DAMI – mock TEST I

1. Let $\{\{ab\}, \{ac\}, \{ae\}, \{bc\}, \{ce\}\}$ be all frequent 2-itemsets. Which of the following statement(s) is(are) correct?

   a) $\{e\}$ is certainly frequent b) perhaps $\{e\}$ is frequent, but this is not certain c) $\{e\}$ is infrequent

   d) $\{bce\}$ is certainly frequent e) perhaps $\{bce\}$ is frequent, but this is not certain f) $\{bce\}$ is infrequent

2. Let us consider the operation of intersecting transaction identifiers' lists (used in the Eclat algorithm). Let tidlist $t(X) = \{1,2,3,4,5\}$ and tidlist $t(Y) = \{4,5,6,7,8,9\}$.

   - Calculate the support $sup(X)$? $sup(X) = 5$
   - Calculate the tidlist $t(XY)$? $t(XY) = \{4,5\}$

3. Let node Y be a right hand side brother of node X in the tree created by the *dEclat* algorithm, differential list $d(X) = \{1,2,3,4,5\}$ and differential list $d(Y) = \{4,5,6,7,8,9\}$.

   - Calculate differential list $d(XY)$?: $d(XY) = \{6,7,8,9\}$
   - Calculate the support $sup(XY)$ provided $sup(X) = 20$? $sup(XY) = 16$

TId	***Items***	$\{e\bar{f}\}$	$\{e\bar{f}\bar{h}\}$
1	$\{abce\}$	√	√
2	$\{abdeh\}$	√	
3	$\{abefh\}$		
4	$\{bcefh\}$		
5	$\{acde\}$	√	√
6	$\{abcdefh\}$		
7	$\{aefh\}$		
8	$\{bcefh\}$		

Fig. 1. Transaction dataset

4. Consider the transaction dataset in Fig. 1. What is the support and confidence of rule with negation $\{e\bar{f}\}\rightarrow\{\bar{h}\}$: $sup(\{e\bar{f}\}\rightarrow\{\bar{h}\}) = 2$, $conf(\{e\bar{f}\}\rightarrow\{\bar{h}\}) = 2/3$.

5. Consider the transaction dataset in Fig. 1. Which itemsets are closures of itemset $\{ach\}$? $\gamma(\{ach\}) = \{abcdefh\}$

6. Consider the transaction dataset in Fig. 1. Which itemsets are generators of itemset $\{ach\}$? $G(\{ach\}) = \{\{ach\}\}$

7. Using the closed itemsets' representation (CR) in Fig. 2, determine whether $\{ach\}$ is frequent. If $\{ach\}$ is frequent, determine its support. If $\{ach\}$ is not frequent, provide the greatest possible value of its support. Frequent, sup = 3.

8. Using the generators' representation (GR) in Fig. 3, determine whether $\{ach\}$ is frequent. If $\{ach\}$ is frequent, determine its support. If $\{ach\}$ is not frequent, provide the greatest possible value of its support. Frequent, sup = 2.

$abcde_2$ $acdeh_3$

$abde_4$ $acde_4$

ade_5 bce_4

be_5 ce_5 de_6

e_7

Fig. 2. CR: Frequent closed itemsets *FC*

I:*abcdefgh*

Negative border *Bd*: *e* *bh*

Frequent generators:

abc_3 bcd_3

ab_4 ac_4 cd_4 bc_4 bd_4

b_5 a_5 c_5 d_5 h_2

$\varnothing_6$

Fig. 3. GR: Frequent generators *FG* & negative border *Bd*

9. Let us consider rules $\varnothing \to \{ach\}$ and $\{c\} \to \{ah\}$.

- Calculate the number of association rules that are covered by the rule $\varnothing \to \{ach\}$? $|C(\varnothing \to \{ach\})| = 3^3 - 2^3$.
- What is the cover of rule $\{c\} \to \{ah\}$: $C(\{c\} \to \{ah\})$ = $\{\{c\} \to \{ah\}, \{ac\} \to \{h\}, \{ch\} \to \{a\}, \{c\} \to \{a\}, \{c\} \to \{h\}\}$

10. Let us assume that the association rules in Fig. 4 are representative rules RR:

rule identifier	rule	support	confidence
1	$\varnothing \to \{abe\}$ [4,4/5]	4	4/5
2	$\varnothing \to \{bcde\}$ [4,4/5]	4	4/5
3	$\{a\} \to \{bcde\}$ [3,3/4]	3	¾
4	$\{c\} \to \{abde\}$ [3,3/4]	3	¾
5	$\{d\} \to \{abce\}$ [3,3/4]	3	¾

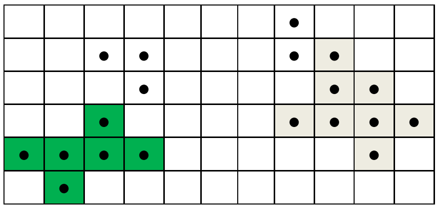
Fig. 4. A set of association rules

- Which of these rules cover rule $\{ae\} \to \{b\}$? #1, #3
- Estimate support and confidence of rule $\{ae\} \to \{b\}$ based on the supports and confidences of the covering representative rules. [≥ 4, ≥ 4/5]

11. Let us assume that the association rules in Fig. 4 are minimal non-redundant rules MNR:

- Which of these rules cover rule $\{ae\} \to \{b\}$? #1, #3
- Determine support and confidence of rule $\{ae\} \to \{b\}$ based on the supports and confidences of the covering minimal non-redundant rules. [4, 4/5]

12. **Related to the DBSCAN algorithm:** Let *minPts* = 4, distance parameter ε = 1. Draw the clusters and noise that would be found by the DBSCAN algorithm provided Euclidean distance is applied (that is, $d(P_1,P_2) = [|x_1-x_2|^2 + |y_1-y_2|^2]^{½}$).



Cluster 1 – the set of points in the green subspace.

Cluster 2 – the set of points in the grey subspace.

Noise – the remaining points.

13. **Related to using the triangle inequality for determining *ε*-neighborhood:** Let D be a set of two dimensional points as shown in Table 3, for which their Euclidean distance to a reference point *r* was calculated.

Table 3. Ordered set of two dimensional points D with their Euclidean distance to reference point r

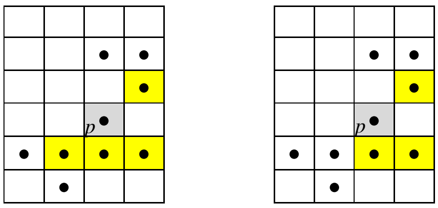
point q	X	Y	distance(q,r)
K	0.9	0.0	0.9
L	1.0	1.5	1.9
G	0.0	2.4	2.4
H	2.4	2.0	3.1
F	1.1	3.0	3.2
C	2.8	3.5	5.0
A	**4.2**	**4.0**	**5.8**
B	5.9	3.9	6.1

Let us also assume that the points in D are sorted with respect to their Euclidean distance to point *r*. Let ε = 1.0 and A be a point for which ε-neighborhood is to be determined by means of the triangle inequality property.

a) For which points in D different from point A a pessimistic estimation of their Euclidean distances to point A would be calculated? B, C, F

b) For which points in D different from point A their real Euclidean distances to point A would be calculated? B, C

14. **Related to the NBC algorithm (4):** Let k = 2.



a) Determine the k^+-neighbourhood of point p in the left-hand side figure.
b) Determine the reversed k^+-neighbourhood of point p in the right-hand side figure.
c) Calculate the neighbourhood density factor of point p: NDF(p) = ¾.
d) Does p play a role of a core point in the NBC algorithm? No.