Northeastern University, Khoury College of Computer Science

# CS 6220 Data Mining — Assignment 4

**Due: March 1, 2023 (100 points)**

**Hamidreza Komaki**
**GitHub:** Hamidreza-Komaki
**Email:** Komaki.H@Northeastern.edu

**Question 1a.)** Done in colab file.
**Question 1b.)** Done in colab file.
**Question 1c.)** Done in colab file.
**Question 2a.)** Done in colab file.
**Question 2b.)** Done in colab file.
**Question 2c.)** Done in colab file.
**Question 2d.)** Done in colab file.

# Market Basket Analysis and Algorithms

Consider $F_3$ as the following set of frequent 3-itemsets:

```
{1, 2, 3}, {1, 2, 4}, {1, 2, 5}, {1, 3, 4},
{2, 3, 4}, {2, 3, 5}, {3, 4, 5}.
```

Assume that there are only five items in the data set.

## Question 3 [25 pts total]

**[10 pts] Question 3a.)** {1, 2, 3, 4}, {1, 2, 3, 5}, {1, 2, 4, 5}, {1, 3, 4, 5}, {2, 3, 4, 5}

**[10 pts] Question 3b.)** {1, 2, 3, 4}, {1, 2, 3, 5}, {1, 2, 4, 5}, {2, 3, 4, 5}

**[5 pts] Question 3c.)** {1, 2, 3, 4}

## Question 4 [25 pts total]

**[3 pts] Question 4a.)** The formula for determining the maximum number of association rules that can be derived is $3^d - 2^{d+1} + 1$, where d denotes the number of items in the set. For this specific set of items with d = 7 (Beer, Diapers, Milk, Bread, Butter, Cookies, Eggs), the maximum number of association rules that can be is calculated as $3^7 - 2^{7+1} + 1 = 1932$.

**[3 pts] Question 4b.)** To calculate the confidence of the rule Milk, Diapers, Butter, we need to find the number of transactions that contain all three items in the antecedent and the consequent, and divide it by the number of transactions that contain the antecedent items. In this case, we have:
The number of transactions containing Milk, Diapers, Butter: 2 (transaction 2 and transaction 7) The number of transactions containing Milk, Diapers: 4 (transactions 2, 3, 5, and 7) Therefore, the confidence of the rule Milk, Diapers, Butter is $2/4 = 0.5$.

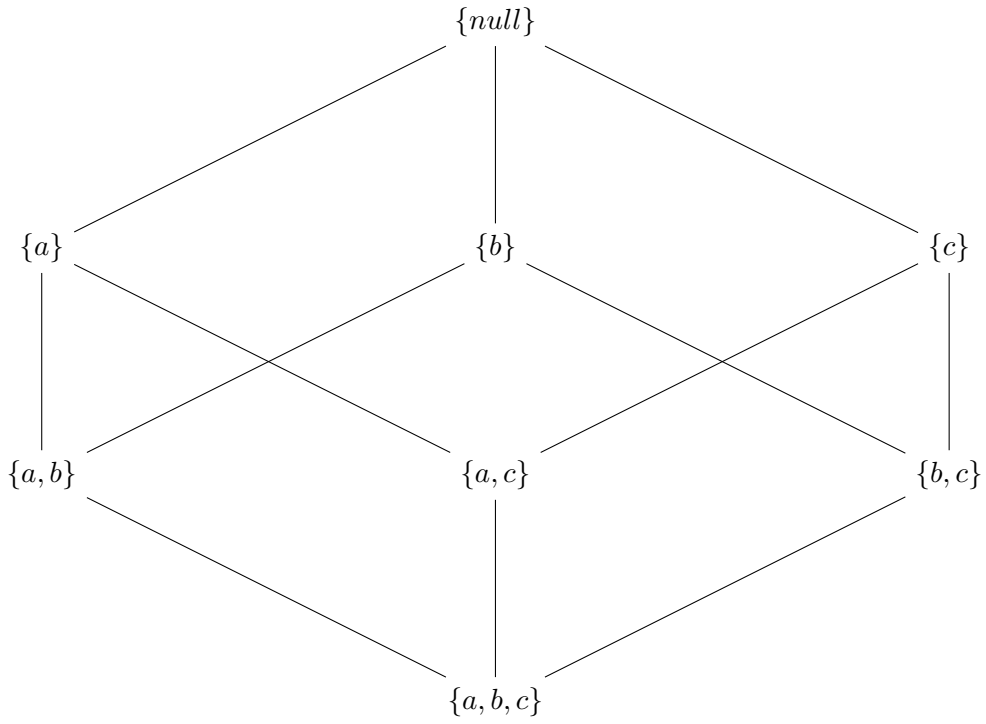
**[3 pts] Question 4c.)** To calculate the support for the rule Milk, Diapers, Butter, we need to find the number of transactions that contain all three items in the antecedent and the consequent, and divide it by the total number of transactions. In this case, we have:
The number of transactions containing Milk, Diapers, Butter: 2 (transaction 2 and transaction 7) The total number of transactions: 10 Therefore, the support for the rule Milk, Diapers, Butter is $2/10 = 0.2$.

**[3 pts] Question 4d.)** `True.` If an itemset is frequent, then all its subsets are also frequent. Therefore, {a,b} is always a frequent itemset if {a,b,c,d} is frequent.

**[3 pts] Question 4e.)** `False.` It is possible for {a,b,c} to be infrequent even if {a,b}, {b,c}, and {a,c} are frequent. This is because {a,b,c} may not appear together frequently.

**[3 pts] Question 4f.)** `False.` We cannot determine the support of {b} from the support of a,b and {b,c}. It is possible that {b} appears in all sets that contain {a,b}, in all sets that contain {b,c}, or in some but not all sets that contain either {a,b} or {b,c}. Therefore, we cannot make a definitive statement about the support of {b}.
**[3 pts] Question 4g.)** `False.` For a dataset with 5 items, there are 5 choose 2 = 10 unique pairs of items. Therefore, the maximum number of size-2 frequent itemsets that can be extracted is 10.

**[4 pts] Question 4h.)** Draw the itemset lattice for the set of unique items $\mathcal{I} = \{a, b, c\}$.


{null}
{a}
{b}
{c}
{a, b}
{a, c}
{b, c}
{a, b, c}