Lab 07 | Similarity measures

IT563 Data Mining, Winter'2019, DAIICT, Gandhinagar; pm_jat

Similarity measure is a core function that is used in many data mining tasks. Let this labe be like a tutorial, where you manually solve following problems.

1. Given $x = (15,20,9,14)$, and y $= (14,18,7,10)$, find Manhattan, Euclidean, L_3-norm, L_∞-norm, distances between x and y.

   What do you observe as degree of norm increases in *Minkowski* distance?

   Figure out one application example for each of this measure.

2. What is your interpretation of Hamming distance, Jaccard similarity, and Cosine similarity? Enumerate suitability and unsuitability of each of these measures.

   Compute these measure for following two binary vectors x, and y, find out Hamming distance, Jaccard similarity, and Cosine similarity.

   x: 0101010001
   y: 0100011000

   a. Suppose x and y are two items. Each dimension is a feature. The values 1/0 indicate that an item has or does not have that feature. Which measure you think is suitable for measuring the distance between items x and y? Justify your answer.

   b. Suppose x and y are two documents, and each dimension is a *word* (or term) count (vectors are no more binary in that case). Which measure you think is appropriate for measuring similarity between document x and y? Justify your answer.

3. Compute the Jaccard similarities of each pair of the following three sets: {Health, Patient, Politics, Shah}, {Health, Gujarat, Tiger, Politics}, and {Hardik, Patel, Politics}.

   Explain three applications (discussed in lecture) of Jaccard Similarity of Documents.

4. Below are given two student objects. Compute the similarity between Pankaj and Mitesh.

	Pankaj	Mitesh
Member Programming Club	Yes	Yes
Sportsman	Yes	No
Watches average number of movies in a month	5	10
Early Riser	Yes	No
Average Attendance in Lectures(%)	40	60
CPI	7.8	6.9