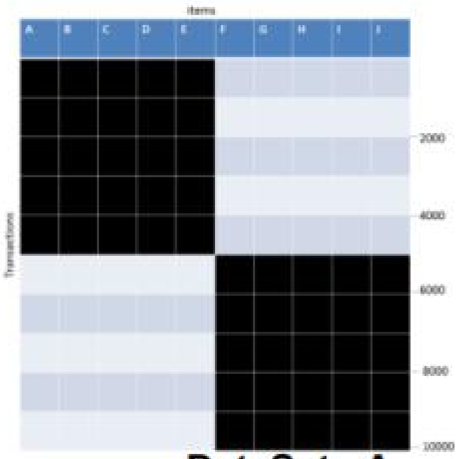
CS 422 Due:
Data Mining Homework 3.0 04/22/2025

Practicum Problems

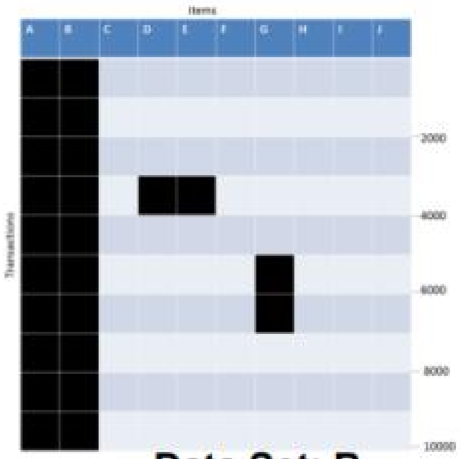
These problems will primarily reference the lecture materials, class examples, the prescribed textbook, or the instructor manual (all available on Canvas). For this assignment, you are required to type your responses and submit them as a single PDF document via Canvas. Students are encouraged to refer to the textbook or credible online resources to answer the questions accurately.

Problem 1

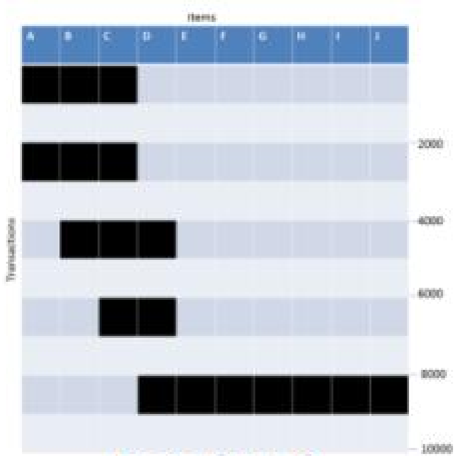
Given the following transaction data sets (dark cells indicate presence of an item in a transaction) and a support threshold of 20%, answer the following questions:



DataSet: A



Data Set: B



Data Set: C

a. What is the number of frequent itemsets for each dataset? Which dataset will produce the most number of frequent itemsets?

b. Which dataset will produce the longest frequent itemset?

c. Which dataset will produce frequent itemsets with highest maximum support?

d. Which dataset will produce frequent itemsets containing items with widely varying support levels (i.e., itemsets containing

items with mixed support, ranging from 20% to more than 70%)?

e. What is the number of maximal frequent itemsets for each dataset? Which dataset will produce the most number of maximal frequent itemsets?

f. What is the number of closed frequent itemsets for each

dataset? Which dataset will produce the most number of closed frequent itemsets?

Problem1：

a. The transaction dataset size is 10,000, and the support threshold is 20%. Therefore, Only those project sets with support degree greater than or equal to 2000 are considered as frequent project sets.

Data Set A：

The number of frequent project sets in the single-item project concentration is 10, in the two-item project concentration it is 20, in the three-item project concentration it is 20, in the four-item project concentration it is 10, and in the five-item project concentration it is 2.So the total number of frequent item sets is 62.

Data Set B：

The number of frequent single-item sets is 3, that of double-item sets is 3, and that of triple-item sets is 1.Therefore, the number of frequent item sets is 7.

Data Set C：

The number of frequent single-item sets is 4, that of double-item sets is 4, and that of triple-item sets is 1.Therefore, the number of frequent item sets is 9.Therefore, the number of frequent data sets in dataset A is the largest.

b.Dataset A will generate the longest frequent itemsets.

c.Dataset B will generate the frequent dataset with the highest maximum support degree.

d.Dataset C will generate frequent itemsets, among which contain items with a wide range of support levels.

e.The number of the maximum frequent itemsets in Dataset A is 2:{A、B、C、D、E}，{E、F、G、H、I、J}, in Dataset B it is 1:{A、B、G}, and in Dataset C it is 2:{A、B、C}，{C、D}. Therefore, dataset A and dataset C will generate the largest number of maximal frequent itemsets.

f.The number of closed itemsets in Dataset A is 2：{A、B、C、D、E}，{E、F、G、H、I、J}, while that in Dataset B is also 2：{A、B、G}，{A、B}. Dataset C has 4 closed itemsets：. {A、B、C}，{B、C}，{C}，{C、D}.Therefore, Dataset C will generate the most closed frequent itemsets.

Problem 2:

Consider the following set of candidate 3-itemsets:

{1, 2, 4}, {1, 3, 5}, {1, 4, 6}, {2, 3, 5}, {2, 5, 6}, {3, 4, 5}, {3, 5, 6}, {2, 4, 6}

a. Construct a hash tree for the above candidate 3-itemsets. Assume the tree uses a hash function where all odd-numbered items are hashed to the left child of a
node, while the even-numbered items are hashed to the right child. A candidate k-itemset is inserted into the tree by hashing on each successive item in the
candidate and then following the appropriate branch of the tree according to the hash value. Once a leaf node is reached, the candidate is inserted based on one of the following conditions:

i. Condition 1: If the depth of the leaf node is equal to k (the root is assumed to be at depth 0), then the candidate is inserted regardless of the number of itemsets already stored at the node.

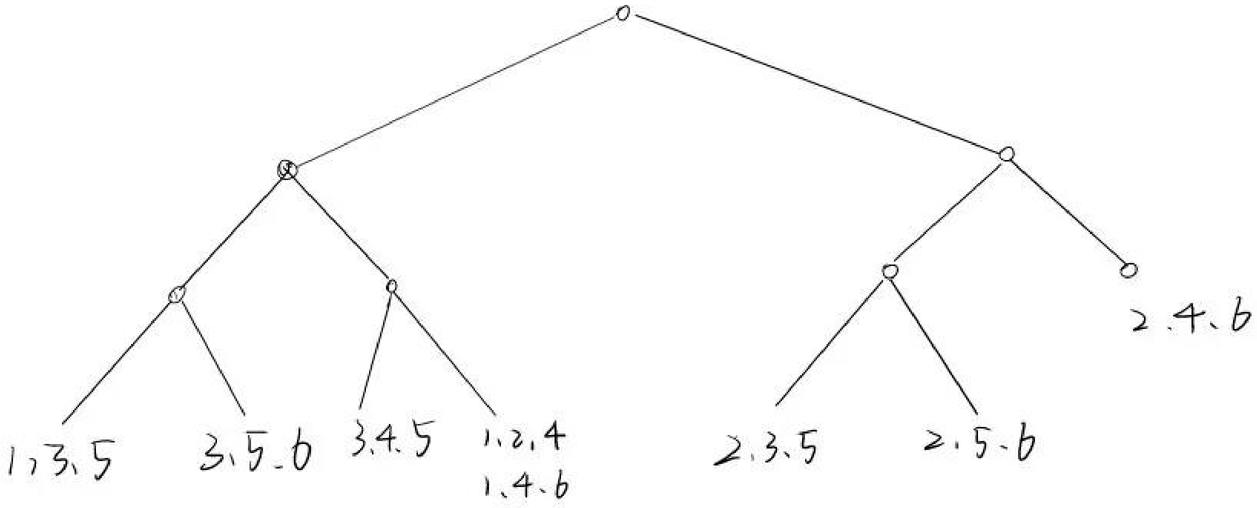
ii. Condition 2: If the depth of the leaf node is less than k, then the candidate can be inserted as long as the number of itemsets stored at the node is less than maxsize. Assume maxsize=2 for this question.

iii. Condition 3: If the depth of the leaf node is less than k and the number of itemsets stored at the node is equal to maxsize, then the leaf node is converted into an internal node. New leaf nodes are created as children of the old leaf node. Candidate itemsets previously stored in the old leaf node are distributed to the children based on their hash values. The new candidate is also hashed to its appropriate leaf node.

b. How many leaf nodes are there in the candidate hash tree? How many internal nodes are there?

c. Consider a transaction that contains the following items: {1, 2, 3, 5, 6}. Using the hash tree constructed in part (a), which leaf nodes will be checked against the transaction? What are the candidate 3-itemsets contained in the transaction?

a.



b.The number of leaf nodes is 7 and the number of internal nodes is 6.

c.Finally, the leaf nodes containing {1, 3, 5}, {2, 3, 5}, {2, 5, 6}, and {3, 5, 6} will be inspected. Finally, the leaf nodes containing {1, 3, 5}, {2, 3, 5}, {3, 5, 6}, and {2, 5, 6} will also be inspected.

Problem 3

The Apriori algorithm uses a generate-and-count strategy for deriving frequent itemsets. Candidate itemsets of size k+1 are created by joining a pair of frequent itemsets of size k (this is known as the candidate generation step). A candidate is discarded if any one of its subsets is found to be infrequent during the candidate pruning step. Suppose the Apriori algorithm is applied to the data set shown in Table 2.0 with minsup=30% , i.e., any itemset occurring in less than 3 transactions is considered to be infrequent.

Table: 2.0

Transaction ID	Items Bought
T1	a, b, x, y
T2	b, x, y
T3	a, y, z
T4	a, b, x, z
T5	x, y
T6	b, z
T7	a, x, y, z
T8	a, b
T9	b, y, z
T10	a, b, x, y

a. Draw an itemset lattice representing the data set given in Table 2.0 . Label each node in the lattice with the following letter(s):

i. N: If the itemset is not considered to be a candidate itemset by the Apriori algorithm. There are two reasons for an itemset not to be considered as a candidate itemset: (1) it is not generated at all during the candidate generation step, or (2) it is generated during the candidate generation step but is subsequently removed during the candidate pruning step because one of its subsets is found to be infrequent.

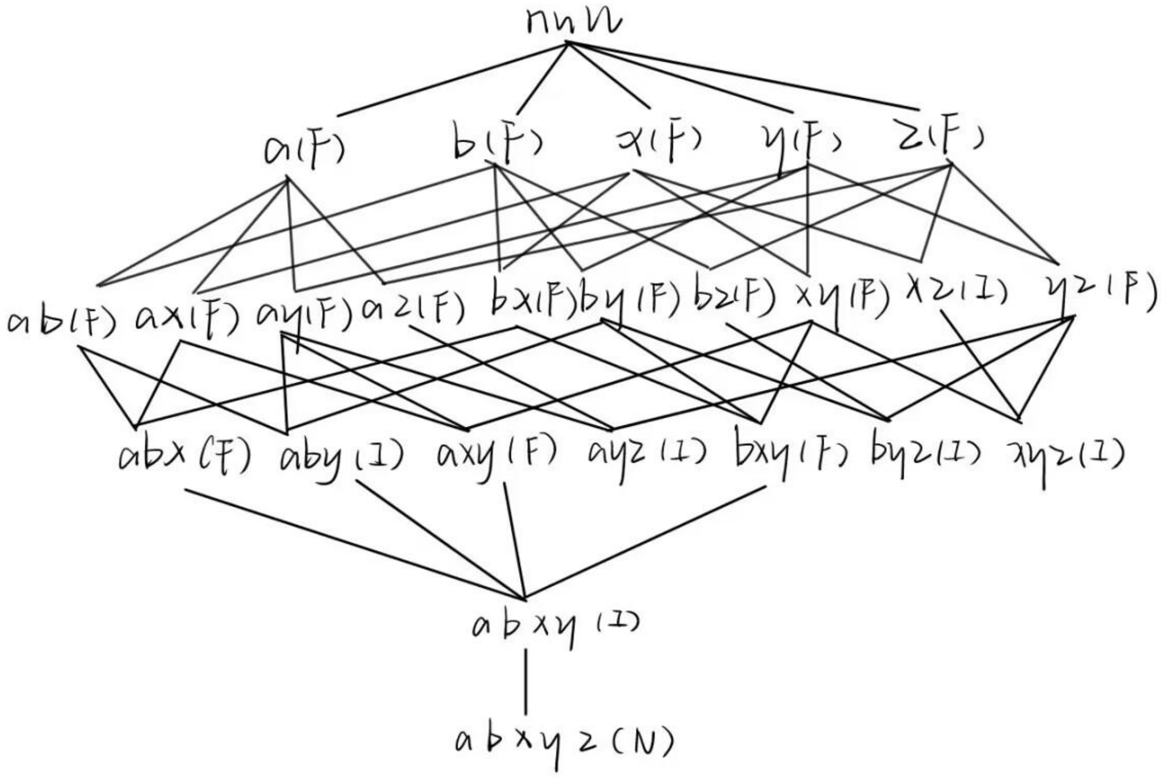
ii. F: If the candidate itemset is found to be frequent by the Apriori algorithm. iii. I: If the candidate itemset is found to be infrequent after support counting.

b. What is the percentage of frequent itemsets (with respect to all itemsets in the lattice)?

c. What is the pruning ratio of the Apriori algorithm on this data set? (Pruning ratio is defined as the percentage of itemsets not considered to be a candidate because (1) they are not generated during candidate generation or (2) they are pruned during the candidate pruning step.)

d. What is the false alarm rate (i.e., percentage of candidate itemsets that are found to be infrequent after performing support counting)?

a.



b.The number of candidate itemsets is C(5,1) + C(5,2) + C(5,3) + C(5,4) + C(5,5) = 5 + 10 + 10 + 5 + 1 = 31.

The number of frequent itemsets: 15

Therefore, the percentage of frequent itemsets: 48.4%

c.The number of itemsets marked as N: Through analysis, four-item and five-item itemsets will not be generated because their subsets are not frequent. There are a total of C(5,4) + C(5,5) = 5 + 1 = 6 such itemsets. During the generation process, the number of itemsets that were pruned because their subsets are not frequent is marked as N, and the number of itemsets removed in the pruning stage because their subsets are not frequent is 10. The pruning rate is 32.3%.

d.Calculate the number of non-frequent itemsets in the candidate itemsets: The number of itemsets marked as I is 7. Total number of candidate itemsets: All itemsets that are not N, that is, 31 - 10 = 21. Miss rate is 33.3%

E.N.D