Distance Measures

Question 1:

Consider the following three vectors u, v, w in a 6-dimensional space:

u = [1, 0.25, 0, 0, 0.5, 0]
v = [0.75, 0, 0, 0.2, 0.4, 0]
w = [0, 0.1, 0.75, 0, 0, 1]

Suppose cos(x,y) denotes the similarity of vectors x and y under the cosine similarity measure. Compute all three pairwise similarities among u,v, w.

Given Vectors:

u = [1, 0.25, 0, 0, 0.5, 0]

v = [0.75, 0, 0, 0.2, 0.4, 0]

w = [0, 0.1, 0.75, 0, 0, 1]

cos(x, y) = $\frac{x.y}{|x||y|}$ = $\frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$

cos(u, v) = $\frac{u.v}{|u||v|}$

The equation is:

= $\frac{0.75+0+0+0+0.2+0}{\sqrt{1+0.0625+0+0+0.25+0}\sqrt{0.5625+0+0+0.04+0.16+0}}$

= **0.9503**

cos(u, w) = $\frac{u.w}{|u||w|}$

The equation is:

= $\frac{0+0.025+0+0+0+0}{\sqrt{1.3125}\sqrt{0+0.01+0.5625+0+0+1}}$

= 0.1742

cos(v, w) = $\frac{v.w}{|v||w|}$

The equation is:

= $\frac{0+0+0+0+0+0}{\sqrt{1.3125}\sqrt{1.5725}}$

= **0**

Question 2:

Here are five vectors in a 10-dimensional space:

1111000000 0100100101 0000011110 0111111111 1011111111

Compute the Jaccard distance (not Jaccard "measure") between each pair of the vectors.

Given Vectors:

V1 = 1111000000

V2 = 0100100101

V3 = 0000011110

V4 = 0111111111

V5 = 1011111111

Jaccard similarity (A, B) = $\frac{A \cap B}{A \cup B}$

Jaccard distance = 1 – Jaccard Similarity

Jaccard similarity (V1, V2) = $\frac{V1 \cap V2}{V1 \cup V2} = \frac{1}{7}$

Jaccard distance (V1, V2) = $1 - \frac{1}{7} = \frac{6}{7}$

Jaccard similarity (V1, V3) = $\frac{V1 \cap V3}{V1 \cup V3} = \frac{0}{7}$

Jaccard distance (V1, V3) = 1 – 0= 1

Jaccard similarity (V1, V4) = $\frac{V1 \cap V4}{V1 \cup V4} = \frac{3}{10}$

Jaccard distance (V1, V4) = $1 - \frac{3}{10} = \frac{7}{10}$

Jaccard similarity (V1, V5) = $\frac{V1 \cap V5}{V1 \cup V5} = \frac{3}{10}$

Jaccard distance (V1, V5) = $1 - \frac{3}{10} = \frac{7}{10}$

Jaccard similarity (V2, V3) = $\frac{V2 \cap V3}{V2 \cup V3} = \frac{1}{7}$

Jaccard distance (V2, V2) = $1 - \frac{1}{7} = \frac{6}{7}$

Question 3:

Here are five vectors in a 10-dimensional space:

1111000000 0100100101 0000011110 0111111111 1011111111

Compute the Manhattan distance (L_1 norm) between each two of these vectors.

Given Vectors:

V1 = 1111000000

V2 = 0100100101

V3 = 0000011110

V4 = 0111111111

V5 = 1011111111

We know that Manhattan distance is the sum of absolute difference of the components of the vectors.

Manhattan distance (x, y) = $|x_1 - y_1| + |x_2 - y_2| + \ldots + |x_n - y_n|$

Manhattan distance (V1, V2)

= |1-0|+|1-1|+|1-0|+|1-0|+|0-1|+|0-0|+|0-0|+|0-1|+|0-0|+|0-1| = 6

Manhattan distance (V1, V3) = 1+1+1+1+0+1+1+1+1+0 = 8

Manhattan distance (V1, V4) = 1+0+0+0+1+1+1+1+1+1 = 7

Manhattan distance (V1, V5) = 0+1+0+0+1+1+1+1+1+1 = 7

Manhattan distance (V2, V3) = 0+1+0+0+1+1+1+0+1+1 = 6

Manhattan distance (V2, V4) = 0+0+1+1+0+1+1+0+1+0 = 5

Manhattan distance (V2, V5) = 1+1+1+1+0+1+1+0+1+0 = 7

Manhattan distance (V3, V4) = 0+1+1+1+1+0+0+0+0+1 = 5

Manhattan distance (V3, V5) = 1+0+1+1+1+0+0+0+0+1 = 5

Manhattan distance (V4, V5) = 1+1+0+0+0+0+0+0+0+0 = 2

Question 4: The edit distance is the minimum number of character insertions and character deletions required to turn one string into another. Compute the edit distance between each pair of the strings **he**, **she**, **his**, and **hers**.

Given words are **he**, **she**, **his**, and **hers**

Edit distance between “he” and “she” = 1

‘s’ should be inserted in “he” in order to turn into “she”

Edit distance between “he” and “his” = 3

‘e’ should be deleted in “is” should be inserted.

Edit distance between “he” and “hers” = 2

‘r’ and ‘s’ should be inserted

Edit distance between “she” and “his” = 4

‘s’ and ‘e’ should be deleted and ‘i’ and ‘s’ should be inserted

Edit distance between “she” and “hers” = 3

‘s’ should be deleted. ‘r’ and ‘s’ should be inserted

Edit distance between “his” and “hers” = 3

‘i’ should be deleted and ‘e’, ‘r’ should be inserted.