DISTANCE MEASURE

SYRACUSE UNIVERSITY
School of Information Studies

DISTANCE MEASURES

Similarity and distance: Two opposite concepts

- Similarity measures how close or similar two examples are.
- Distance measures how far or different two examples are.

The definitions of distance functions are dependent on variable types: numeric, nominal. Many data sets contain mixed types of attributes.

- Example: How similar are these two people?

  i = (Refund = No, Married, Income = 120K)

  j = (Refund = Yes, Married, Income = 90K)

NUMERIC ATTRIBUTES

If the data have all numeric attributes, distance measures can compare the numeric values of the attributes.

Some popular ones include *Minkowski distance*

$$d(i,j)=\sqrt[q]{(|x_{i_1} \quad x_{j_1}|^q+|x_{i_2} \quad x_{j_2}|^q+\ldots+|x_{i_p} \quad x_{j_p}|^q)}$$

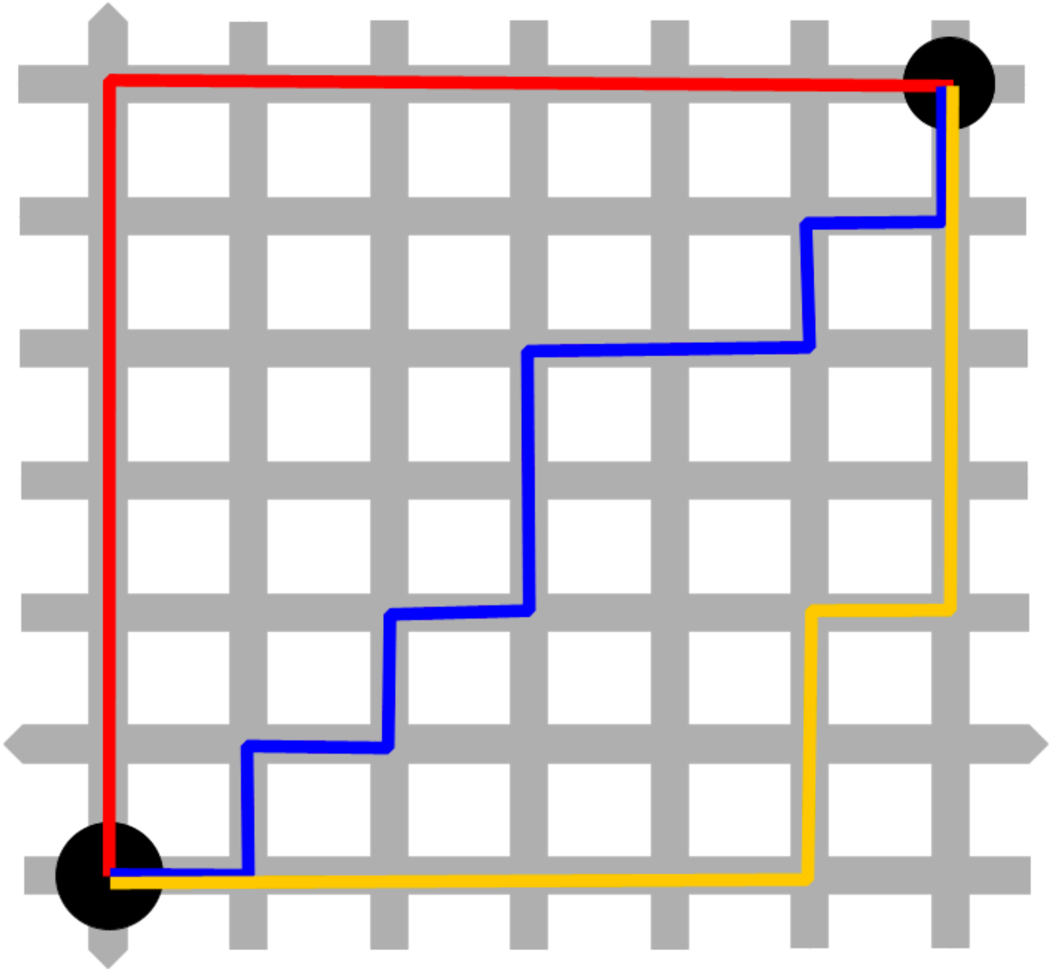
where *i* = (x_{i1}, x_{i2}, … , x_{ip}) and *j* = (x_{j1}, x_{j2}, … , x_{jp}) are two *p*-dimensional data instances, and *q* is a positive integer.

If *q* = 1, *d* is *Manhattan distance.*

Taking the absolute value of the differences between attribute values

$$d(i,j)=|x_{i_1} \quad x_{j_1}|+|x_{i_2} \quad x_{j_2}|+\ldots+|x_{i_p} \quad x_{j_p}|$$

MANHATTAN DISTANCE

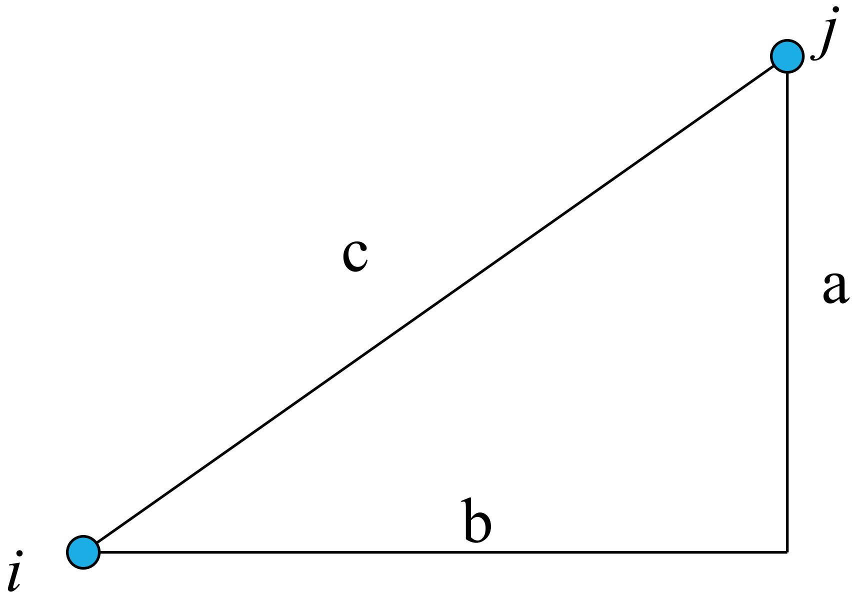


https://upload.wikimedia.org/wikipedia/commons/d/de/Manhattan_distance_bgiu.png

EUCLIDEAN DISTANCE

When *q* = 2, *d* is *Euclidean distance*:

$$d(i,j)=\sqrt{(|x_{i_1}\quad x_{j_1}|^2+|x_{i_2}\quad x_{j_2}|^2+...+|x_{i_p}\quad x_{j_p}|^2)}$$



$$d_1(i,j)=a+b$$

$$d_2(i,j)=\sqrt{a^2+b^2}=c$$

PROPERTIES OF DISTANCE MEASURE

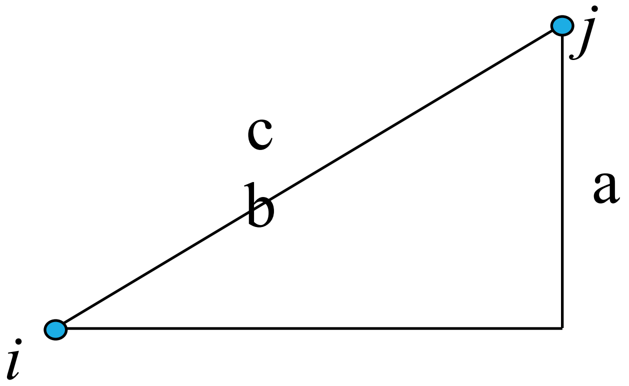
A distance measure should satisfy the following requirements:

$d(i, j) \geq 0$ (nonnegative value)

$d(i, i) = 0$ (zero distance to itself)

$d(i, j) = d(j, i)$ (symmetric measure)

$d(i, j) \leq d(i, k) + d(k, j)$ (shortest distance between two points)



DISTANCE BETWEEN NOMINAL VALUES

Example: How similar are these two people?

i = (Refund = Yes, Married, Income = 120K)

j = (Refund = No, Divorced, Income = 90K)

Taxpayer	Refund	Marital Status	Income in Thousands
i	Yes	Married	120
j	No	Divorced	90

METHOD 1: SIMPLE MATCHING

Taxpayer	Refund	Marital Status	Income in Thousands
i	Yes	Married	120
j	No	Divorced	90

m: Number of matches; *p*: Total number of nominal variables

$$d(i,j) = \frac{p \quad m}{p}$$

METHOD 2: CONVERT NOMINAL TO BINARY VARIABLES

Taxpayer	Refund	Marital Status	Income in Thousands
i	Yes	Married	120
j	No	Divorced	90

Convert a nominal attribute to multiple binary attributes, and treat binary attributes as numeric (0 or 1).

Taxpayer	Refund	Married?	Divorced?	Single?	Income
1	1	1	0	0	120
2	0	0	1	0	90

BINARY VARIABLES: SYMMETRIC OR ASYMMETRIC

All patients run through many tests.

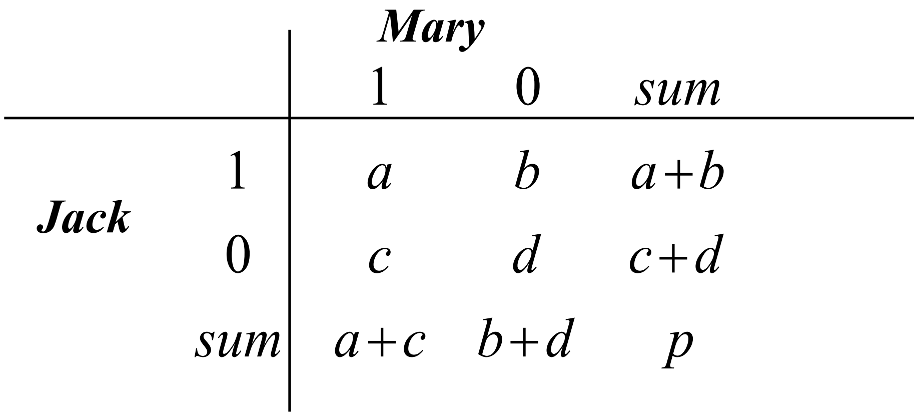
How different are their test results?

Patient	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6
Jack	1	0	1	0	0	0
Mary	1	0	1	0	1	0

BINARY VARIABLES: SYMMETRIC OR ASYMMETRIC

A contingency table for binary data

- Gives the number of attributes of each pair of values



		Mary		
		1	0	*sum*
Jack	1	a	b	$a+b$
	0	c	d	$c+d$
	sum	$a+c$	$b+d$	p

Patient	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6
Jack	1	0	1	0	0	0
Mary	1	0	1	0	1	0

SYMMETRIC BINARY ATTRIBUTES

Distance measure for symmetric binary attributes:

		Object *j*		
		1	0	*sum*
Object *i*	1	a	b	$a+b$
	0	c	d	$c+d$
	sum	$a+c$	$b+d$	p

$$d(i,j) = \frac{b+c}{a+b+c+d}$$

ASYMMETRIC BINARY ATTRIBUTES

If most test results are negative, *d* will be much greater than *a*, *b*, and *c*. Sharing many negative test results is not that informative to doctors.

		Object *j*		
		1	0	*sum*
Object *i*	1	a	b	$a+b$
	0	c	d	$c+d$
	sum	$a+c$	$b+d$	p

Distance measure for asymmetric binary attributes:

$$d(i,j)=\frac{b+c}{a+b+c}$$

DISTANCE BETWEEN ORDINAL VALUES

Method 1: Treat as nominal.

Method 2: Treat as numeric.

ATTRIBUTES OF MIXED TYPES

A database may contain different types of attributes: Symmetric binary, asymmetric binary, nominal, ordinal, numerical

How to compute the distance between examples with heterogeneous attributes?

Calculate distance for each type of attribute and aggregate.

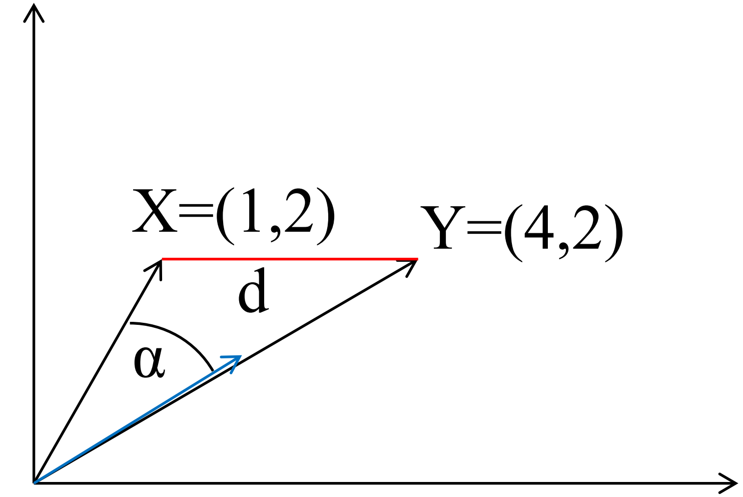
SIMILARITY MEASURE

If defining a distance measured in [0,1] range, similarity can be defined as 1 – d.

Other similarity measures:

Cosine similarity measure

VECTOR SPACE REPRESENTATION AND COSINE SIMILARITY



Distance and similarity measures

Euclidean distance

$$d = x \quad y = \sqrt{(x_1 \quad y_1)^2 + (x_2 \quad y_2)^2}$$

$$= \sqrt{(1 \quad 4)^2 + (2 \quad 2)^2} = 3$$

Cosine similarity

$$\cos(\quad) = \frac{x \quad y}{|x||y|} = \frac{x_1 y_1 + x_2 y_2}{\sqrt{x_1{}^2 + x_2{}^2} \quad \sqrt{y_1{}^2 + y_2{}^2}}$$

$$= \frac{1 \cdot \ 4 + 2 \cdot \ 2}{\sqrt{1^2 + 2^2} \quad \sqrt{4^2 + 2^2}} = \frac{8}{\sqrt{5} \quad \sqrt{20}} = 0.8$$

COSINE SIMILARITY

In the range of [0,1]:

“0” means two vectors are perpendicular to each other.

“1” means same vector direction and length.

Commonly used in information retrieval and text mining to compare document similarity

High-dimensional space

Each word in the vocabulary is a dimension.