

Similarity / Distance Measures

CSE 4334 / 5334 Data Mining
Spring 2019

Won Hwa Kim

(Slides courtesy of Pang-Ning Tan, Michael Steinbach and Vipin Kumar)



Similarity and Dissimilarity



Similarity

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

Dissimilarity

- Numerical measure of how different are two data objects
- Lower when objects are more alike
- Minimum dissimilarity is often 0
- Upper limit varies

Proximity refers to a similarity or dissimilarity

Similarity and Dissimilarity



Similarity and Dissimilarity of Simple Attributes

Dissimilarity between Objects

- Distance
- Set Difference
- ...

Similarity between Objects

- Binary Vectors
- Vectors
- ...

3

Copyright ©2007-2017 The University of Texas at Arlington. All Rights Reserved.

Similarity/Dissimilarity for Simple Attributes



p and q are the attribute values for two data objects.

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{ p-q }{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{ p-q }{n-1}$
Interval or Ratio	$d = p - q $	$s = -d, s = \frac{1}{1+d}$ or $s = 1 - \frac{d - \min_d}{\max_d - \min_d}$

Table 5.1. Similarity and dissimilarity for simple attributes

4

Copyright ©2007-2017 The University of Texas at Arlington. All Rights Reserved.

Euclidean Distance



Euclidean Distance

$$dist = \sqrt{\sum_{k=1}^n (p_k - q_k)^2}$$

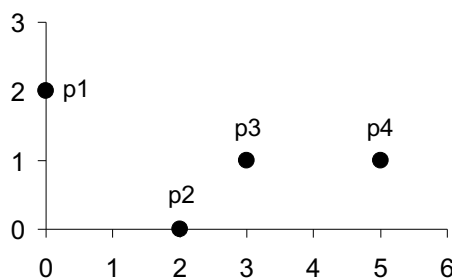
Where n is the number of dimensions (attributes) and p_k and q_k are, respectively, the k^{th} attributes (components) or data objects p and q .

Standardization is necessary, if scales differ.

5

Copyright ©2007-2017 The University of Texas at Arlington. All Rights Reserved.

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

6

Copyright ©2007-2017 The University of Texas at Arlington. All Rights Reserved.

Minkowski Distance



Minkowski Distance is a generalization of Euclidean Distance

$$dist = \left(\sum_{k=1}^n |p_k - q_k|^r \right)^{\frac{1}{r}}$$

Where r is a parameter, n is the number of dimensions (attributes) and p_k and q_k are, respectively, the k -th attributes (components) or data objects p and q .

7

Copyright ©2007-2017 The University of Texas at Arlington. All Rights Reserved.

Minkowski Distance: Examples



- ❖ $r = 1$. City block (Manhattan, taxicab, L_1 norm) distance.

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

- ❖ $r = 2$. Euclidean distance

- ❖ $r \rightarrow \infty$. “supremum” (L_{\max} norm, L_{∞} norm) 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.

8

Copyright ©2007-2017 The University of Texas at Arlington. All Rights Reserved.

Minkowski Distance



Data

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

Distance Matrix

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

L2	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

L_∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

9

Copyright ©2007-2017 The University of Texas at Arlington. All Rights Reserved.

Common Properties of a Distance



❖ Distances, such as the Euclidean distance, have some well known properties.

1. $d(p, q) \geq 0$ for all p and q and $d(p, q) = 0$ only if $p = q$. (Positive definiteness)
2. $d(p, q) = d(q, p)$ for all p and q . (Symmetry)
3. $d(p, r) \leq d(p, q) + d(q, r)$ for all points p, q , and r . (Triangle Inequality)

where $d(p, q)$ is the distance (dissimilarity) between points (data objects), p and q .

❖ A distance that satisfies these properties is a **metric**

10

Copyright ©2007-2017 The University of Texas at Arlington. All Rights Reserved.

Common Properties of a Similarity



❖ Similarities, also have some well known properties.

1. $s(p, q) = 1$ (or maximum similarity) only if $p = q$.
2. $s(p, q) = s(q, p)$ for all p and q . (Symmetry)

where $s(p, q)$ is the similarity between points (data objects), p and q .

Similarity Between Binary Vectors



❖ Common situation is that objects, p and q , have only binary attributes

❖ Compute similarities using the following quantities

M_{01} = the number of attributes where p was 0 and q was 1
 M_{10} = the number of attributes where p was 1 and q was 0
 M_{00} = the number of attributes where p was 0 and q was 0
 M_{11} = the number of attributes where p was 1 and q was 1

❖ Simple Matching and Jaccard Coefficients

SMC = number of matches / number of attributes
= $(M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$

J = number of 11 matches / number of not-both-zero attributes values
= $(M_{11}) / (M_{01} + M_{10} + M_{11})$



SMC versus Jaccard: Example

$$p = 1000000000$$

$$q = 0000001001$$

$M_{01} = 2$ (the number of attributes where p was 0 and q was 1)

$M_{10} = 1$ (the number of attributes where p was 1 and q was 0)

$M_{00} = 7$ (the number of attributes where p was 0 and q was 0)

$M_{11} = 0$ (the number of attributes where p was 1 and q was 1)

$$SMC = (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00}) = (0+7) / (2+1+0+7) = 0.7$$

$$J = (M_{11}) / (M_{01} + M_{10} + M_{11}) = 0 / (2 + 1 + 0) = 0$$

13

Copyright ©2007-2017 The University of Texas at Arlington. All Rights Reserved.



Cosine Similarity

If d_1 and d_2 are two document 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:

$$d_1 = 3205000200$$

$$d_2 = 1000000102$$

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

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

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

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

14

Copyright ©2007-2017 The University of Texas at Arlington. All Rights Reserved.

Pearson Correlation Coefficient



❖ Correlation measures the linear relationship between objects

population correlation

$$\rho_{X,Y} = \text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y},$$

$$\begin{aligned} E[X] &= \mu, \quad \sigma = \sqrt{E[(X - \mu)^2]} \\ &= \sqrt{E[X^2] + E[-2\mu X] + E[\mu^2]} = \sqrt{E[X^2] - 2\mu E[X] + \mu^2} \\ &= \sqrt{E[X^2] - 2\mu^2 + \mu^2} = \sqrt{E[X^2] - \mu^2} \\ &= \sqrt{E[X^2] - (E[X])^2} \end{aligned}$$

covariance

sample correlation

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

$$\begin{aligned} \text{cov}(X, Y) &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY - X E[Y] - E[X] Y + E[X] E[Y]] \\ &= E[XY] - E[X] E[Y] - E[X] E[Y] + E[X] E[Y] \\ &= E[XY] - E[X] E[Y]. \end{aligned}$$

$$r = r_{xy} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}.$$

$$r = r_{xy} = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) \leftarrow \text{dot product}$$

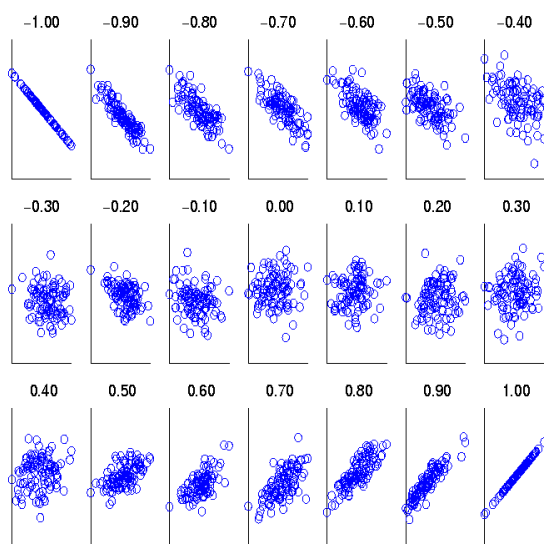
$$r = r_{xy} = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sqrt{(\sum x_i^2 - n \bar{x}^2)} \sqrt{(\sum y_i^2 - n \bar{y}^2)}}.$$

$$r = r_{xy} = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{(n-1) s_x s_y} \quad s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

15

Copyright ©2007-2017 The University of Texas at Arlington. All Rights Reserved.

Visually Evaluating Correlation



Scatter plots showing the similarity from -1 to 1 .

16

Copyright ©2007-2017 The University of Texas at Arlington. All Rights Reserved.

General Approach for Combining Similarities



- ❖ Sometimes attributes are of many different types, but an overall similarity is needed.

1. For the k^{th} attribute, compute a similarity, s_k , in the range $[0, 1]$.
2. Define an indicator variable, δ_k , for the k^{th} attribute as follows:

$$\delta_k = \begin{cases} 0 & \text{if the } k^{th} \text{ attribute is a binary asymmetric attribute and both objects have} \\ & \text{a value of 0, or if one of the objects has a missing values for the } k^{th} \text{ attribute} \\ 1 & \text{otherwise} \end{cases}$$

3. Compute the overall similarity between the two objects using the following formula:

$$similarity(p, q) = \frac{\sum_{k=1}^n \delta_k s_k}{\sum_{k=1}^n \delta_k}$$

Using Weights to Combine Similarities



- ❖ May not want to treat all attributes the same.
 - o Use weights w_k which are between 0 and 1 and sum to 1.

$$similarity(p, q) = \frac{\sum_{k=1}^n w_k \delta_k s_k}{\sum_{k=1}^n \delta_k}$$

$$distance(p, q) = \left(\sum_{k=1}^n w_k |p_k - q_k|^r \right)^{1/r}$$