

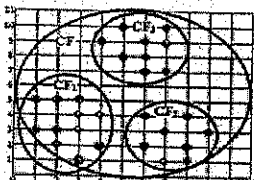
B.E. INFORMATION TECHNOLOGY 4TH YEAR 1ST SEMESTER EXAMINATION - 2019

Subject: Data Mining

Time: 3 Hours

Full Marks: 100

(Note: Answers of all parts/subparts of a question should be written together)

| CO1 (10) | <p>Q.1 Answer either (a) or (b):</p> <p>a. i) Define with suitable examples: nominal attributes and ordinal attributes. ii) What are the major issues in data mining? iii) Explain various types of data visualization techniques.</p> <p>b. i) Describe various methods/tests/techniques with suitable example(s) used in four levels of data preprocessing. ii) Differentiate between data sets and data objects with suitable examples.</p> | 3+4+3 6+(2+2) | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|-------------|--|------------------|---------|----|---------|----|---------|----|------------|----|------|----|---------|----|------|------|-------|------|------|-------|------|-----|-----------------------------------|---|---|----|---|---|-----|---|---|----|---|---|-----|---|---|----|---|---|-----|---|---|-----------------------|
| CO2 (25) | <p>Q.2 Answer (a) and any one from (b) and (c):</p> <p>a. <table border="1" data-bbox="231 694 486 862"> <thead> <tr> <th>Objects</th> <th>X</th> <th>Y</th> <th>Objects</th> <th>X</th> <th>Y</th> </tr> </thead> <tbody> <tr> <td>X1</td> <td>4</td> <td>5</td> <td>X7</td> <td>-1</td> <td>2</td> </tr> <tr> <td>X2</td> <td>-2</td> <td>3</td> <td>X8</td> <td>8</td> <td>9</td> </tr> <tr> <td>X3</td> <td>1</td> <td>-1</td> <td>X9</td> <td>6</td> <td>7</td> </tr> <tr> <td>X4</td> <td>3</td> <td>8</td> <td>X10</td> <td>3</td> <td>0</td> </tr> <tr> <td>X5</td> <td>4</td> <td>5</td> <td>X11</td> <td>2</td> <td>1</td> </tr> <tr> <td>X6</td> <td>2</td> <td>0</td> <td>X12</td> <td>1</td> <td>1</td> </tr> </tbody> </table> <p>i) The following table consists of twelve objects (i.e., points) in 2-D system. Find best possible two clusters formed by the objects using the algorithms: k-means and k-medoids. ii) Describe generalized steps of CLIQUE algorithm. Hence find out its strengths and weaknesses.</p> <p>b. i) Find out the outliers w.r.t. the data set; $D = \{2.37, 2.16, 14.82, 1.73, 41.04, 0.23, 1.32, 2.91, 39.41, 0.11, 27.44, 4.51, 60.05, 0.51, 4.50, 0.18, 14.68, 4.66, 1.30, 2.06, 49.09, 1.19\}$ using boxplots. ii) Explain the drawbacks of k-means and k-medoids algorithms (with suitable examples). iii) Define relative interconnectivity and relative closeness w.r.t. multiphase hierarchical clustering.</p> <p>c.  <p>i) The following figure represents three disjoint clusters and their merger C_1, C_2, C_3 and C respectively in 2-D system. Construct a CF-tree using C_1, C_2, C_3, and C. ii) Write a short note on CALARANS algorithm.</p> </p></p> | Objects | X | Y | Objects | X | Y | X1 | 4 | 5 | X7 | -1 | 2 | X2 | -2 | 3 | X8 | 8 | 9 | X3 | 1 | -1 | X9 | 6 | 7 | X4 | 3 | 8 | X10 | 3 | 0 | X5 | 4 | 5 | X11 | 2 | 1 | X6 | 2 | 0 | X12 | 1 | 1 | 5x2+5 4+3+3 6+4 |
| Objects | X | Y | Objects | X | Y | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| X1 | 4 | 5 | X7 | -1 | 2 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| X2 | -2 | 3 | X8 | 8 | 9 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| X3 | 1 | -1 | X9 | 6 | 7 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| X4 | 3 | 8 | X10 | 3 | 0 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| X5 | 4 | 5 | X11 | 2 | 1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| X6 | 2 | 0 | X12 | 1 | 1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| CO3 (20) | <p>Q.3 Answer (a) and any one from (b) and (c):</p> <p>a. The following table consists of five transactions with $\text{min_sup} = 55\%$ and $\text{min_conf} = 75\%$.</p> <table border="1" data-bbox="231 1299 375 1444"> <thead> <tr> <th>Tid</th> <th>Items</th> </tr> </thead> <tbody> <tr> <td>10</td> <td>1, 3, 4</td> </tr> <tr> <td>20</td> <td>2, 3, 5</td> </tr> <tr> <td>30</td> <td>1, 2, 3, 5</td> </tr> <tr> <td>40</td> <td>2, 5</td> </tr> <tr> <td>50</td> <td>1, 3, 5</td> </tr> </tbody> </table> <p>i) Find all frequent itemsets using <i>Apriori</i> and <i>FP-growth</i> respectively. Compare the efficiency of two mining processes. ii) List all strong association rules (with support $s = 30\%$ and confidence $c = 60\%$) matching the following <i>metarule</i>, where <i>item_i</i> denotes variables representing items: $\forall X \in \text{transaction}, \text{buys}(X, \text{item}_1) \wedge \text{buys}(X, \text{item}_2) \Rightarrow \text{buys}(X, \text{item}_3) \quad [s, c]$</p> <p>b. The following contingency table summarizes supermarket transaction data. Suppose that the association rule "game \Rightarrow video" is mined with a minimum support threshold of 30% and a minimum confidence threshold of 55%.</p> <table border="1" data-bbox="231 1590 391 1691"> <thead> <tr> <th></th> <th>game</th> <th>game</th> </tr> </thead> <tbody> <tr> <td>video</td> <td>4000</td> <td>3500</td> </tr> <tr> <td>video</td> <td>2000</td> <td>500</td> </tr> </tbody> </table> <p>i) Check whether the association rule is misleading strong or not. ii) If it is misleading then what measure(s) will you take and how (in different ways) to filter the misleading strong association rule?</p> <p>c. i) Suppose a frequent itemset contains 200 items. What is the maximum number of frequent itemsets that can be generated from the data? ii) What is antimonotonicity? How can it be used in data mining? iii) How can frequent itemsets be mined using vertical data format?</p> | Tid | Items | 10 | 1, 3, 4 | 20 | 2, 3, 5 | 30 | 1, 2, 3, 5 | 40 | 2, 5 | 50 | 1, 3, 5 | | game | game | video | 4000 | 3500 | video | 2000 | 500 | (3+5+1)+3 2+(3+3) 2+(2+2)+2 | | | | | | | | | | | | | | | | | | | | | |
| Tid | Items | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 10 | 1, 3, 4 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 20 | 2, 3, 5 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 30 | 1, 2, 3, 5 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 40 | 2, 5 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 50 | 1, 3, 5 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | game | game | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| video | 4000 | 3500 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| video | 2000 | 500 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

| CO4 (20) | Q.4 Answer either (a) or (b): | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|-------------|--|---|---------|---------------|---------------|---------------|---------------|---|--------|------|-----|------|-----|---|-------|--------|----|-----------|-----|---|--------|------|----|-----------|----|---|--------|-----|----|------|-----|---|-------|-----|-----|------|----|---|--------|--------|-----|-----------|-----|---|--------|-----|----|-----------|----|---|-------|------|-----|------|----|--|
| | <table border="1"> <thead> <tr> <th>RID</th> <th>age</th> <th>income</th> <th>student</th> <th>credit rating</th> <th>buys computer</th> </tr> </thead> <tbody> <tr><td>1</td><td>senior</td><td>high</td><td>yes</td><td>fair</td><td>yes</td></tr> <tr><td>2</td><td>youth</td><td>medium</td><td>no</td><td>excellent</td><td>yes</td></tr> <tr><td>3</td><td>middle</td><td>high</td><td>no</td><td>excellent</td><td>no</td></tr> <tr><td>4</td><td>senior</td><td>low</td><td>no</td><td>fair</td><td>yes</td></tr> <tr><td>5</td><td>youth</td><td>low</td><td>yes</td><td>fair</td><td>no</td></tr> <tr><td>6</td><td>senior</td><td>medium</td><td>yes</td><td>excellent</td><td>yes</td></tr> <tr><td>7</td><td>middle</td><td>low</td><td>no</td><td>excellent</td><td>no</td></tr> <tr><td>8</td><td>youth</td><td>high</td><td>yes</td><td>fair</td><td>no</td></tr> </tbody> </table> | RID | age | income | student | credit rating | buys computer | 1 | senior | high | yes | fair | yes | 2 | youth | medium | no | excellent | yes | 3 | middle | high | no | excellent | no | 4 | senior | low | no | fair | yes | 5 | youth | low | yes | fair | no | 6 | senior | medium | yes | excellent | yes | 7 | middle | low | no | excellent | no | 8 | youth | high | yes | fair | no | <p>The following table consists of training data from an employee database. Assuming "buys_computer" as class label attribute, select appropriate splitting attribute and draw the decision tree accordingly using:</p> <ol style="list-style-type: none"> Information gain, Gain ratio, and Gini index |
| RID | age | income | student | credit rating | buys computer | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 1 | senior | high | yes | fair | yes | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 2 | youth | medium | no | excellent | yes | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 3 | middle | high | no | excellent | no | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 4 | senior | low | no | fair | yes | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 5 | youth | low | yes | fair | no | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 6 | senior | medium | yes | excellent | yes | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 7 | middle | low | no | excellent | no | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 8 | youth | high | yes | fair | no | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | <p>b. i) Why do we need tree pruning in making decision tree? Explain with suitable examples of pre pruning and post pruning.</p> <p>ii) How does tree pruning suffer from repetition and replication? Explain them with suitable example of training data set.</p> <p>iii) What is/are the problem(s) and solution(s) of overfitting data in making decision tree? Explain with suitable examples.</p> <p>iv) Write the definitions of coverage and accuracy. Find their values using the above database, (in Q.4 a) and satisfying the rule R, where $R: (age=youth) \wedge (student=yes) \Rightarrow (buys_computer=yes)$</p> | <p style="text-align: right;">6x3+2</p> <p style="text-align: right;">(2+3)+(3+3)+3+(4+2)</p> | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| CO5 (10) | <p>Q.5 Answer all:</p> <p>a. Explain: data mining for i) intrusion detection and prevention and ii) science and technology?</p> <p>b. How can mining be done on i) multimedia data and ii) web data?</p> | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| CO6 (15) | <p>Q.6 Answer (a) and any one from (b) and (c):</p> <p>a. Suppose that a data warehouse for Big University consists of the four dimensions student, course, semester, and instructor, and two measures count and avg_grade. At the lowest conceptual level (e.g., for a given student, course, semester, and instructor combination), the avg_grade measure stores the actual course grade of the student. At higher conceptual levels, avg_grade stores the average grade for the given combination.</p> <ol style="list-style-type: none"> Draw a snowflake schema diagram for the data warehouse. Starting with the base cuboid [student, course, semester, instructor], what specific OLAP operations (e.g., roll-up from semester to year) should you perform in order to list the average grade of CS courses for each Big University student. 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)? <p>b. Why do we need a separate Data Warehouse? Describe the multi-tier architecture of Data Warehouse with proper schematic diagram.</p> <p>c. Describe the working procedure(s) of the following OLAP operations with proper example: drill-up, roll-down, drill-across, pivot, and slice-dice.</p> | <p style="text-align: right;">5x2</p> <p style="text-align: right;">5+3+2</p> <p style="text-align: right;">2+3</p> <p style="text-align: right;">1x5</p> | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

-: Course Objectives :-

- CO1: Explain the basics and illustrate the major steps of preprocessing techniques of data mining.
- CO2: Illustrate, analyze and compare various functionalities (including methods) of data mining and apply them to solve problems.
- CO3: Explain, discuss, analyze the concept of mining frequent patterns and solve related problems.
- CO4: Explain and analyze various approaches of classification and prediction and employ them to solve problems.
- CO5: Discuss and illustrate various mining trends, applications and research frontiers of data.
- CO6: Review and outline the concept of data warehousing and its related tools (such as OLAP) and apply them to solve problems.