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.

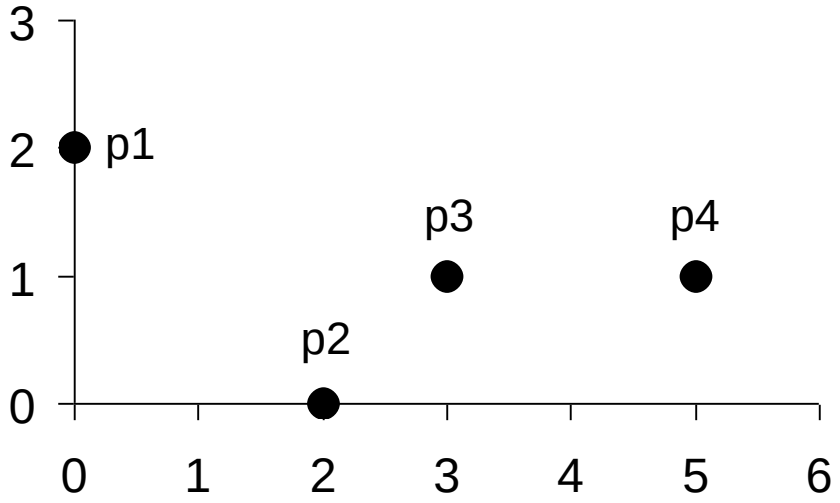
Attribute Type	Dissimilarity	Similarity
Nominal	d={ 0if x=y1if x≠y	s={ 1if x=y0if x≠y
Ordinal	d=|x−y|/(n−1) (values mapped to integers 0 to n−1, where *n* is the number of values)	s=1−d
Interval or Ratio	d=|x−y|	s=−d, s=11+d, s=e−d,s=1−d−min_dmax_d−min_d

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***

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

Common Properties of a Similarity

- Similarities, have some well known properties.
  - s(x, y) = 1 (or maximum similarity) only if x = y.
  - s(x, y) = s(y, x) for all x and y. (Symmetry)
    where s(x, y) is the similarity between points (data objects), x and y.

A Non-symmetric Similarity Measure:

- Consider an experiment in which people are asked to classify a small set of characters as they flash on a screen.
- The confusion matrix for this experiment records how often each character is classified as itself, and how often each is classified as another character.
- Using the confusion matrix, we can define a similarity measure between a character x and a character y as the number of times that x is misclassified as y, but note that this measure is not symmetric.
- For example, suppose that “0” appeared 200 times and was classified as a “0” 160 times, but as an “o” 40 times.
- Likewise, suppose that “o” appeared 200 times and was classified as an “o” 170 times, but as “0” only 30 times.
- Then, s(0,o)=40, but s(o, 0)=30. In such situations, the similarity measure can be made symmetric by setting: s’(x,y)=s’(y,x)=(s(x,y)+s(y,x))/2 where s’ indicates the new similarity measure.

Similarity Measures for Binary Data

- Similarity measures between objects that contain only binary attributes are called **similarity coefficient**s, and typically have values between 0 and 1.
- A value of 1 indicates that the two objects are completely similar, while a value of 0 indicates that the objects are not at all similar.
- There are many rationales for why one coefficient is better than another in specific instances.
- Let x and y be two objects that consist of n binary attributes.
- The comparison of two such objects, i.e., two binary vectors, leads to the following four quantities (frequencies):

  f01 = the number of attributes where x is 0 and y is 1
  f10 = the number of attributes where x is 1 and y is 0
  f00 = the number of attributes where x is 0 and y is 0
  f11 = the number of attributes where x is 1 and y is 1

Simple Matching Coefficient

- One commonly used similarity coefficient is the simple matching coefficient (SMC), which is defined as –

  SMC = number of matches / number of attributes

  = (f11 + f00) / (f01 + f10 + f11 + f00)

- Example:

  x = 1 0 0 0 0 0 0 0 0 0

  y = 0 0 0 0 0 0 1 0 0 1

  f01 = 2 (the number of attributes where x was 0 and y was 1)

  f10 = 1 (the number of attributes where x was 1 and y was 0)

  f00 = 7 (the number of attributes where x was 0 and y was 0)

  f11 = 0 (the number of attributes where x was 1 and y was 1)

  SMC = (f11 + f00) / (f01 + f10 + f11 + f00) = (0+7) / (2+1+0+7) = 0.7

Jaccard Coefficient

- Suppose that x and y are data objects that represent two rows (two transactions) of a transaction matrix.
- If each asymmetric binary attribute corresponds to an item in a store, then a 1 indicates that the item was purchased, while a 0 indicates that the product was not purchased.
- Because the number of products not purchased by any customer far outnumbers the number of products that were purchased, a similarity measure such as SMC would say that all transactions are very similar.
- As a result, the Jaccard coefficient is frequently used to handle objects consisting of asymmetric binary attributes.
- The Jaccard coefficient , which is often symbolized by J, is given by the following equation:

  J = number of 11 matches / number of non-zero attributes

  = (f11) / (f01 + f10 + f11)
- Previous example: x = 1 0 0 0 0 0 0 0 0 0, y = 0 0 0 0 0 0 1 0 0 1

  f01 = 2, f10 = 1, f00 = 7, f11 = 0

  J = (f11) / (f01 + f10 + f11) = 0 / (2 + 1 + 0) = 0

Cosine Similarity

- Documents are often represented as vectors, where each component (attribute) represents the frequency with which a particular term (word) occurs in the document.
- Even though documents have thousands or tens of thousands of attributes (terms), each document is sparse since it has relatively few nonzero attributes.
- Thus, as with transaction data, similarity should not depend on the number of shared 0 values because any two documents are likely to “not contain” many of the same words, and therefore, if 0–0 matches are counted, most documents will be highly similar to most other documents.
- Therefore, a similarity measure for documents needs to ignores 0–0 matches like the Jaccard measure, but also must be able to handle non-binary vectors.
- The cosine similarity is one of the most common measures of document similarity.
- If x and y are two document vectors, then –

  cos(d1, d2) = <d1,d2> / ||d1|| ||d2||

  where <d1,d2> indicates inner product or vector dot product of vectors, d1 and d2, and || d || is the length of vector d.

Cosine Similarity

- Example:

  d1 = 3 2 0 5 0 0 0 2 0 0

  d2 = 1 0 0 0 0 0 0 1 0 2

  <d1,d2> = 3*1+2*0+0*0+5*0+0*0+0*0+0*0+2*1+0*0+0*2 = 5

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

  $= (42)^{\mathbf{0.5}} = 6.481$

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

  $= (6)^{\mathbf{0.5}} = 2.449$

  $\cos(\mathbf{d_1}, \mathbf{d_2}) = 0.3150$

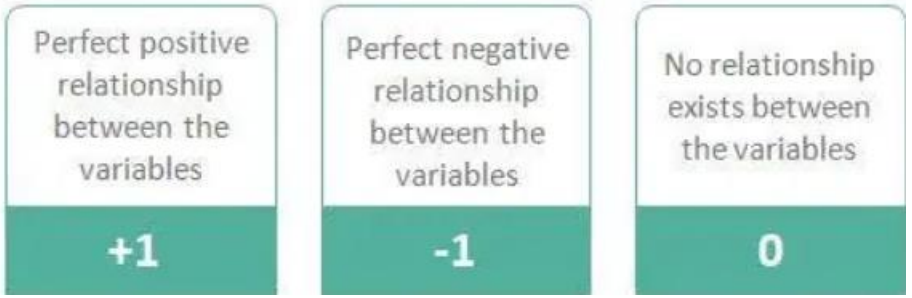
Extended Jaccard Coefficient

- Also known as Extended Tanimoto coefficient.
- The extended Jaccard coefficient can be used for document data and that reduces to the Jaccard coefficient in the case of binary attributes.
- This coefficient, which we shall represent as EJ, is defined by the following equation:

$$EJ(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \mathbf{x} \cdot \mathbf{y}}$$

Correlation

- Correlation is frequently used to measure the linear relationship between two sets of values that are observed together.
- Thus, correlation can measure the relationship between two variables (height and weight) or between two objects (a pair of temperature time series).
- Correlation is used much more frequently to measure the similarity between attributes since the values in two data objects come from different attributes, which can have very different attribute types and scales.
- There are many types of correlation, and indeed correlation is sometimes used in a general sense to mean the relationship between two sets of values that are observed together.
- In this discussion, we will focus on a measure appropriate for numerical values.



Correlation

$$\text{corr}(\mathbf{x}, \mathbf{y}) = \frac{\text{covariance}(\mathbf{x}, \mathbf{y})}{\text{standard\_deviation}(\mathbf{x}) * \text{standard\_deviation}(\mathbf{y})} = \frac{s_{xy}}{s_x \; s_y}, \quad (2.11$$

where we are using the following standard statistical notation and definitions

$$\text{covariance}(\mathbf{x}, \mathbf{y}) = s_{xy} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})(y_k - \overline{y}) \quad (2.12$$

$$\text{standard\_deviation}(\mathbf{x}) = s_x = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})^2}$$

$$\text{standard\_deviation}(\mathbf{y}) = s_y = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (y_k - \overline{y})^2}$$

$$\overline{x} = \frac{1}{n} \sum_{k=1}^{n} x_k \text{ is the mean of } \mathbf{x}$$

$$\overline{y} = \frac{1}{n} \sum_{k=1}^{n} y_k \text{ is the mean of } \mathbf{y}$$

Pearson's Correlation

- Pearson's correlation between two sets of numerical values, i.e., two vectors, x and y, is defined by the following equation:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

- **Find out the number of pairs of variables denoted by n. Suppose x consists of 3 variables – 6, 8, 10. Suppose y consists of corresponding three variables: 12, 10, and 20.**

x	y	x*y	x^2	y^2
6	12	72	36	144
8	10	80	64	100
10	20	200	100	400
24	**42**	**352**	**200**	**644**

r = 3*352-24*42 / √(3*200-24^2)*(3*644-42^2)

= 0.7559

Correlation vs Cosine vs Euclidean Distance

- Compare the three proximity measures according to their behavior under variable transformation:;
  - Scaling: multiplication by a value ($\mathbf{y_s = y * 2}$)
  - Translation: adding a constant ($\mathbf{y_t = y + 5}$)
- Consider the example: x = (1, 2, 4, 3, 0, 0, 0), y = (1, 2, 3, 4, 0, 0, 0)
  and, ys = y * 2 (scaled version of y), yt = y + 5 (translated version).

Measure	(*x* , *y*)	(*x* , y_s)	(*x* , y_t)
Cosine	0.9667	0.9667	0.7940
Correlation	0.9429	0.9429	0.9429
Euclidean Distance	1.4142	5.8310	14.2127

Property	Cosine	Correlation	Euclidean Distance
Invariant to scaling (multiplication)	Yes	Yes	No
Invariant to translation (addition)	No	Yes	No

Correlation vs Cosine vs Euclidean Distance

- Choice of the right proximity measure depends on the domain.
- What is the correct choice of proximity measure for the following situations?

1. Comparing documents using the frequencies of words – cosine.
   - Documents are considered similar if the word frequencies are similar
2. Comparing the temperature in Celsius of two locations – correlation.
   - Two locations are considered similar if the temperatures are similar in magnitude.
3. Comparing amount of precipitation (in cm) in several locations - Euclidean Distance.

Practice Problem

- Exercise: 2, 3, 6, 7, 8, 14, 18, 19, 20 (CH-02).

Thank You!