HW1

Question 1:

For the following vectors, x and y, calculate the indicated similarity or distance measures.

1. x=(1,1,1,1), y = (2,2,2,2) Euclidean, Jaccard
2. x=(0,1,0,1), y = (1,0,1,0) Euclidean, Jaccard, city block (manhattan) distance
3. x=(2, -1, 0, 2, 0), y = (-1, 1, -1, 0, 0, -1) Euclidean, supremum distance
4. x=(2, -1, 0, 2, 0, -3), y = (-1, 1, -1, 0, 0, -1) Euclidean, supremum distance, city block

Question 2:

Many sciences rely on observation instead of designed experiments. Compare and discuss the data quality issues involved in observational science withthose of experimental science and data mining.

Question 3:

You are given a set of m objects that is divided into K groups, where the ith group is of size m\_i. If the goal is to obtain a sample of size n<m, what is the difference between the following two sampling schemes? (assume sampling with replacement.)

1. We randomly select nxm\_i/m elements from each group.
2. We randomly select n elements from the dataset, without regard for the group to which an object belongs.

Question 4:

This exercise compares some similarity and distance measures.

1. For binary data, compute the simple matching (SMC) and Jaccard coefficients

x = 0101010001

y = 0100011000

1. Suppose that you are comparing how similar two organisms of different species are in terms of the number of genes they share. Describe which measure, SMC or Jaccard, you think would be more appropriate for comparing the genetic makeup of two organisms. Explain.

(assume that each animal is represented as a binary vector, where each attribute is 1 if a particular gene is present in the organism and 0 otherwise).

1. If you wanted to compare the genetic makeup of two organisms of the same species, e.g., two human beings, would you use the SMC or Jaccard, or a different measure of similarity or distance? Example (Note that two human beings share > 99.9% of the same genes).

Question 5 Consider the dataset shown in below table.

1. Compute the support for itemsets {e}, {b, d}, and {b, d, e} by treating each customer ID as a market basket.
2. Use the results in part(a) to compute the confidence for the association rules {b, d} -> {e} and {e} -> {b, d}. Is confidence a symmetric measure?

|  |  |
| --- | --- |
| Customer ID | Items bought |
| 1 | {a, d, e} |
| 2 | {a, b, c, e} |
| 3 | {a, b, d, e} |
| 4 | {a, c, d, e} |
| 5 | {b, c, e} |
| 6 | {b, d, e} |
| 7 | {c, d} |
| 8 | {a, b, c} |
| 9 | {a, d, e} |
| 10 | {a, b, e} |

Question 6 Given the above table,

(a). what is the maximum number of association rules that can be extracted from this data (including rules that have zero support)?

(b). what is the maximum size of frequent itemsets that can be extracted (assuming minsup > 0)?

(c). find an itemset (of size 2 or larger) that has the largest support.

(d). Find a pair of items, j and j, such that the rules {i} -> {j} and {j} -> {i} have the same confidence.

Question 7. Suppose the Apriori algorithm is applied to the table table with minsup = 30%, i.e., any itemset occurring in less than 3 transactions is considered to be infrequent.

(a). Draw an itemset lattice representing the dataset given in the table, label each node in the lattice with the following letter(s):

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
2. It is generated during the candidate generation step but is subsequently removed during the candidate pruning step because one of tis subsets is found to be infrequent.

F: if the candidate itemset is found to be frequent by the Apriori algorithm

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)?

1. What is the pruning ratio of the Apriori algorithm on this dataset?

(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)