Name : Pawan Suresh Bathe Roll No.:CS18M519

CS6745: Mining Massive DataSets

Tutorial 6

October 13, 2019

- Write your name and roll number in the space provided
- Be neat, and use the space judiciously.
- **Rough sheets won't be evaluated.**

1. (2 marks) Find the L_1 and L_2 distances between the points (5, 6, 7) and (8, 2, 4)

Answer:

a) L_1 distance between the points (5, 6, 7) and (8, 2, 4) = $|5-8| + |6-2| + |7-4| = 3 + 4 + 3 = 13$

b) L_2 distance between the points (5, 6, 7) and (8, 2, 4) = $\sqrt{(5-8)^2 + (6-2)^2 + (7-4)^2} = \sqrt{9 + 16 + 9} = \sqrt{3}4 = 5.83$

2. (5 marks) Suppose we are trying to perform entity resolution among bibliographic references, and we score pairs of references based on the similarities of their titles, list of authors, and place of publication. Suppose also that all references include a year of publication, and this year is equally likely to be any of the ten most recent years. Further, suppose that we discover that among the pairs of references with a perfect score, there is an average difference in the publication year of 0.1. Suppose that the pairs of references with a certain score s are found to have an average difference in their publication dates of 2. What is the fraction of pairs with score s that truly represent the same publication? *Note:* Do not make the mistake of assuming the average difference in publication date between random pairs is 5 or 5.5. You need to calculate it exactly, and you have enough information to do so.

Answers: We are given that, among the pairs of references with a perfect score, there is an average difference in the publication year of 0.1. Also, the pairs of references with a certain score s are found to have an average difference in their publication dates of 2.

10 most recent years 2010- 2019: Sum of difference between a pair with one year of pair being 2019 is : 0+1+2+3+4+5+6+7+8+9 = 45 Average difference between a pair with one year of pair being 2018 (excluding 2019, since it is covered above) is : 0+1+2+3+4+5+6+7+8 = 36 Similarly for 2018,2017,2016,2015,2014,2013,2012,2011,2010 it is 21, 15, 10, 6, 3, 1, 0 consecutively.

So average difference in publication date between random pairs is (45 + 36 + 21+15 + 10 + 6 + 3 + 1 + 0) / 55 (number of year pairs) = 137/55 = 2.49

So fraction of pairs with score s that truly represent the same publication. = (2.49-2)/(2.49-0.1) = 0.49/2.39 = 0.205

3. (3 marks) Let us compute sketches using the following four "random" vectors:

$$v_1 = [+1, +1, +1, -1]; v_2 = [+1, +1, -1, +1]$$

$$v_3 = [+1, -1, +1, +1]; v_4 = [-1, +1, +1, +1]$$

Compute the sketches of the following vectors,

a) $[2, 3, 4, 5]$

b) $[-2, 3, -4, 5]$

c) $[2, -3, 4, -5]$

For each pair, what is the estimated angle between them, according to the sketches? What are the true angles?

Answer:

a) Sketch for [2, 3, 4, 5] :
let x = [2, 3, 4, 5]
v1.x= 2 + 3 + 4 - 5 = 4, since v1.x is positive, first component of sketch is +1,
v2.x= 2 + 3 - 4 + 5 = 6, since v2.x is positive, second component of sketch is +1,
v3.x= 2 - 3 + 4 + 5 = 8 since v3.x is positive, third component of sketch is +1,
v4.x= -2 + 3 + 4 + 5 = 10 since v4.x is positive, fourth component of sketch is +1
so the sketch is [+1, +1, +1, +1]

b) Sketch for [-2, 3, -4, 5] :
let x = [-2, 3, -4, 5]
v1.x= -2 + 3 -4 - 5 = -8, since v1.x is negative, first component of sketch is -1,
v2.x= -2 + 3 + 4 + 5 = 10, since v2.x is positive, second component of sketch is +1,
v3.x= -2 - 3 - 4 + 5 = -4 since v3.x is negative, third component of sketch is -1,
v4.x= +2 + 3 -4 + 5 = 6 since v4.x is positive, fourth component of sketch is +1
so the sketch is [-1, +1, -1, +1]

c) Sketch for [2, -3, 4, -5] :
let x = [2, -3, 4, -5]
v1.x= 2 - 3 + 4 + 5 = 8 , since v1.x is positive, first component of sketch is +1,
v2.x= 2 - 3 - 4 - 5 = -10, since v2.x is negative, second component of sketch is -1,
v3.x= 2 + 3 + 4 - 5 = 4 since v3.x is positive, third component of sketch is +1,
v4.x= -2 - 3 + 4 - 5 = -6 since v4.x is negative, fourth component of sketch is -1
so the sketch is [+1, -1, +1, -1]

Since , sketches for a and b agree in 2/4 positions i.e 1/2 positions, we estimate angle between them is 180 degrees.
Since , sketches for b and c agree in 0 positions, we estimate angle between them is 0 degrees.
Since , sketches for a and c agree in 2/4 positions i.e 1/2 positions, we estimate angle between them is 180 degrees.

We will calculate magnitudes of a, b and c, and products a.b, b.c and a.c
a.b = 5*5 + (4*-4) + 3*3 +(2*-2) = 14
b.c = (5*-5) + (-4*4) + (3*-3) + (-2*2) = 54
a.c = (5*-5) + 4*4 + (3*-3) + 2*2 = -14

Magnitude of a = $\sqrt{5^2 + 4^2 + 3^2 + 2^2}$ = 7.348

Since, vector components have just sign difference , we know that a = b = c = 7.348

So, cosine of the angle between a and b is = 4/ 7.348*7.348 = 0.259 hence angle between a and b is 0.259 = 74.989 degrees
So, cosine of the angle between b and c is = 54/ 7.348*7.348 = 1.000 hence angle between a and b is 0.259 = 0 degrees
So, cosine of the angle between a and c is = 14/ 7.348*7.348 = 0.259 hence angle between a and b is 0.259 = 74.989 degrees