**Quiz 1 – Spring 2018**

**CS583: Data Mining and Text Mining**

Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ UID\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

This is a closed-book quiz.

|  |  |
| --- | --- |
|  | **Marks** |
| Q1 |  |
| Q2 |  |
| **Total** |  |

1. (10 marks) Circle all correct answers for the following questions.

(a). Association rule mining requires

1. Numeric data
2. Sparse data (1.5)
3. Dense data
4. Transaction data (0.5)

(b). Given a dataset, a minsup and a minconf

1. Association rule mining is complete.
2. Association rule mining has a unique solution.
3. Association mining can be done efficiently.
4. Association rule mining is often done with data in memory.
5. For class association rule mining,
6. Items that can appear on the right hand side of rules can also appear on the left hand side.
7. Items that can appear on the right hand side of rules cannot appear on the left hand side.
8. We can control whether we want rules of a particular class.
9. We cannot use its resulting rules for prediction.
10. For mining association rules with multiple minimum supports,
11. We can set a minimum support for every item.
12. The complexity is lower than the general association rule mining.
13. We do tail counting for efficiency
14. We do tail counting for rule generation.
15. For sequential pattern mining,
16. <{2, 3}{3, 4, 5, 6}> is a supersequence of <{2}{3}>
17. <{2, 3}{3, 4, 5, 6}> is a supersequence of <{4}{5}>
18. Pruning of a candidate k-sequence is done based on all its subsequences
19. Generation of candidate k-sequences need frequent (k-1)-sequences.

2. (20 marks) Answer the following questions.

(a). (5 marks) Let F3 = {{1, 2, 3}, {1, 3, 4}, {1, 2, 5}, {1, 3, 5} {1, 3, 7}, {1, 4, 7}, {3, 4, 7}}. Compute the candidate 4-itemsets after the join step and after the pruning step separately based on the Apriori candidate generation algorithm.

Join: {1, 2, 3, 5}, {1, 3, 4, 5}, {1, 3, 4, 7}, {1, 3, 5, 7}

Prune: {1, 3, 4, 7}

(b). (2 marks) Given the user-specified MIS values for the three items 1, 2, and 3:

MIS(1) = 2%, MIS(2) = 0.1%, MIS(3) = 0.3%

does each of the following rules satisfy its minsup?

1 → 2 [sup=0.15%, conf =70%] Yes

1 → 3 [sup=0.15%, conf =70%] No

(c). (2 marks) Assume we have *k* items in *I* (the set of all items). What is the complexity of association rule mining?

O(2k)

(d). (3 marks) Suppose {2, 3, 4} is a frequent itemset with sup=30%. Itemsets {2, 3}, {2, 4}, {3, 4}, {2}, {3}, and {4} have sup=50%, 50%, 75%, 75%, 75%, 75% respectively. What is the support and conference of each rule?

2, 3 → 4 support = 30% confidence = 30/50

3, 4 → 2 support = 30% confidence = 30/75

(e). (3 marks) Given the following documents with attached classes:

doc 1: Student, Teach, School : Education

doc 2: Student, School, Team : Education

doc 3: Teach, School, City, Game : Education

doc 4: Baseball, Student, Basketball, School : Sport

doc 5: Basketball, Player, Spectator : Sport

doc 6: Baseball, Coach, Game, Team : Sport

What is the support and conference of each rule?

Game → Sport support = 1/6 confidence = 1/2

School, Student → Sport support = 1/6 confidence = 1/3

(f). (5 marks) Fill the Table below with candidate sequences based on the GSP algorithm.

|  |  |  |
| --- | --- | --- |
| **Frequent**  **3-sequences** | **Candidate 4-sequences** | |
| **after joining** | **after pruning** |
| 〈{a, b} {c}〉 | <{a, b}{c, d}> | <{a, b}{c, d}> |
| 〈{a, b} {d}〉 | <{a, b}{c} {e}> |  |
| 〈{a} {c, d}〉 |  |  |
| 〈{a, c} {e}〉 |  |  |
| 〈{b} {c, d}〉 |  |  |
| 〈{b} {c} {e}〉 |  |  |