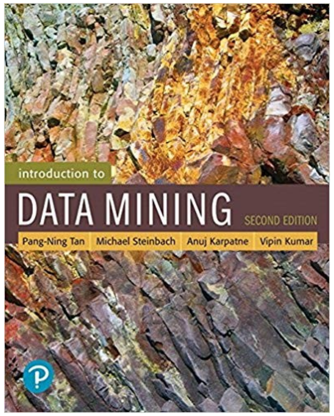


CIS 530—Advanced Data Mining

7- Similarity and Distance

Thomas W Gyeera, Assistant Professor
Computer and Information Science
University of Massachusetts Dartmouth

Courtesy to Prof. Panayiotis Tsaparas

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

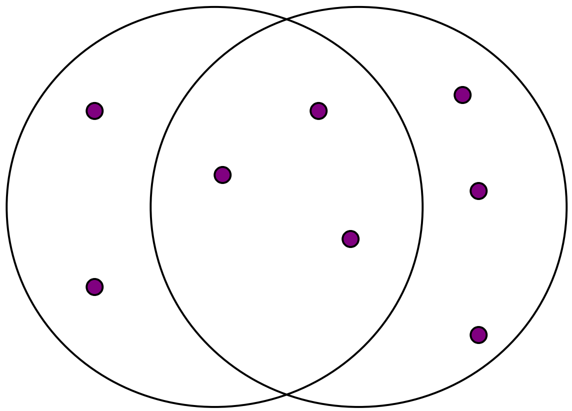
- Sim(D,D) = 3, Sim(D,D) = Sim(D,D) =2
- What about this document?

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	10	20	0	0
D2	30	60	0	0
D2	0	0	10	20

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
D2	0	0	1/3	2/3

Documents D1, D2 are in the “same direction”

Document D3 is orthogonal to these two

Cosine Similarity

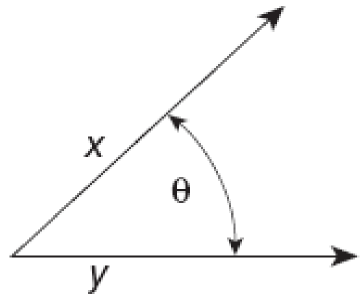


Figure 2.16. Geometric illustration of the cosine measure.

- 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 $\bullet$ 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	10	20	0	0
D2	30	60	0	0
D2	0	0	10	20

$$\cos(D1,D2) = 1$$
$$\cos(D1,D3) = \cos(D2,D3) = 0$$

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).

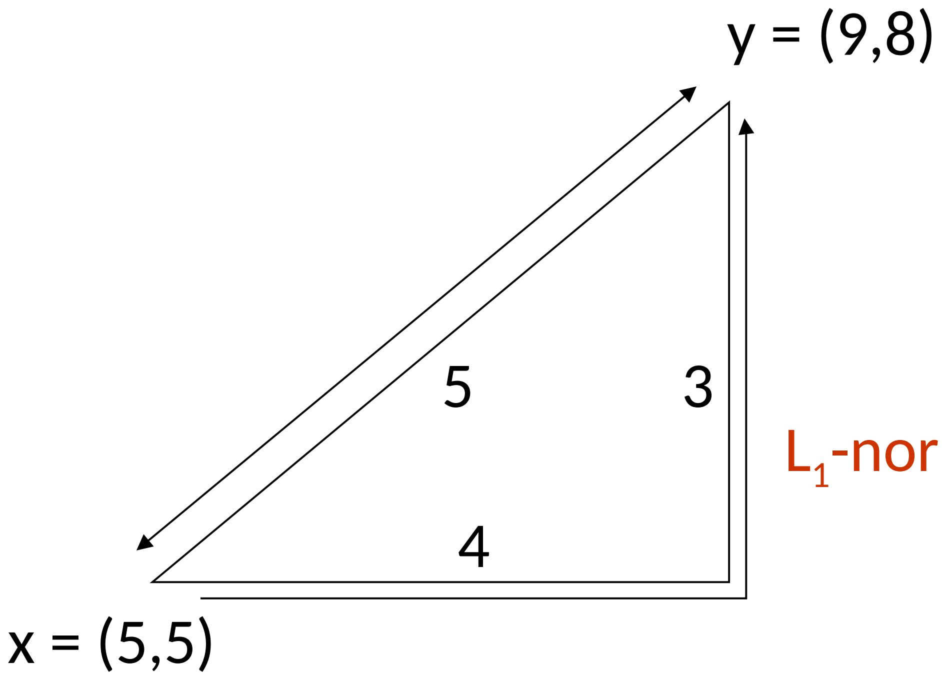
Distances for real vectors

- Vectors and
- L_p norms or Minkowski distance:
- L_2 norm: Euclidean distance:
- L_1 norm: Manhattan distance:
- L norm:
  - The limit of L_p as p goes to infinity.

L_p norms are known to be distance metrics

Example of Distances

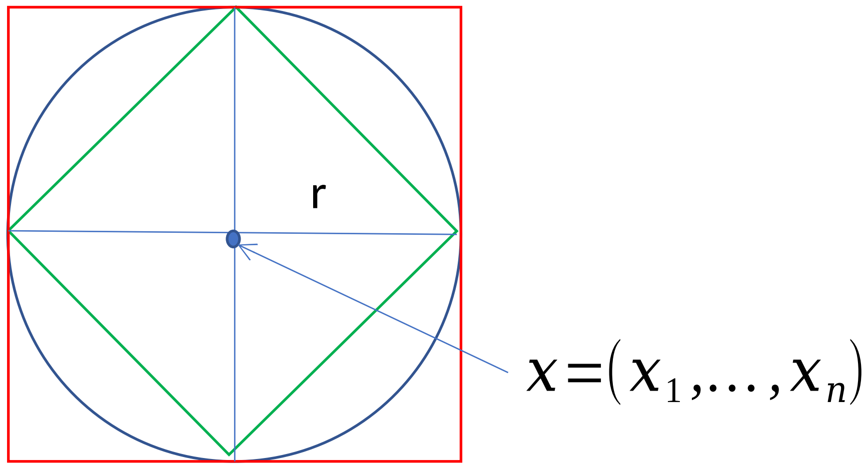
L_2-norm:



L_1-norm:

L-norm:

Example



Green: All points y at distance $L_1(x,y) = r$ from point x

Blue: All points y at distance $L_2(x,y) = r$ from point x

Red: All points y at distance $L(x,y) = r$ from point x

L_p distances for sets

- We can apply all the L_p distances to the cases of sets of attributes, with or without counts, if we represent the sets as vectors
  - E.g., a transaction is a 0/1 vector
  - E.g., a document is a vector of counts.

Similarities into distances

- Jaccard distance:

- Jaccard Distance is a metric
- Cosine distance:

- Cosine distance is a metric

Why Jaccard Distance is a Distance Metric?

- JDist(x,x) = 0
  - since JSim(x,x) = 1
- JDist(x,y) = JDist(y,x)
  - by symmetry of intersection
- JDist(x,y) $\geq$ 0
  - since intersection of x,y cannot be bigger than the union.
- Triangle inequality:
  - Follows from the fact that JSim(x,y) is the probability of randomly selected element from the union of x and y to belong to the intersection

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.
- For binary vectors if follows from the fact that L_1 norm is a metric

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.

Edit Distance for strings

- The edit distance of two strings is the number of inserts and deletes of characters needed to turn one into the other.
- Example: x = abcde ; y = bcduve.
  - Turn x into y by deleting a, then inserting u and v after d.
  - Edit distance = 3.
- Minimum number of operations can be computed using dynamic programming
- Common distance measure for comparing DNA sequences

Why Edit Distance is a Distance Metric?

- $d(x,x) = 0$ because 0 edits suffice.
- $d(x,y) = d(y,x)$ because insert/delete are inverses of each other.
- $d(x,y) \geq 0$: no notion of negative edits.
- Triangle inequality: changing *x* to *z* and then to *y* is one way to change *x* to *y*. The minimum is no more than that

Variant Edit Distances

- Allow insert, delete, and mutate.
  - Change one character into another.
- Minimum number of inserts, deletes, and mutates also forms a distance measure.

- Same for any set of operations on strings.
  - Example: substring reversal or block transposition OK for DNA sequences
  - Example: character transposition is used for spelling

Distances between distributions

- We can view a document as a distribution over the words

document	Apple	Microsoft	Obama	Election
D1	0.35	0.5	0.1	0.05
D2	0.4	0.4	0.1	0.1
D2	0.05	0.05	0.6	0.3

- KL-divergence (Kullback-Leibler) for distributions P,Q

- KL-divergence is asymmetric. We can make it symmetric by taking the average of both sides
- JS-divergence (Jensen-Shannon)

+