Data Mining & Knowledge Discovery

Similarity and Dissimilarity Measures

Md. Biplob Hosen

Lecturer, IIT-JU

Email: biplob.hosen@juniv.edu

Similarity and Dissimilarity Measures

Similarity measure:

- Numerical measure of how alike two data objects are.
- Is higher when objects are more alike.
- Often falls in the range [0,1]

Dissimilarity measure:

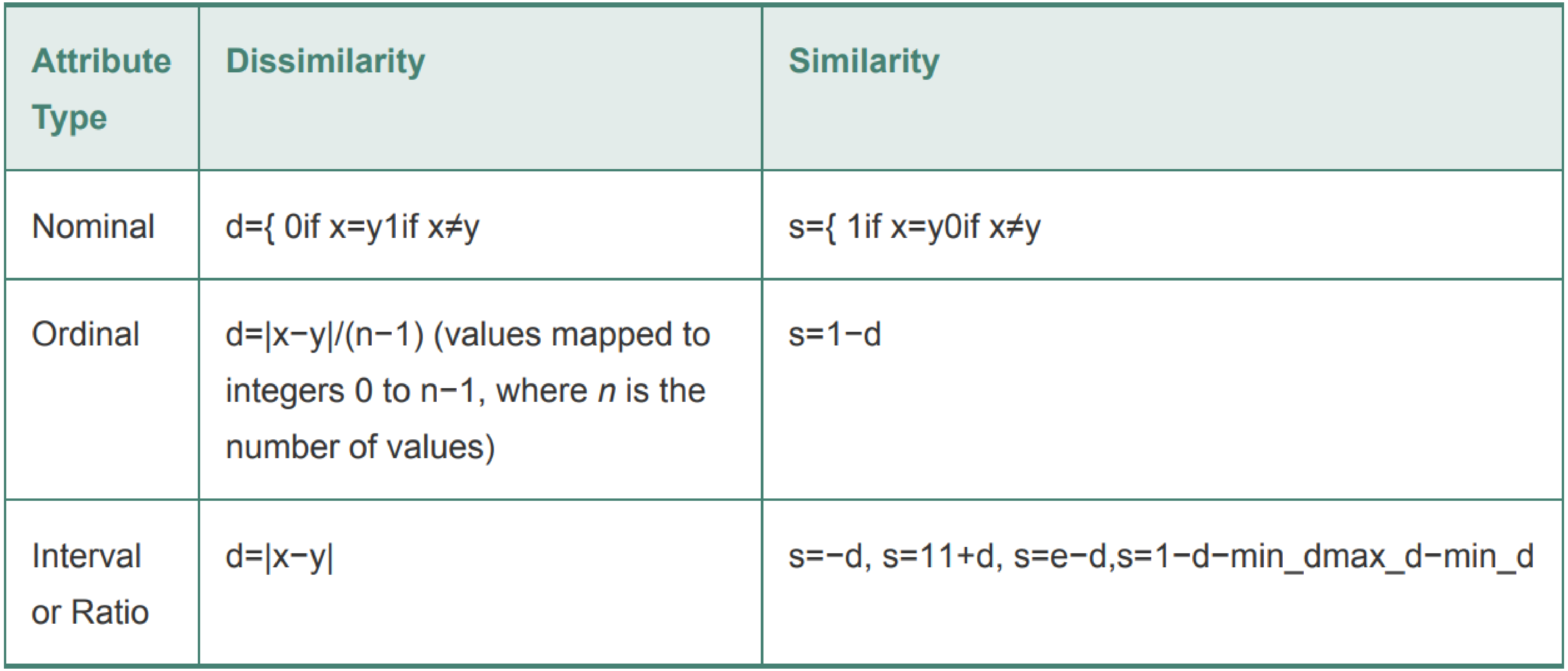
- Numerical measure of how different two data objects are.
- Lower when objects are more alike.
- Minimum dissimilarity is often 0.
- Upper limit varies.
- Proximity refers to a similarity or dissimilarity.

Transformations

- Transformations are often applied to convert a similarity to a dissimilarity, or vice versa, or to transform a proximity measure to fall within a particular range, such as [0,1].
- For instance, we may have similarities that range from 1(not at all similar) to 10 (completely similar), but the particular algorithm or software package may be designed to work only with dissimilarities, or it may work only with similarities in the interval [0,1].
- We can make them fall within the range [0, 1] by using the transformation s’=(s-1)/9 where s and s′ are the original and new similarity values, respectively.
- In the more general case, the transformation of similarities to the interval [0, 1] is given by the expression s’=(s-min_s)/(max_s-min_s).
- Likewise, dissimilarity measures with a finite range can be mapped to the interval [0,1] by using the formula d’=(d-min_d)/(max_d-min_d).
- This is an example of a linear transformation, which preserves the relative distances between points.

Similarity/Dissimilarity for Simple Attributes

- The following table shows the similarity and dissimilarity between two objects, x and y, with respect to a single, simple attribute.



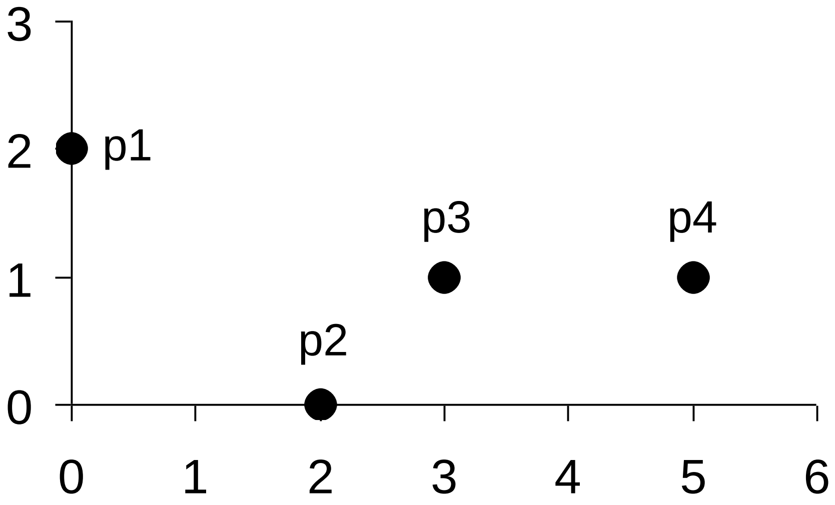
Attribute Type	Dissimilarity	Similarity
Nominal	d={ 0if x=y1if x≠y	s={ 1if x=y0if x≠y
Ordinal	d=|x−y|/(n−1) (values mapped to integers 0 to n−1, where *n* is the number of values)	s=1−d
Interval or Ratio	d=|x−y|	s=−d, s=11+d, s=e−d,s=1−d−min_dmax_d−min_d

Dissimilarity - Euclidean Distance

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

- where n is the number of dimensions (attributes) and xk and yk are, respectively, the kth attributes (components) or data objects x and y.

Dissimilarity - Euclidean Distance



point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Distance Matrix

Dissimilarity - Minkowski Distance

- Minkowski Distance is a generalization of Euclidean Distance:

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r}$$

Where r is the order parameter, n is the number of dimensions (attributes) and xk and yk are, respectively, the kth attributes (components) or data objects x and y.

Minkowski Distance: Examples

r = 1. Manhattan distance.

- A common example of this for binary vectors is the Hamming distance, which is just the number of bits that are different between two binary vectors

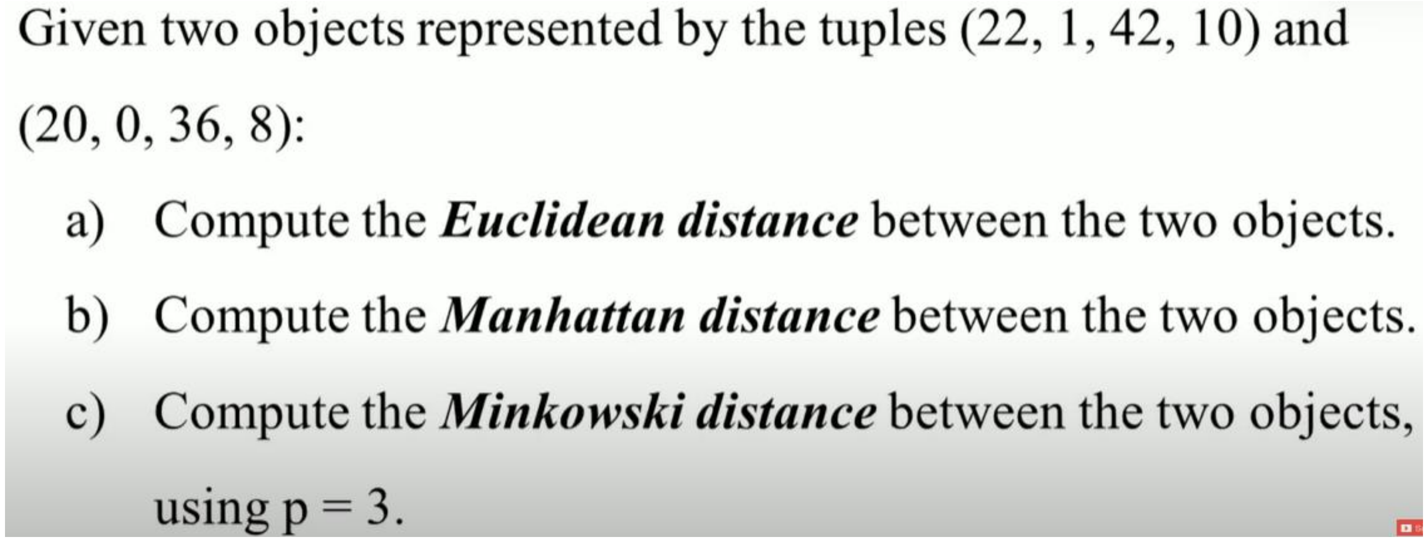
r = 2. Euclidean distance

r = ∞. “supremum” distance.

- This is the maximum difference between any component of the vectors

- Do not confuse r with n, i.e., all these distances are defined for all numbers of dimensions.

Minkowski Distance: Examples (2)

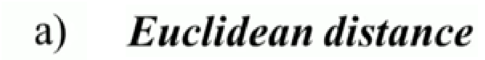


Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8):

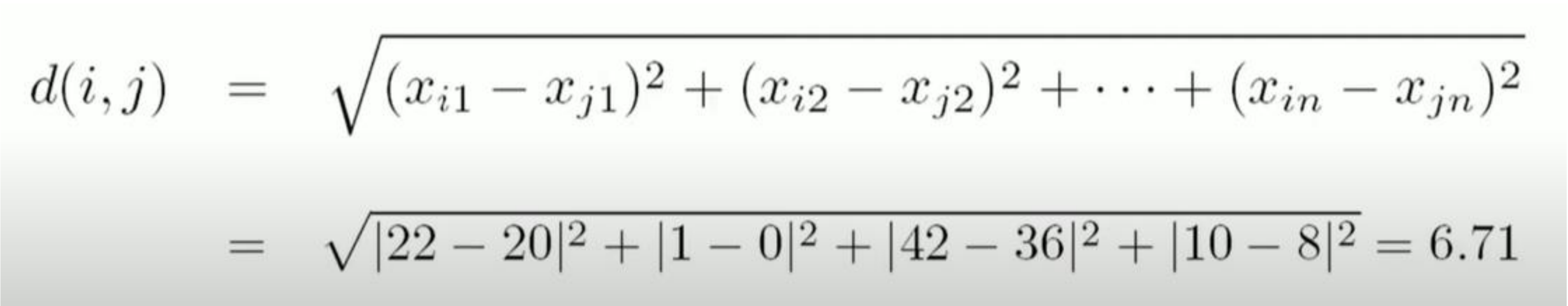
a) Compute the ***Euclidean distance*** between the two objects.

b) Compute the ***Manhattan distance*** between the two objects.

c) Compute the ***Minkowski distance*** between the two objects, using p = 3.



a) ***Euclidean distance***

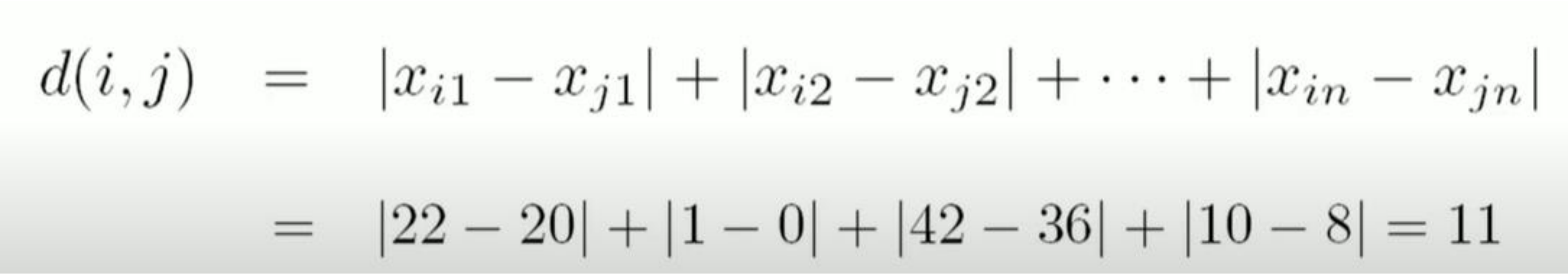


$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{in} - x_{jn})^2}$$

$$= \sqrt{|22-20|^2 + |1-0|^2 + |42-36|^2 + |10-8|^2} = 6.71$$

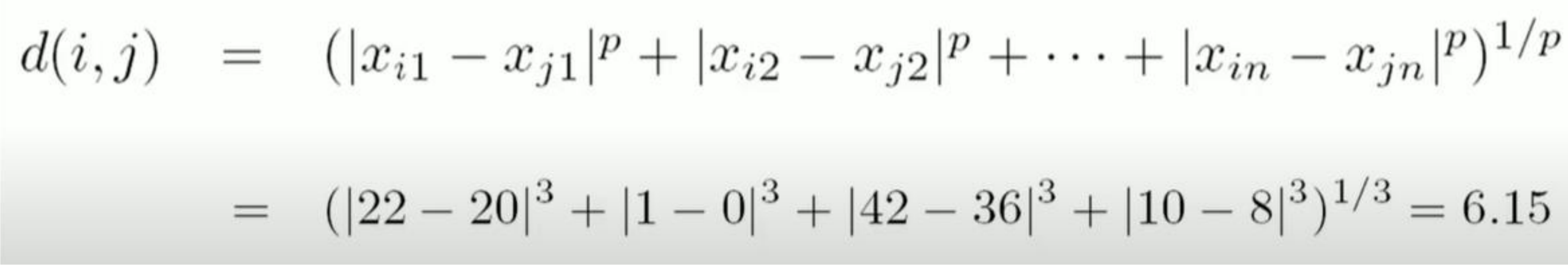
Minkowski Distance: Examples (2)

b) ***Manhattan distance***



$$\begin{aligned} d(i,j) &= |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{in} - x_{jn}| \\ &= |22 - 20| + |1 - 0| + |42 - 36| + |10 - 8| = 11 \end{aligned}$$

c) ***Minkowski distance***



$$\begin{aligned} d(i,j) &= (|x_{i1} - x_{j1}|^p + |x_{i2} - x_{j2}|^p + \cdots + |x_{in} - x_{jn}|^p)^{1/p} \\ &= (|22 - 20|^3 + |1 - 0|^3 + |42 - 36|^3 + |10 - 8|^3)^{1/3} = 6.15 \end{aligned}$$

Thank You!