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.

Sol:

Assignment -3 / Distance measures, Nearest-neighbor learning.

Q1: 3 vectors are:

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

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

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

Cosine similarity

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

$$= \frac{\Sigma u \cdot v}{\sqrt{\Sigma u^2} \times \sqrt{\Sigma v^2}}$$

$$\cos(u,v) = \frac{1 \times 0.75 + 0.25 \times 0 + 0 \times 0 + 0 \times 0.2 + 0.5 \times 0.4 + 0}{\sqrt{1^2 + 0.25^2 + 0^2 + 0^2 + 0.5^2 + 0} \cdot \sqrt{0.75^2 + 0^2 + 0^2 + 0.2^2 + 0.4^2}}$$

$$= \frac{0.75 + 0.2}{1.1456 \times 0.8732} = 0.95$$

$$\cos(u,w) = \frac{0.25 \times 0.1}{\sqrt{1.3125} \cdot \sqrt{1.5725}}$$

$$= \frac{0.025}{1.146 \cdot 1.254}$$

$$= \frac{0.025}{1.437} = 0.02$$

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.

Sol:

Q2: five vectors are

A = 1111000000

B = 0100100101

C = 0000011110

D = 0111111111

E = 1011111111

Jaccards distance = 1 - jaccards similarity.

(A) Between A and B

$$\text{jaccard sim} = \frac{|A \cap B|}{|A \cup B|} = 1/7$$

Jaccard distance = $1 - 1/7 = \frac{6}{7}$

(b) Between A and C

Jaccard sim = 0

Jaccard distance = 1 - 0 = 1

(c) Between A and D

Jaccard sim = 3/10

Jaccard dist = $1 - 3/10 = 7/10$

(d) A and E

(e) between B and C

Jaccard sim = 1/7

Jaccard dist = 1 - 1/7 = 6/7

(f) between B and D

Jaccard sim = 4/9

Jaccard distance = 1 - 4/9 = 5/9

(g) between B and E

Jaccard sim = 3/10

Jaccard dist = 1 - 3/10 = 7/10

(h) between C and D

Jaccard sim = 4/9

Jaccard dist = 1 - 4/9 = 5/9

(i) between C and E

Jaccard sim = 4/9

Jaccard dist = 1 - 4/9 = 5/9

(j) between D and E vectors

Jaccard sim = 8/10

Jaccard dist = 1 - 8/10 = 2/10

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.

Sol:

Q3: A = 1111000000

B = 0100100101

C = 0000011110

D = 0111111111

E = 1011111111

Manhattan distance is absolute sum of differences b/w vectors.

between A and B

Manhattan dist = 6

between A and C

Manhattan dist = 8

between A and D

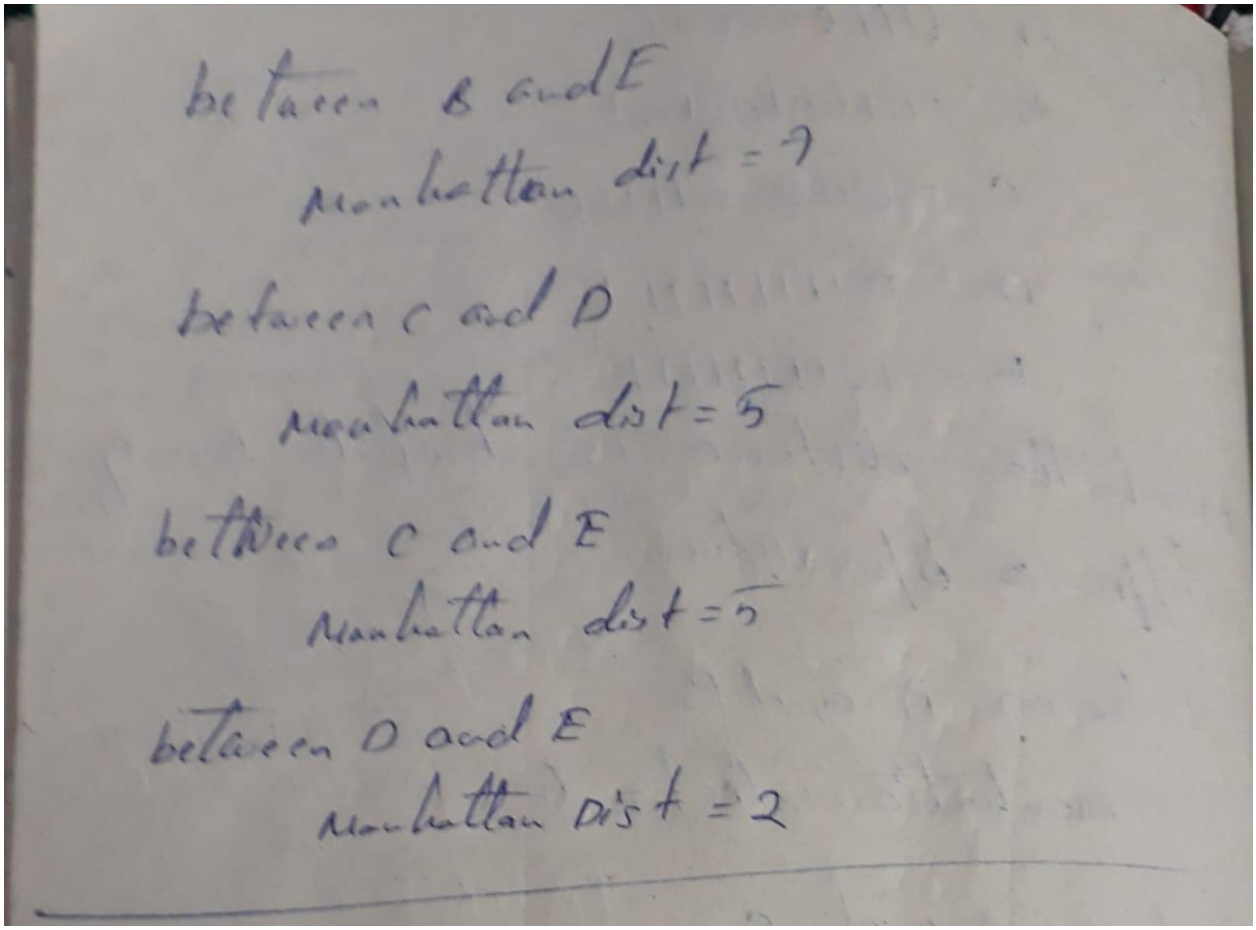
Manhattan dist = 7

between A and E

Manhattan dist = 7

between B and C

Manhattan dist = 5



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**.

Sol:

Q.4. 'He', 'she', 'his', 'hers'

Edit distance = $x + y - 2[lcs(x, y)]$

1st pair 'He' and 'she'

lcs = 2

edit distance = $2 + 3 - 2(2)$

= 1

2nd pair 'he' and 'his'

lcs = 1

edit dist = $2 + 3 - 2(1)$

= 3

3rd pair 'he' and 'hers'

LCS = 2

edit distance = 2 + 4 - 2(2)

= 2

4th pair 'she' and 'his'

LCS = 1

edit distance = 3 + 3 - 2(1)

= 4

5th pair 'she' and 'hers'

LCS = 2

edit distance = 3 + 4 - 2(2)

= 3

6th pair 'his' and 'hers'

LCS = 2

edit distance = 3 + 4 - 2(2)

= 3