

Data Similarity

Anna Monreale
Computer Science Department

Introduction to Data Mining, 2nd Edition
Chapter I

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/Dissimilarity for one Attribute

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	$d = \frac{ p-q }{n-1}$ <p>(values mapped to integers 0 to $n-1$, where n is the number of values)</p>	$s = 1 - \frac{ p-q }{n-1}$
Interval or Ratio	$d = p - q $	$s = -d, s = \frac{1}{1+d} \text{ or } s = 1 - \frac{d - \min_d}{\max_d - \min_d}$

Table 5.1. Similarity and dissimilarity for simple attributes

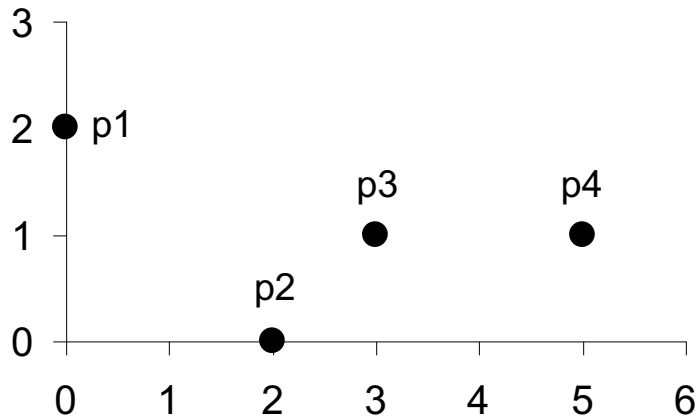
Euclidean Distance

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

where n is the number of dimensions (attributes) and x_k and y_k are, respectively, the k^{th} attributes (components) or data objects \mathbf{x} and \mathbf{y} . Standardization is necessary, if scales differ.

- Standardization is necessary, if scales differ.

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

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 a parameter, n is the number of dimensions (attributes) and x_k and y_k are, respectively, the k^{th} attributes (components) or data objects \mathbf{x} and \mathbf{y} .

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.

Minkowski Distance

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

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

Distance Matrix

Common Properties of a Distance

- Distances, such as the Euclidean distance, have some well-known properties.
 1. $d(\mathbf{x}, \mathbf{y}) \geq 0$ for all \mathbf{x} and \mathbf{y} and $d(\mathbf{x}, \mathbf{y}) = 0$ only if $\mathbf{x} = \mathbf{y}$. (Positive definiteness)
 2. $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ for all \mathbf{x} and \mathbf{y} . (Symmetry)
 3. $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$ for all points \mathbf{x} , \mathbf{y} , and \mathbf{z} . (Triangle Inequality)

where $d(\mathbf{x}, \mathbf{y})$ is the distance (dissimilarity) between points (data objects), \mathbf{x} and \mathbf{y} .

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

Common Properties of a Similarity

Similarities, also have some well-known properties.

1. $s(\mathbf{x}, \mathbf{y}) = 1$ (or maximum similarity) only if $\mathbf{x} = \mathbf{y}$.
(does not always hold, e.g., cosine)
2. $s(\mathbf{x}, \mathbf{y}) = s(\mathbf{y}, \mathbf{x})$ for all \mathbf{x} and \mathbf{y} . (Symmetry)

where $s(\mathbf{x}, \mathbf{y})$ is the similarity between points (data objects), \mathbf{x} and \mathbf{y} .

Binary Data

Categorical	insufficient	sufficient	good	very good	excellent
p1	0	0	1	0	0
p2	0	0	1	0	0
p3	1	0	0	0	0
p4	0	1	0	0	0
item	bread	butter	milk	apple	tooth-past
p1	1	1	0	1	0
p2	0	0	1	1	1
p3	1	1	1	0	0
p4	1	0	1	1	0

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)

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

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

Document Data

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

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:

$$d_1 = 3 \ 2 \ 0 \ 5 \ 0 \ 0 \ 0 \ 2 \ 0 \ 0$$

$$d_2 = 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 2$$

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

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

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

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

Using Weights to Combine Similarities

- May not want to treat all attributes the same.
 - Use non-negative weights ω_k

- $similarity(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^n \omega_k \delta_k s_k(\mathbf{x}, \mathbf{y})}{\sum_{k=1}^n \omega_k \delta_k}$

- Can also define a weighted form of distance

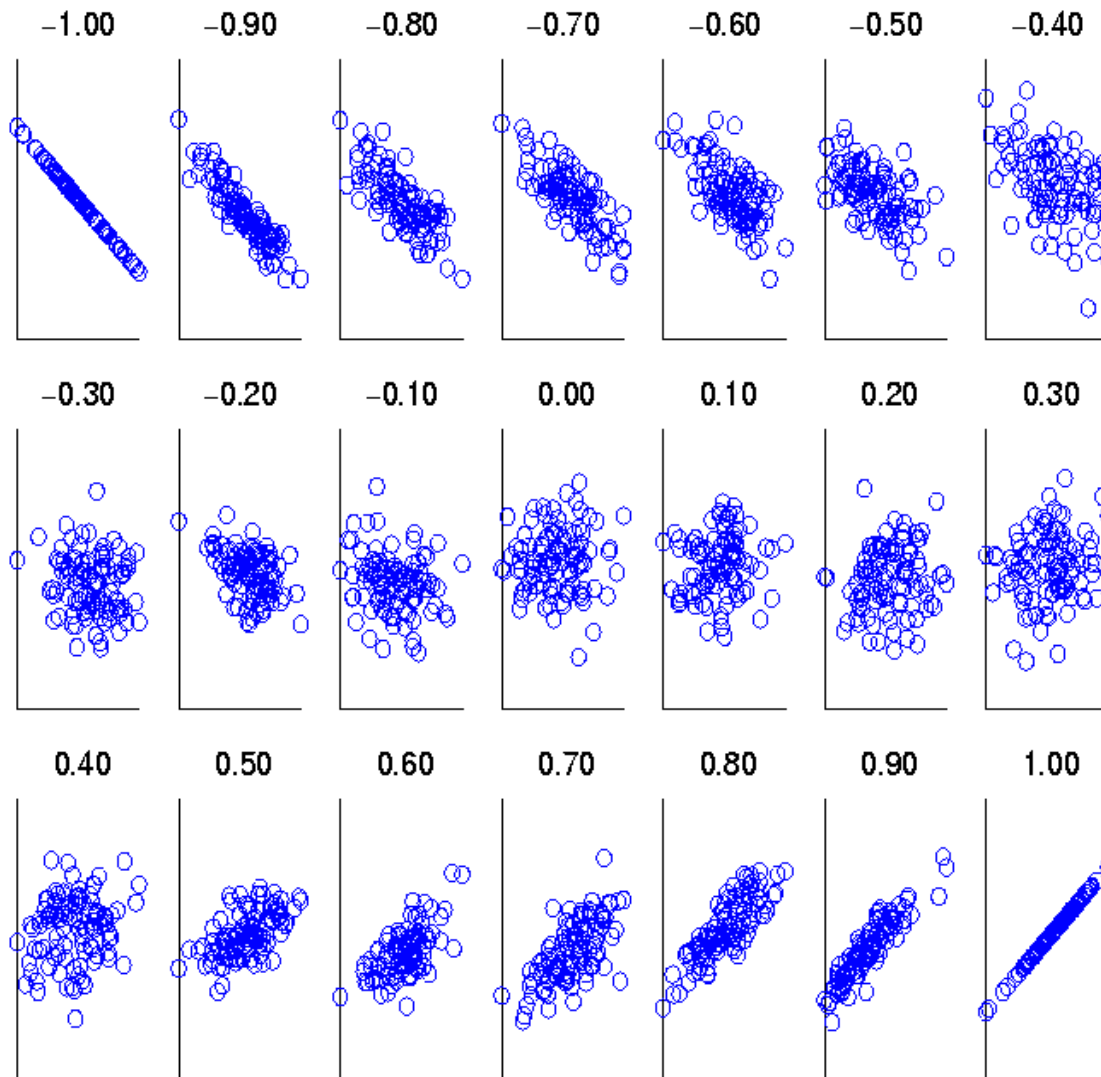
$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^n \omega_k |x_k - y_k|^r \right)^{1/r}$$

Correlation

- Correlation measures the linear relationship between objects (binary or continuous)
- To compute correlation, we standardize data objects, p and q , and then take their dot product (covariance/standard deviation)

$$\text{corr}(\mathbf{x}, \mathbf{y}) = \frac{\text{covariance}(\mathbf{x}, \mathbf{y})}{\text{standard_deviation}(\mathbf{x}) * \text{standard_deviation}(\mathbf{y})} = \frac{s_{xy}}{s_x s_y},$$

Visually Evaluating Correlation



Scatter plots showing the similarity from -1 to 1.

Information and Probability

- Information relates to possible outcomes of an event
 - transmission of a message, flip of a coin, or measurement of a piece of data



- The more certain an outcome, the less information that it contains and vice-versa
 - For example, if a coin has two heads, then an outcome of heads provides no information
 - More quantitatively, **the information is related the probability of an outcome**
 - **The smaller the probability** of an outcome, **the more information** it provides and vice-versa
 - Entropy is the commonly used measure

Entropy

- For
 - a variable (event), X ,
 - with n possible values (outcomes), $x_1, x_2 \dots, x_n$
 - each outcome having probability, $p_1, p_2 \dots, p_n$
 - the entropy of X , $H(X)$, is given by

$$H(X) = - \sum_{i=1}^n p_i \log_2 p_i$$

- Entropy is between 0 and $\log_2 n$ and is measured in bits
 - Thus, entropy is a measure of how many bits it takes to represent an observation of X on average

Entropy Examples

- For a coin with probability p of heads and probability $q = 1 - p$ of tails

$$H = -p \log_2 p - q \log_2 q$$

- For $p = 0.5, q = 0.5$ (fair coin) $H = 1$
- For $p = 1$ or $q = 1, H = 0$

Entropy for Sample Data

- Suppose we have
 - a number of observations (m) of some attribute, X , e.g., the hair color of students in the class,
 - where there are n different possible values
 - And the number of observation in the i^{th} category is m_i
 - Then, for this sample

$$H(X) = - \sum_{i=1}^n \frac{m_i}{m} \log_2 \frac{m_i}{m}$$

Mutual Information

- Information one variable provides about another

Formally, $I(X, Y) = H(X) + H(Y) - H(X, Y)$, where

$H(X, Y)$ is the joint entropy of X and Y ,

$$H(X, Y) = - \sum_i \sum_j p_{ij} \log_2 p_{ij}$$

Where p_{ij} is the probability that the i^{th} value of X and the j^{th} value of Y occur together

- For discrete variables, this is easy to compute
- Maximum mutual information for discrete variables is $\log_2(\min(n_X, n_Y))$, where n_X (n_Y) is the number of values of X (Y)

Mutual Information Example

Student Status	Count	p	$-p \log_2 p$
Undergrad	45	0.45	0.5184
Grad	55	0.55	0.4744
Total	100	1.00	0.9928

Grade	Count	p	$-p \log_2 p$
A	35	0.35	0.5301
B	50	0.50	0.5000
C	15	0.15	0.4105
Total	100	1.00	1.4406

Student Status	Grade	Count	p	$-p \log_2 p$
Undergrad	A	5	0.05	0.2161
Undergrad	B	30	0.30	0.5211
Undergrad	C	10	0.10	0.3322
Grad	A	30	0.30	0.5211
Grad	B	20	0.20	0.4644
Grad	C	5	0.05	0.2161
Total		100	1.00	2.2710

Mutual information of Student Status and Grade = $0.9928 + 1.4406 - 2.2710 = 0.1624$