Midterm exam

Question 1 (10 points):

Calculate city block, Euclidean and supremum distances for the below data.

(a). x = (1, -1, 10, 3, 4), y = (10, -1, 4, 5, 2)

(b). x1 = (5, 4), x2 = (-2, -3)

Question 2 (15 points):

Discuss the advantages and disadvantages of using sampling to reduce the number of data objects. Would simple random sampling (without replacement) be a good approach to sampling? Why or why not? What kind of sampling method that you would like to use?

Question 3 (10 points):

The Apriori algorithm uses a hash tree data structure to efficiently count the support of candidate itemsets. Consider theh hash tree for candidate 3-itemsets as show in the below figure.

Given a transaction that contains items {1,3,4,5,8}, which of the hash tree leaf nodes will be visited when finding the candidates of the transaction?

Diagram

Description automatically generated

Question 4 (15 points):

Consider the following set of candidate 3-itemsets:

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

Construct a hash tree for the above candidate 3-itemsets. Assume the tree uses a hash function where:

1. all odd-numbered items are hashed to the left child of a node

2. the even-numbered items are hashed to the right child.

Question 5 (30 points)

Given the lattice structure in the below picture and the transactions given in the below table, label each node with the following letter(s): I if it is infrequent; F if it is frequent; M if the node is a maximal frequent itemset. Assume that the support threshold (minimum support) is 30%.

Table

Description automatically generated

A picture containing text, tool, scissors

Description automatically generated

Question 6 (20 points)

Consider the dataset in the below table. A rule is considered to be strong if its support exceeds 15%. The dataset in the Table supports the following two strong rules:

(a).

(b).

Compute the support and confidence for both rules.

Table

Description automatically generated

Extra points (10 points)

Consider the following set of frequent 3-itemsets:

Assume that there are only 5 items in the dataset.

(a). List all candidate 4-itemsets obtained by the candidate generation procedure in Apriori.

(b). List all the candidate 4-itemsets that survive the candidate pruning step of the Apriori algorithm.

