, a53, …, a99, a100
   T3: a1, a2, a3, …., a99, a100
   T4: a1, a2, a3, …, a50

2. Explain the technique CAR.
3. Justify the statement: The introduction/truncation of null transactions does not affect the result of Apriori algorithm.
4. Compare the various data warehouse schemas.

*

Marks: 50

Computer Engineering Department, S V N I T, Surat
Make-Up Mid Sem Examinations, November 2019
B Tech IV (CO) – 7th Semester
Course: Data Warehousing and Mining (CO415)

Date: 28th November 2019 Time: 16.00 hrs to 17.30 hrs Max Marks: 30

Instructions: 1. Please start the answer to each question on new page ONLY of your answer sheets.
2. Please write your correct exam no without fail on the answer sheets as well as the question papers.

Q.1 Answer the following questions:

1. What is data mining functionality? Explain different types of data mining functionality with examples. [06]
2. Consider 30 students who appeared for a class test. Determine the z-test score for the 4th student based on the marks scored by the students out of 100 : 55, 67, 84, 65, 59, 68, 77, 95, 88, 78, 53, 81, 73, 66, 65, 52, 54, 83, 86, 94, 85, 72, 62, 64, 74, 82, 58, 57, 51, 91. [04] [02]

Q.2 Answer the following questions:

1. For the given dataset, Apply Naïve-Bayes' algorithm and predict the outcome for the car where color = Red, Type = SUV, Origin = Domestic. [08]

Color	Type	Origin	Stolen
Red	Sport	Domestic	Yes
Red	Sport	Domestic	No
Red	Sport	Domestic	Yes
Yellow	Sport	Domestic	No
Yellow	Sport	Imported	Yes
Yellow	SUV	Imported	No
Yellow	SUV	Imported	Yes
Yellow	SUV	Domestic	No
Red	SUV	Imported	No
Red	Sports	Imported	Yes

2. Explain difference between K-means and k-medoids clustering algorithm with appropriate example.

Q.3 Plot a Dendrogram using average linkage Agglomerative clustering for the following data set. [06]

	P1	P2	P3	P4	P5	P6
P1	0					
P2	0.23	0				
P3	0.22	0.15	0			
P4	0.37	0.20	0.15	0		
P5	0.34	0.14	0.28	0.29	0	
P6	0.23	0.25	0.11	0.22	0.39	0

Q.4 Answer the following questions: [10]

1. Enlist the thresholds other than the Support and Confidence of the association rule mining and also explain any one.
2. Justify the statement: "Performance of Association rule mining approach is dependent upon the thresholds".
3. Explain the maximal itemset and also explain the application which will be suitable for it.
4. Explain the different types of OLAP server.
5. Explain the different types of cube and default indexing available with data warehouse.

*

Computer Engineering Department, S V N I T, Surat
Mid-Semester Examinations, Sept. - Oct. 2019

B Tech IV (CO) – 7th Semester
Course: Data Warehousing and Mining (CO415)

Date: 4th October 2019 Time: 16.00 hrs to 17.30 hrs Max Marks: 30

Instructions: 1. Please start the answer to each question on new page ONLY of your answer sheets.
2. Please write your correct exam no without fail on the answer sheets as well as the question papers.

Q. 1 Answer the following questions:

1. Discuss the issues of data mining in detail. **[06]**
2. Find the Five Number Summary for the following data set: [04]
   10 11 12 25 25 27 31 33 34 34 35 36 43 50 59
   Draw a box-and-whisker plot for the given data set. [02]

OR

1. With a neat diagram explain the architecture of data mining. [04]
2. The daily sale of sugar in a certain grocery shop is given below: [02]

Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
75 kg	120 kg	12 kg	50 kg	70.5 kg	140.5 kg

The average daily sale is 78 Kg. Calculate the variance and the standard deviation of the above data.

Q. 2 Answer the following questions:

1. Find the Least Square Regression line, $y = ax + b$ for the following data. Also, estimate the value of y when $x = 10$. **[08]**

X	0	1	2	3	4
Y	2	3	5	4	6

2. Explain two techniques of finding a good subset of the original attributes with example. Also write difference between those two techniques.

Q. 3 Perform Agglomerative algorithm on the following data and plot a dendrogram using Single link approach. The given data indicates the distance in kilometres between some Italian cities. **[06]**

	BA	FI	MI	NA	RM	TO
BA	0	662	877	255	412	996
FI	662	0	295	468	268	400
MI	877	295	0	754	564	138
NA	255	468	754	0	219	869
RM	412	268	564	219	0	669
TO	996	400	138	869	669	0

Q. 4 Answer the following questions [Any Two]: **[04]**

1. Explain the Apriori property and the thresholds of the association rule mining.
2. Justify the statement: "The change in dataset format from horizontal to vertical, may impact the performance of Association rule mining approach".
3. Write the drawback of Apriori algorithm and the ways to overcome that.

Q. 5 For the given Table, perform: **[06]**

TID	Items
1	Red, Blue
2	Red, White, Blue
3	White, Orange
4	Red, White, Blue, Green
5	Red, White, Green
6	White, Orange
7	White, Blue
8	Red, White, Orange
9	Yellow
10	Red, White, Blue

1. Identify the transaction length range for the given transactions and derive the maximum size (K) of frequent itemsets that can be extracted, if support = 1. Make the generalized statement for your answer. [02]
2. With the help of steps, generate the FP-Tree for support = 30% and also generate conditional FP-Tree for the item with last item of f-list. [04]

OR

Discover the rules {a} → {b} and {b} → {a} have the same confidence threshold from the existence of a 2-itemsets {a, b} with support = 50%.

*