2018

1	N	T							

NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY

An Autonomous Institution Approved by UGC/AICTE/Govt. of Karnataka
Accredited by NBA (Tier – I) and NAAC 'A+' Grade
Affiliated to Visveswaraya Technological University, Belagavi
Post Box No. 6429, Yelahanka, Bengaluru – 560 064, Karnataka, India

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

MID SEMESTER EXAMINATION-III – Scheme & solution

Course Title with code	Data Mining, 18CS54	**Maximum Marks**	**30 Marks**
Date and Time	31/01/2022, 9.30am to 10.30am	**No. of Hours**	1.0
Course Instructor(s)	Dr. Vijaya Shetty S, Dr. Sujata Joshi, Dr. Vani V		
Instructions to Students			

1. Answer any **two full questions.**
2. Any missing data may assume suitably.

<table>
<tr><th>Q. No</th><th>Question</th><th>MAX MARKS</th><th>CO</th><th>BL</th><th>PO/ PSO</th></tr>
<tr><td>1. a</td><td>Draw a contingency table for the rules
1. {b} → {c},
2. {a} → {d},
3. {b} → {d} using the transactions shown in Table 1a.
Table 1a. Market basket transactions.
<table>
<tr><th>Transaction ID</th><th>Items Bought</th></tr>
<tr><td>1</td><td>$\{a, b, d, e\}$</td></tr>
<tr><td>2</td><td>$\{b, c, d\}$</td></tr>
<tr><td>3</td><td>$\{a, b, d, e\}$</td></tr>
<tr><td>4</td><td>$\{a, c, d, e\}$</td></tr>
<tr><td>5</td><td>$\{b, c, d, e\}$</td></tr>
<tr><td>6</td><td>$\{b, d, e\}$</td></tr>
<tr><td>7</td><td>$\{c, d\}$</td></tr>
<tr><td>8</td><td>$\{a, b, c\}$</td></tr>
<tr><td>9</td><td>$\{a, d, e\}$</td></tr>
<tr><td>10</td><td>$\{b, d\}$</td></tr>
</table>
Answer:
<table>
<tr><th></th><th>c</th><th>$\overline{c}$</th></tr>
<tr><td>b</td><td>3</td><td>4</td></tr>
<tr><td>$\overline{b}$</td><td>2</td><td>1</td></tr>
</table>
<table>
<tr><th></th><th>d</th><th>$\overline{d}$</th></tr>
<tr><td>a</td><td>4</td><td>1</td></tr>
<tr><td>$\overline{a}$</td><td>5</td><td>0</td></tr>
</table>
<table>
<tr><th></th><th>d</th><th>$\overline{d}$</th></tr>
<tr><td>b</td><td>6</td><td>1</td></tr>
<tr><td>$\overline{b}$</td><td>3</td><td>0</td></tr>
</table>
Scheme: 2 x 3 = 6 marks</td><td>6</td><td>4</td><td>3</td><td>1,2, 3/2</td></tr>
<tr><td>1. b</td><td>Use the contingency tables in 1.a to compute and rank the rules in decreasing order according to the following measures.
i)Support.
ii). Confidence.
iii) Interest X →Y)
iv) IS (X → Y)</td><td>9</td><td>4</td><td>3</td><td>1,2, 3/2</td></tr>
</table>

<table>
<tr>
<td></td>
<td>
i. Support.
Answer:
<table>
<tr><th>Rules</th><th>Support</th><th>Rank</th></tr>
<tr><td>$b \longrightarrow c$</td><td>0.3</td><td>3</td></tr>
<tr><td>$a \longrightarrow d$</td><td>0.4</td><td>2</td></tr>
<tr><td>$b \longrightarrow d$</td><td>0.6</td><td>1</td></tr>
</table>
ii. Confidence.
Answer:
<table>
<tr><th>Rules</th><th>Confidence</th><th>Rank</th></tr>
<tr><td>$b \longrightarrow c$</td><td>3/7</td><td>3</td></tr>
<tr><td>$a \longrightarrow d$</td><td>4/5</td><td>2</td></tr>
<tr><td>$b \longrightarrow d$</td><td>6/7</td><td>1</td></tr>
</table>
iii. Interest$(X \longrightarrow Y) = \frac{P(X,Y)}{P(X)}P(Y)$.
Answer:
<table>
<tr><th>Rules</th><th>Interest</th><th>Rank</th></tr>
<tr><td>$b \longrightarrow c$</td><td>0.214</td><td>3</td></tr>
<tr><td>$a \longrightarrow d$</td><td>0.72</td><td>2</td></tr>
<tr><td>$b \longrightarrow d$</td><td>0.771</td><td>1</td></tr>
</table>
iv. IS$(X \longrightarrow Y) = \frac{P(X,Y)}{\sqrt{P(X)P(Y)}}$.
Answer:
<table>
<tr><th>Rules</th><th>IS</th><th>Rank</th></tr>
<tr><td>$b \longrightarrow c$</td><td>0.507</td><td>3</td></tr>
<tr><td>$a \longrightarrow d$</td><td>0.596</td><td>2</td></tr>
<tr><td>$b \longrightarrow d$</td><td>0.756</td><td>1</td></tr>
</table>
Scheme :
Support :0.5 x 3 = 1.5 Marks
Confidence: 0.5 x 3 = 1.5 Marks
Interest: 1 x 3 = 3 Marks
IS: 1 x 3 = 3 Marks
</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. a</td>
<td>
Compare partition clustering and hierarchical clustering algorithms
<u>Partitioning Algorithms</u>:
The simplest and most fundamental version of cluster analysis is partitioning, which organizes the objects of a set into several exclusive groups or clusters.
The number of clusters is given as background knowledge. This parameter is the starting point for partitioning methods.
Given a data set, D, of n objects, and k, the number of clusters to form, a partitioning algorithm organizes the objects into k partitions k <=n, where each partition represents a cluster. The clusters are formed to optimize an objective partitioning criterion, such as a dissimilarity function based on distance, so that the objects within a cluster are “similar” to one another and “dissimilar” to objects in other clusters in terms of the data set attributes.
Typical methods: k-means, k-medoids, PAM, CLARA and CLARANS
<u>Hierarchical Algorithm</u>:
While partitioning methods meet the basic clustering requirement of organizing a set of objects into several exclusive groups, in some situations we may want to partition our data into groups at different levels such as in a hierarchy. A hierarchical clustering method works by grouping data objects into a hierarchy or “tree”(dendrogram) of clusters. Representing data objects in the form of a hierarchy is useful for data summarization and visualization.
Typical methods: DIANA, AGNES, BIRCH, and CAMELEON
Scheme: Each type description : 2 x 2 = 4 Marks
Typical methods for each : 2 x 0.5 = 1 Mark
</td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>1,2,
3/2</td>
</tr>
</table>

2. b | 10 | 5 | 3 | 1,2, 3/2

Consider the dataset given in the Table2.b with five objects characterized by a single continuous feature.

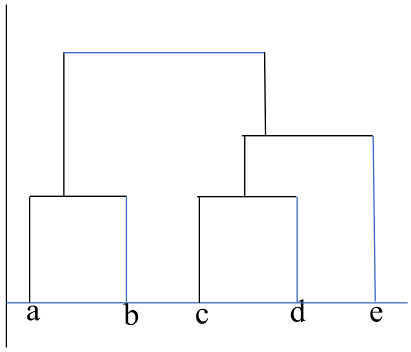
Table2.b Dataset

	a	b	c	d	e
Feature	1	2	4	5	6

Apply the agglomerative algorithm with single-link cluster distance measures to produce a dendrogram tree

Note : dist(c1,c2) = abs(c1-c2)

	a	b	c	d	e
a	0	1	3	4	5
b	1	0	2	3	4
c	3	2	0	1	2
d	4	3	1	0	1
e	5	4	2	1	0



	a, b	c	d	e
a, b	0			
C	2	0		
D	3	1	0	
E	4	2	1	0

	a, b	C, d	e
a, b	0	2	4
c, d	2	0	1
e	4	1	0

	a, b	C, d, e
a, b	0	2
c, d, e	2	0

Initial distance matrix + 3 iterations: 2 x 4 = 8 marks

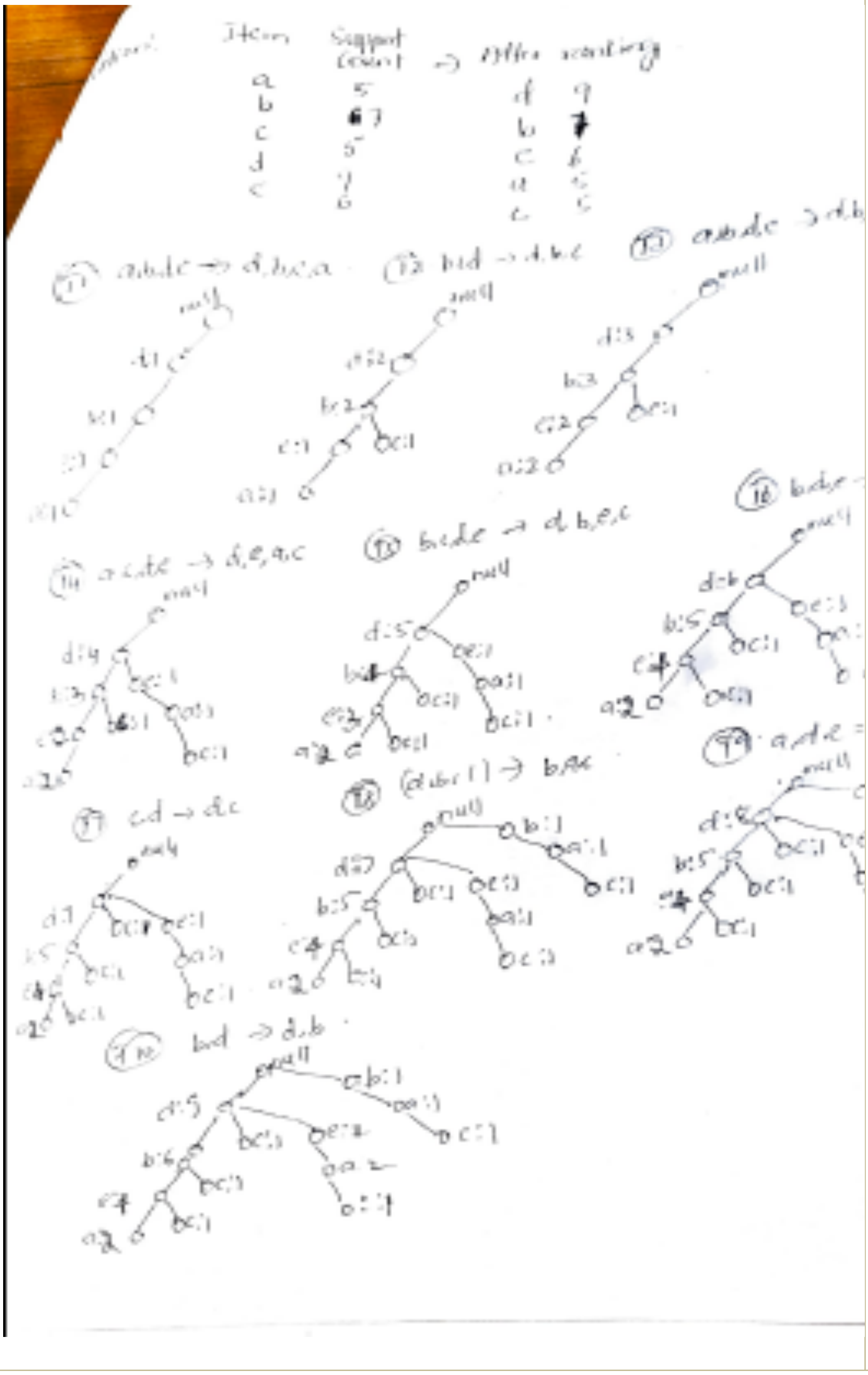
Dendrogram tree: 2 marks

3. a | 5 | 4 | 3 | 1,2, 3/2

Construct FP tree for the Table 3.a.

Market basket transactions

Transaction ID	Items Bought
1	$\{a, b, d, e\}$
2	$\{b, c, d\}$
3	$\{a, b, d, e\}$
4	$\{a, c, d, e\}$
5	$\{b, c, d, e\}$
6	$\{b, d, e\}$
7	$\{c, d\}$
8	$\{a, b, c\}$
9	$\{a, d, e\}$
10	$\{b, d\}$



Item	Support count	→ After sorting	
a	5	d	9
b	~~6~~ 7	b	7
c	5	c	6
d	9	a	5
e	6	e	5

T1 a,b,d,e → d,b,e,a

T2 b,c,d → d,b,c

T3 a,b,d,e → d,b

T4 a,c,d,e → d,e,a,c

T5 b,c,d,e → d,b,e,c

T6 b,d,e

T7 c,d → d,c

T8 (a,b,c) → b,a,c

T9 a,d,e

T10 b,d → d,b

Scheme: Each item with correct support count in FP tree carries 1 mark
5 x 1 = 5 marks

3. b	Suppose that the data mining task is to cluster points (with (x, y) representing location) into three clusters, where the points are A1(2, 10), A2(2, 5), A3(8, 4), B1(5, 8), B2(7, 5), B3(6, 4), C1(1, 2), C2(4, 9).	10	5	3	1,2, 3/2

The distance function is Manhattan distance. Suppose initially we assign A1, B1, and C1 as the center of each cluster, respectively. Use the Manhattan algorithm to show only

(a) the three cluster centers after the first round of execution.

Answer:
After the first round, the three new clusters are: (1) $\{A_1\}$, (2) $\{B_1,A_3,B_2,B_3,C_2\}$, (3) $\{C_1,A_2\}$, and their centers are (1) (2, 10), (2) (6, 6), (3) (1.5, 3.5).

Initial	M1(2,10)	M2(5,8)	M3(1,2)
A1(2, 10)	0	5	9
A2(2, 5)	5	6	4
A3(8, 4)	12	7	9
B1(5, 8)	5	0	10
B2(7, 5)	10	5	10
B3(6, 4)	10	5	7
C1(1, 2)	9	10	0
C2(4, 9)	3	2	10
	A1	B1,A3,B2,B3,C2	C1,A2

(b) the final three clusters once the algorithm converged

Answer:
The final three clusters are: (1) $\{A_1,C_2,B_1\}$, (2) $\{A_3,B_2,B_3\}$, (3) $\{C_1,A_2\}$.

Iteration1	M1(2,10)	M2(6,6)	M3(1.5,3.5)
A1(2, 10)	0	8	7
A2(2, 5)	5	5	2
A3(8, 4)	12	4	7
B1(5, 8)	5	3	8
B2(7, 5)	10	2	7
B3(6, 4)	10	2	5
C1(1, 2)	9	9	2
C2(4, 9)	3	5	8
	A1,C2	A3,B1,B2,B3	A2,C1
Iteration2	M1(3,9.5)	M2(6.5,5.25)	1.5,3.5
A1(2, 10)	1.5	9.25	7
A2(2, 5)	5.5	4.75	2

A3(8, 4)	10.5	2.74	7
B1(5, 8)	3.5	5.25	8
B2(7, 5)	8.5	0.75	7
B3(6, 4)	8.5	1.75	5
C1(1, 2)	9.5	8.75	2
C2(4, 9)	1.5	10.25	8
	A1,B1,C2	A3,B2,B3	A2,C1
Iteration3	M1(3.67,9)	M2(7,4.33)	1.5,3.5
A1(2, 10)	2.67	10.67	7
A2(2, 5)	5.67	5.67	2
A3(8, 4)	9.33	1.33	7
B1(5, 8)	2.33	5.67	8
B2(7, 5)	7.33	0.67	7
B3(6, 4)	7.33	1.33	5
C1(1, 2)	7.33	8.33	2
C2(4, 9)	0.33	7.67	8

No change in the data points in each cluster, hence algorithm converged.
Scheme:

(a) 2.5 Marks
(b) Iteration 1 – 3 , 3 x 2.5 = 7.5 marks

Faculty Signature	**Course Co-Ordinator/Mentor Signature**	**HoD Signature**