

Final Exam

Data Mining: Techniques and Applications

Exercise 1 (7 pts):

Use the dissimilarity matrix in the following table to perform single link hierarchical clustering. Show your results by drawing a dendrogram. The dendrogram should clearly show the order in which the points are merged.

	p1	p2	p3	p4	p5
p1	0.00				
p2	0.10	0.00			
p3	0.41	0.64	0.00		
p4	0.55	0.47	0.44	0.00	
p5	0.35	0.98	0.85	0.76	0.00

Exercise 2 (10 pts):

Consider the dataset shown in the following table.

(a) Compute the support for itemsets $\{e\}$, $\{b, d\}$, and $\{b, d, e\}$ by treating each transaction ID as a market basket.

(b) Use the results in question (a) to compute the confidence for the association rules $\{b, d\} \rightarrow \{e\}$ and $\{e\} \rightarrow \{b, d\}$.

(c) Repeat question (a) by treating each customer ID as a market basket. Each item should be treated as a binary variable (1 if an item appears in at least one transaction bought by the customer, and 0 otherwise.)

Customer ID	Transaction ID	Items Bought
1	0001	$\{a, d, e\}$
1	0024	$\{a, b, c, e\}$
2	0012	$\{a, b, d, e\}$
2	0031	$\{a, c, d, e\}$
3	0015	$\{b, c, e\}$
3	0022	$\{b, d, e\}$
4	0029	$\{c, d\}$
4	0040	$\{a, b, c\}$
5	0033	$\{a, d, e\}$
5	0038	$\{a, b, e\}$

(d) Use the results in question (c) to compute the confidence for the association rules $\{b, d\} \rightarrow \{e\}$ and $\{e\} \rightarrow \{b, d\}$.

(e) Suppose s_1 and c_1 are the support and confidence values of an association rule r when treating each transaction ID as a market basket. Also, let s_2 and c_2 be the support and confidence values of r when treating each customer ID as a market basket. Discuss whether there are any relationships between (s_1, c_1) and (s_2, c_2) .

Exercise 3 (3 pts):

You are given a dataset with 100 records and are asked to cluster the data. You use K-means to cluster the data, but for all values of K (Number of clusters), $1 \leq K \leq 100$, the K-means algorithm returns only one non-empty cluster. How is this possible? How would another method, like hierarchical clustering, handle such data?

Dr. Tahar Mehenni

Solution of Final Exam

Data Mining: Techniques and Applications

Exercise 1 (7 pts):

	p1	p2	p3	p4	p5
p1	0.00				
p2	0.10	0.00			
p3	0.41	0.64	0.00		
p4	0.55	0.47	0.44	0.00	
p5	0.35	0.98	0.85	0.76	0.00

merge (p1, p2)
(1.0 pt)

	p1p2	p3	p4	p5
p1p2	0.00			
p3	0.41	0.00		
p4	0.47	0.44	0.00	
p5	0.35	0.85	0.76	0.00

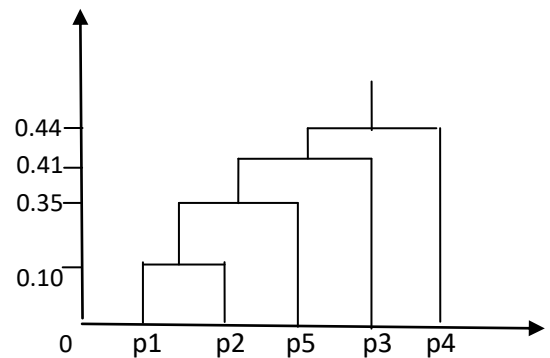
merge (p5, p1p2)
(1.5 pts)

	p5p1p2	p3	p4
p5p1p2	0.00		
p3	0.41	0.00	
p4	0.47	0.44	0.00

merge (p3, p5p1p2)
(1.5 pts)

	p4	p3p5p1p2
p4	0	
p3p5p1p2	0.44	0

merge (p4, p3p5p1p2)
(1.5 pts)



Dendrogram (1.5 pts)

Exercise 2 (10 pts):

(a) Compute the support for itemsets $\{e\}$, $\{b, d\}$, and $\{b, d, e\}$ by treating each transaction ID as a market basket.

Answer:

$$s(\{e\}) = 8/10 = 0.8 \quad (0.5 \text{ pt})$$

$$s(\{b, d\}) = 2/10 = 0.2 \quad (0.5 \text{ pt})$$

$$s(\{b, d, e\}) = 2/10 = 0.2 \quad (0.5 \text{ pt})$$

(b) Use the results in part (a) to compute the confidence for the association rules $\{b, d\} \rightarrow \{e\}$ and $\{e\} \rightarrow \{b, d\}$.

Answer:

$$c(bd \rightarrow e) = 0.2/0.2 = 100\% \quad (1.5 \text{ pts})$$

$$c(e \rightarrow bd) = 0.2/0.8 = 25\% \quad (1.5 \text{ pts})$$

(c) Repeat part (a) by treating each customer ID as a market basket. Each item should be treated as a binary variable (1 if an item appears in at least one transaction bought by the customer, and 0 otherwise.)

Answer:

$$s(\{e\}) = 4/5 = 0.8 \quad (0.5 \text{ pt})$$

$$s(\{b, d\}) = 5/5 = 1 \quad (0.5 \text{ pt})$$

$$s(\{b, d, e\}) = 4/5 = 0.8 \quad (0.5 \text{ pt})$$

(d) Use the results in part (c) to compute the confidence for the association rules $\{b, d\} \rightarrow \{e\}$ and $\{e\} \rightarrow \{b, d\}$.

Answer:

$$c(bd \rightarrow e) = 0.8/1 = 80\% \quad (1.5 \text{ pts})$$

$$c(e \rightarrow bd) = 0.8/0.8 = 100\% \quad (1.5 \text{ pts})$$

(e) Suppose s_1 and c_1 are the support and confidence values of an association rule r when treating each transaction ID as a market basket. Also, let s_2 and c_2 be the support and confidence values of r when treating each customer ID as a market basket. Discuss whether there are any relationships between s_1 and s_2 or c_1 and c_2 .

Answer:

There are no apparent relationships between s_1 , s_2 , c_1 , and c_2 . (1 pt)

Exercise 3 (3 pts):

(a) The data consists completely of duplicates of one object. (1.5 pts)

(b) Hierarchical clustering would depend on the ordering of the points and the exact algorithm using single, median or complete link. However, if the dendrogram were plotted showing the proximity at which each object is merged, then it would be obvious that the data consisted of duplicates. (1.5 pts)

Dr. Tahar Mehenni