## Similarity Measures

Similarity and dissimilarity are important because they are used by a number of data mining techniques, such as clustering nearest neighbor classification and anomaly detection.

The term **proximity** is used to refer to either similarity or dissimilarity.

Definitions:

The **similarity** between two objects is a numeral measure of the degree to which the two objects are alike. Consequently, similarities are *higher* for pairs of objects that are more alike. Similarities are usually non-negative and are often between 0 (no similarity) and 1(complete similarity).

The **dissimilarity** between two objects is the numerical measure of the degree to which the two objects are different. Dissimilarity is *lower* for more similar pairs of objects.

Frequently, the term **distance** is used as a synonym for dissimilarity. Dissimilarities sometimes fall in the interval [0,1], but it is also common for them to range from 0 to $\infty$

**Proximity Measures:**

Proximity measures, especially similarities, are defined to have values in the interval [0,1]. If the similarity between objects can range from 1 (not at all similar) to 10 (completely similar), we can make them fall into the range [0,1] by using the formula:

s’=(s-1)/9, where s and s’ are the original and the new similarity values, respectively.

The more general case, s’ is calculated as

s’=(s-min_s)/(max_s-min_s), where min_s and max_s are the minimum and maximum similarity values respectively.

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).

If the proximity measure originally takes values in the interval [0, ∞], then we usually use the formula: d’= d/(1+d) for such cases and bring the dissimilarity measure between [0,1].

**<u>Similarity and dissimilarity between simple attributes:</u>**

The proximity of objects with a number of attributes is defined by combining the proximities of individual attributes.

-Attribute Types and Similarity Measures:

1) For interval or ratio attributes, the natural measure of dissimilarity between two attributes is the absolute difference of their values. For example, we might compare our current weight to our weight one year ago. In such cases the dissimilarities range from 0 to ∞.

2) For objects described with one nominal attribute, the attribute value describes whether the attribute is present in the object or not. Comparing two objects with one nominal attribute means comparing the values of this attribute. In that case, similarity is

traditionally defined as 1 if attribute values match and as 0 otherwise. A dissimilarity would be defined in the opposite way: 0 if the attribute values match, 1 if they do not.

3) For objects with a single ordinal attribute, information about order should be taken into account. Consider an attribute that measures the quality of a product on the scale {*poor, fair, OK, good, wonderful*}. It would be reasonable that a product P1 which was rated *wonderful* would be closer to a product P2 rated *good* rather than a product P3 rated *OK.*

To make this observation quantitative, the values of the ordinal attribute are often mapped to successive integers, beginning at 0 or 1, e.g.{*poor=0, fair=1, OK=2, good=3, wonderful=4*}.

Then d(P1-P2) =4-3 =1

In the table below, x and y are two objects that have one attribute of the indicated type, and d(x,y) and s(x,y) are the dissimilarity and similarity between x and y, respectively.

| Attribute Type | Dissimilarity | Similarity |
|---|---|---|
| Interval or ratio | d=$\lvert$x-y$\rvert$ | s=-d, $s=\frac{1}{1+d}$,<br>s=$e^{-d}$,<br>s=1-$\frac{d-\min\_d}{\max\_d-\min\_d}$ |
| Nominal | d=$\begin{cases}0\ if\ x=y\\1\ if\ x\ \neq y\end{cases}$ | s= $\begin{cases}1\ if\ x=y\\0\ if\ x\ \neq y\end{cases}$ |
| Ordinal | d= $\frac{\lvert x-y\rvert}{(n-1)}$<br>(values mapped to integers 0 to n-1 where n is the number of values) | s= 1-d |

Dissimilarities between Data Objects:

Distances:

Distances are dissimilarities with certain properties. The **Euclidian distance,** *d,* between two points , x and y in one , two or higher dimensional space is given by the formula:

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

where n is the number of dimensions and $x_k$ and $y_k$ are, respectively, the k$^{th}$ attribute (component) of *x* and *y.*

The Euclidian distance measure is given generalized by the Minkowski distance metric shown as:

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

where r is a parameter.

The following are the 3 most common examples of Minkowski distances:

- r = 1 also known as City block (Manhattan or L1 norm) distance. A common example is the **Hamming distance**, which is the number of bits that are different between two objects that only have binary attributes (i.e., binary vectors)
- r=2. Euclidian distance (L2 norm).
- r= ∞. Supremum, ($L_{max}$ or $L_\infty$ norm) distance. This is the maximum difference between any attributes of the objects. The $L_\infty$ is defined more formally by:

  d(x, y)= $\lim_{r\to\infty}(\sum_{k=1}^{n} |x_k - y_k|^r)^{1/r}$

  *Note: r should not be confused with the number of dimensions or attributes n.*

| Point | x coordinate | y coordinate |
|---|---|---|
| p1 | 0 | 2 |
| p2 | 1 | 0 |
| p3 | 2 | 1 |

L2 distance:

| | **p1** | **p2** | **p3** |
|---|---|---|---|
| **p1** | **0** | $\sqrt{5}$ | $\sqrt{5}$ |
| **p2** | $\sqrt{5}$ | **0** | $\sqrt{2}$ |
| **p3** | $\sqrt{5}$ | $\sqrt{2}$ | **0** |

L1 distance:

| | **p1** | **p2** | **p3** |
|---|---|---|---|
| **p1** | **0** | | **3** |
| **p2** | **3** | **0** | **2** |
| **p3** | **3** | **2** | **0** |

## Similarities between Data Objects:

s(x, y) is the similarity between points x and y, then typically we will have

1. s(x, y) =1 only if x=y. ($0 \le s \le 1$)
2. s(x, y) = s (y, x) for all x and y. (Symmetry)

## Non-symmetric Similarity Measures – confusion matrix-

Consider an experiment in which people are asked to classify a small set of characters as they flash on the screen. The **confusion matrix** for this experiment records how often each character is classified as itself, and how often it is classified as another character. For example, suppose “0” appeared 200 times and was classified as “0” 160 times but as “o” 40 times. Likewise, suppose that “o” appeared 200 times and was classified as “o” 170 times and as “0” 30 times.

If we take these counts as a measure of similarity between the two characters, then we have a similarity measure, but **<u>not a symmetric one.</u>**

s(“0”, “o”) =40/2 = 20%

s(“o”, “0”) = 30/2 = 15%

In such situations, the similarity measure is often made symmetric by setting

s′(x,y) = s′(y,x) = (s(x,y)+ (s(y,x))/2

s′(“0”, “o”)= s′(“o”, “0”)= (20+15)/2 = 17.5%

## <u>Similarity Measures for Binary Data</u>

Similarity measures between objects that contain only binary attributes are called **similarity coefficients**, 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.

Let **x** and **y** be two objects that consist of n binary attributes. The comparison of 2 such objects, i.e. two binary vectors, leads to the following four quantities (frequencies):

$f_{00}$= the number of attributes where x is 0 and y is 0

$f_{01}$= the number of attributes where x is 0 and y is 1

$f_{10}$= the number of attributes where x is 1 and y is 0

$f_{11}$= the number of attributes where x is 1and y is 1

**Simple Matching Coefficient** (SMC) One commonly used similarity coefficient is defined as:

$$SMC = \frac{number\ of\ matching\ attribute\ values}{number\ of\ attributes}$$
$$= \frac{f_{11} + f_{00}}{f_{11} + f_{00} + f_{10} + f_{01}}$$

This measure counts both presences and absences equally.

**Jaccard Coefficient** In some situations, only the presence of an item is relevant, supposed we want to do a basket market study, if we were to count all items that a person did not buy, their number will outnumber the items purchased by far and it can compromise our study. The **Jaccard coefficient (J)** is often used in this type of situation in which we have asymmetric binary attributes.

$$J = \frac{number\ of\ matching\ presences}{number\ of\ attributes\ not\ involed\ in\ 00\ matches}$$
$$= \frac{f_{11}}{f_{11} + f_{10} + f_{01}}$$

## <u>Cosine Similarity:</u>

Documents are often represented as vectors, in which each attribute represents the frequency with which a particular term (word) occurs in the document. Even though documents may have thousands or tens of thousands of words, each document is sparse since it has few non zero attributes. Therefore, a similarity measure for documents needs to ignore 0-0 matches like he Jaccard measure, but must handle non-binary vectors. The cosine similarity is the most common measure of document similarity.

if x and y are 2 documents, then:

cos(x,y)=$\frac{x.y}{||x||||y||'}$

where **.** indicates the vector **.** product*, x***.***y* = $\sum_{k=1}^{n} x_k . y_k$

and ||x|| is the length of the vector x, ||x||= $\sqrt{\sum_{k=1}^{n} {x_k}^2}$

Cosine similarity is a measure of the (cosine of the) angle between x and y. Thus if the cosine similarity is 1, the angle between x and y is 0 and x and y are the same except for magnitude.

If the cosine similarity is 0, then the angle between x and y is 90, then they do not share any terms (words).

## Correlations:

The correlation between two data objects that have binary or continuous variables is a measure of the linear relationship between the attributes of the objects. More precisely, **Pearson's correlation** coefficient between two data objects , x and y, is defined by the following equation:

*corr(x,y)*=$\frac{covariance(x,y)}{standard\ deviation(x)*standard\ deviation(y)}=\frac{s_{xy}}{s_x s_y}$

where the standard statistical notations are used.

*covariance(x,y)* =$s_{xy}$= $\frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{x})(y_k-\bar{y})$

standard deviation(x) = $s_x$= $\sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{x})^2}$

standard deviation(y) = $s_y$= $\sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(y_k-\bar{y})^2}$

$\bar{x}=\frac{1}{n}\sum_{k=1}^{n}x_k$   is the mean of *x (average)*

$\bar{y}=\frac{1}{n}\sum_{k=1}^{n}y_k$   is the mean of *y (average)*

**Perfect correlation:**

Correlation is always in the range of -1 to 1. A correlation of 1 (-1) means that x and y have a perfect positive (negative) linear relationship; that is

$x_k$= $ay_k$+b

The following two sets of values for x and y indicate cases where the correlation is -1 and +1

1)x = (-3, 6, 0, 3, -6)  y= (1, -2, 0, -1, 2)

2)x= (3,6, 0, 3, 6)  y= (1, 2,0,1,2)

Correlation measures are only useful if/when the relationship between attributes is linear. So if the correlation is 0, then there is no linear relationship between the two data objects. However, a non-linear relationship may still exist. For example

x=(-3, -2, -1, 0, 1, 2, 3)

y=(9, 4, 1, 0,1, 4, 9)

The relationship between x and y is ???

[redacted]

Correlation =0