DATA SIMILARITY AND DISTANCE

LECTURE 3

Dr. Mostafa Elmasry

Similarity and Distance

- For many different problems we need to quantify how close two objects are.
- Examples:
  - For an item bought by a customer, find other similar items
  - Group together the customers of site so that similar customers are shown the same ad.
  - Group together web documents so that you can separate the ones that talk about politics and the ones that talk about sports.
  - Find all the near-duplicate mirrored web documents.
  - Find credit card transactions that are very different from previous transactions.
- To solve these problems we need a definition of similarity, or distance.
  - The definition depends on the type of data that we have

Similarity

- Numerical measure of how alike two data objects are.
  - A function that maps pairs of objects to real values
  - Higher when objects are more alike.
- Often falls in the range [0,1], sometimes in [-1,1]

- Desirable properties for similarity
  1. s(p, q) = 1 (or maximum similarity) only if p = q. (Identity)
  2. s(p, q) = s(q, p) for all p and q. (Symmetry)

Similarity between sets

- Consider the following documents

apple releases new ipod	apple releases new ipad	new apple pie recipe

- Which ones are more similar?
- How would you quantify their similarity?

Similarity: Intersection

- Number of words in common

apple releases new ipod	apple releases new ipad	new apple pie recipe

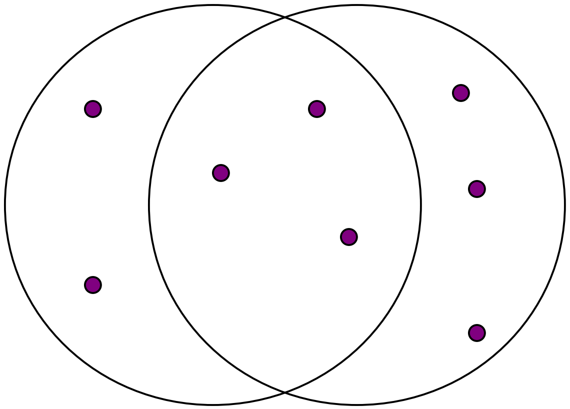
- Sim(D,D) = 3, Sim(D,D) = Sim(D,D) =2
- What about this document?

Vefa releases new book with apple pie recipes

- Sim(D,D) = Sim(D,D) = 3

Jaccard Similarity

- The Jaccard similarity (Jaccard coefficient) of two sets S_1, S_2 is the size of their intersection divided by the size of their union.
  - JSim $(C_1, C_2) = |C_1 \cap C_2| / |C_1 \cup C_2|$.



3 in intersection.
8 in union.
Jaccard similarity
= 3/8

- Extreme behavior:
  - Jsim(X,Y) = 1, iff X = Y
  - Jsim(X,Y) = 0 iff X,Y have not elements in common
- JSim is symmetric

Similarity: Intersection

- Number of words in common

apple releases new ipod	apple releases new ipad	new apple pie recipe	Vefa releases new book with apple pie recipes

- JSim(D,D) = 3/5
- JSim(D,D) = JSim(D,D) = 2/6
- JSim(D,D) = JSim(D,D) = 3/9

Similarity between vectors

Documents (and sets in general) can also be represented as vectors

document	Apple	Microsoft	Obama	Election
D1	1	2	0	0
D2	3	6	0	0
D3	0	0	1	2

How do we measure the similarity of two vectors?

How well are the two vectors aligned?

Example

document	Apple	Microsoft	Obama	Election
D1	1/3	2/3	0	0
D2	1/3	2/3	0	0
D3	0	0	1/3	2/3

Documents D1, D2 are in the “same direction”
Document D3 is orthogonal to these two

Cosine Similarity

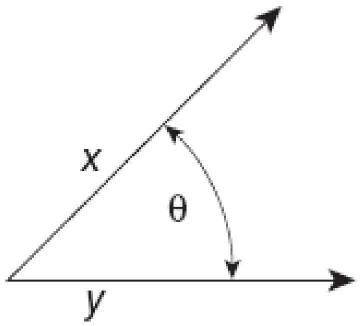


Figure 2.16. Geometric illustration of the cosine measure.

- Sim(X,Y) = cos(X,Y)
  - The cosine of the angle between X and Y

- If the vectors are aligned (correlated) angle is zero degrees and cos(X,Y)=1
- If the vectors are orthogonal (no common coordinates) angle is 90 degrees and cos(X,Y) = 0

- Cosine is commonly used for comparing documents, where we assume that the vectors are normalized by the document length.

Cosine Similarity - math

- If d_1 and d_2 are two vectors, then

  $$\cos(d_1, d_2) = (d_1 \bullet d_2) / ||d_1||\ ||d_2||\ ,$$

  where • indicates vector dot product and || *d* || is the length of vector *d*.

- Example:

  d_1 **= 3 2 0 5 0 0 0 2 0 0**

  d_2 **= 1 0 0 0 0 0 0 1 0 2**

$d_1 \bullet d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$

$||d_1|| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$

$||d_2|| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5} = (6)^{0.5} = 2.245$

$\cos(d_1, d_2) = .3150$

Similarity between vectors

document	Apple	Microsoft	Obama	Election
D1	1	2	0	0
D2	3	6	0	0
D3	0	0	1	2

$$\cos(D1,D2) = 1$$

$$\cos(D1,D3) = \cos(D2,D3) = 0$$

Cosine similarity between two sentences

- osine similarity between two sentences can be found as a dot product of their vector representation.
- Their are various ways to represent sentences/paragraphs as vectors.

1.Julie loves me more than Linda loves me
2.Jane likes me more than Julie loves me

me Julie loves Linda than more likes Jane

Cosine similarity between two sentences

me 2 2
Jane 0 1
Julie 1 1
Linda 1 0
likes 0 1
loves 2 1
more 1 1
than 1 1

The two vectors are, again:
a: [2, 1, 0, 2, 0, 1, 1, 1]
b: [2, 1, 1, 1, 1, 0, 1, 1]

Distance

- Numerical measure of how different two data objects are
  - A function that maps pairs of objects to real values
  - Lower when objects are more alike
- Minimum distance is 0, when comparing an object with itself.
- Upper limit varies

Distance Metric

- A distance function d is a distance metric if it is a function from pairs of objects to real numbers such that:
  1. $d(x,y) \geq 0$. (non-negativity)
  2. $d(x,y) = 0$ iff $x = y$. (identity)
  3. $d(x,y) = d(y,x)$. (symmetry)
  4. $d(x,y) \leq d(x,z) + d(z,y)$ (triangle inequality).

Triangle Inequality

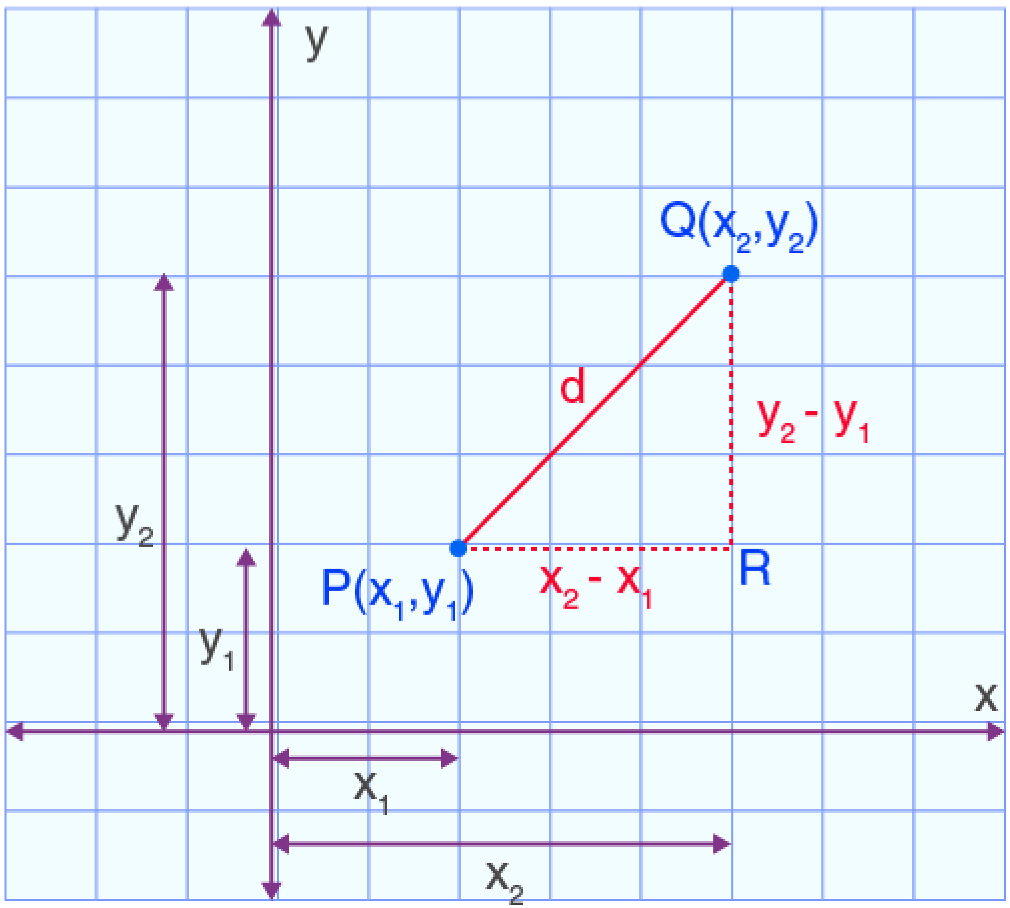
- Triangle inequality guarantees that the distance function is well-behaved.
  - The direct connection is the shortest distance

- It is useful also for proving properties about the data
  - For example, suppose I want to find an object that minimizes the sum of distances to all points in my dataset

Euclidean Distance

- In Mathematics, the Euclidean distance is defined as the distance between two points.
- In other words, the Euclidean distance between two points in the Euclidean space is defined as the length of the line segment between two points.
- **d =**$\sqrt{[(x_2 - x_1)^2 + (y_2 - y_1)^2]}$

Euclidean Distance



Euclidean Distance example

- **Example 1:** Find the distance between points P(3, 2) and Q(4, 1).
- **Solution:**
- Given:
- $PQ = \sqrt{[(4 - 3)^2 + (1 - 2)^2]}$
- $PQ = \sqrt{[(1)^2 + (-1)^2]}$
- $PQ = \sqrt{2}$ units.

Higher dimensions

In three dimensions, for points given by their Cartesian coordinates, the distance is

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_3 - q_3)^2}.$$

In general, for points given by Cartesian coordinates in n-dimensional Euclidean space, the distance is

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_i - q_i)^2 + \cdots + (p_n - q_n)^2}.$$

Hamming Distance

- Hamming distance is the number of positions in which bit-vectors differ.
  - Example: $p_1 = 10101$
    $p_2 = 10011$.
    - $d(p_1, p_2) = 2$ because the bit-vectors differ in the 3rd and 4th positions.
    - The L_1 norm for the binary vectors

- Hamming distance between two vectors of categorical attributes is the number of positions in which they differ.
  - Example: x = (married, low income, cheat),
    y = (single, low income, not cheat)
  - d(x,y) = 2

Why Hamming Distance Is a Distance Metric

- $d(x,x) = 0$ since no positions differ.
- $d(x,y) = d(y,x)$ by symmetry of "different from."
- $d(x,y) \geq 0$ since strings cannot differ in a negative number of positions.
- Triangle inequality: changing *x* to *z* and then to *y* is one way to change *x* to *y*.

Hamming Distance Calculation

Calculation of Hamming Distance

In order to calculate the Hamming distance between two strings, and , we perform their XOR operation, (a$\oplus$ b), and then count the total number of 1s in the resultant string.

Example

Suppose there are two strings 1101 1001 and 1001 1101.

11011001 $\oplus$ 10011101 = 01000100. Since, this contains two 1s, the Hamming distance, d(11011001, 10011101) = 2.

Distance between strings

- How do we define similarity between strings?

weird	wierd
intelligent	unintelligent
Athena	Athina

- Important for recognizing and correcting typing errors and analyzing DNA sequences.