PSG College of Technology, Coimbatore -641 004

Department of Applied Mathematics and Computational Sciences

8th Semester MSc TCS

18XT87 Data Mining Lab

Problem Sheet - 2

1.

It is important to define or select similarity measures in data analysis. However, there is no commonly accepted subjective similarity measure. Results can vary depending on the similarity measures used. Nonetheless, seemingly different similarity measures may be equivalent after some transformation.

Suppose we have the following 2-D data set:

	A_1	A_2
$\boldsymbol{x_1}$	1.5	1.7
$\boldsymbol{x_2}$	2	1.9
$\boldsymbol{x_3}$	1.6	1.8
$\boldsymbol{x_4}$	1.2	1.5
$\boldsymbol{x_5}$	1.5	1.0

(a) Consider the data as 2-D data points. Given a new data point, $\boldsymbol{x} = (1.4, 1.6)$ as a query, rank the database points based on similarity with the query using Euclidean distance, Manhattan distance, supremum distance, and cosine similarity.

2. Consider the following Term-Frequency vector.

Document Vector or Term-Frequency Vector

Document	team	coach	hockey	baseball	soccer	penalty	score	win	loss	season
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	2	1	0	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

1) Find Cosine similarity matrix of the given documents.
2) Also, find dissimilarity matrix by using
    a. Euclidean distance
    b. Manhattan distance
    c. Minkowski distance (h=3)
    d. Supremum distance

So finally, you are getting one similarity matrix and 4 dissimilarity matrix for the given four documents.

3. **Dissimilarity between binary attributes:**

Find the distance between objects (patients) by considering only on asymmetric attributes. (Gender is symmetric attribute.)

Relational Table Where Patients Are Described by Binary Attributes

name	*gender*	*fever*	*cough*	*test-1*	*test-2*	*test-3*	*test-4*
Jack	M	Y	N	P	N	N	N
Jim	M	Y	Y	N	N	N	N
Mary	F	Y	N	P	N	P	N

For the given data, find d(Jack, Jim), d(Jack, Mary), d(Jim, Mary).

Also, ensure that your program should work correctly for any other possible size of data.