Measuring Data similarity and dissimilarity :-

In data mining applications, such as clustering, outlier analysis, and nearest-neighbour classification, we need ways to assess how alike or unlike objects are in comparision to one another. For Example, a store may want to search for clusters of customer objects, resulting in groups of customers with similar characteristic (eg. similar income, area of residence and age). Such information can then be used for marketing.

A cluster is a collection of data objects such that the objects within a cluster are similar to one another and dissimilar to the objects in other clusters.

Outlier analysis also employs clustering based techniques to identify potential outliers as objects that are highly dissimilar to others. Knowledge of object similarities can also be used in nearest-neighbour classification schemes where a given object (eg. a patient) is assigned a class label (relating to say, a diagnosis) based on its similarity toward other objects in the model.

The measurements of similarity and dissimilarity are referred to as measure of proximity. Similarity and dissimilarity are related. A similarity measure for two objects, i and j, will typically return the value 0 if the objects are unlike. The higher the similarity value, the greater the similarity b/w objects. Typically, a value of 1 indicates complete similarity, that is the object are identical.

A dissimarity measure works in the opposite way. It returns a value of 0 if the objects are the same. The higher the dissimilarity value, the more dissimilar the two objects are.

Dissimilarity of Numeric Data: Minkowski Distance:→

We describe distance measures that are commonly used for computing the dissimilarity of objects described by numeric attributes. These measure include the Euclidean, Manhattan, and Minkowski distance.

The most popular distance measure is Euclidean distance (i.e. straight line). Let $i = (x_{i1}, x_{i2} \dots x_{ip})$ and $j = (x_{j1}, x_{j2}, \dots x_{jp})$ be two objects described by p numeric attributes. The Euclidean distance between objects i and j is defined as

$$d(i,j) = \sqrt{(x_{i1}-x_{j1})^2 + (x_{i2}-x_{j2})^2 + \dots + (x_{ip}-x_{jp})^2}$$

Another well known distance measure is the Manhattan (or city block) distance, named so because it is the distance in blocks b/w any two points in a city (such as 2 blocks down and 3 blocks over for a total of 5 blocks). It is defined as

$$d(i,j) = |x_{i1}-x_{j1}| + |x_{i2}-x_{j2}| + \dots + |x_{ip}-x_{jp}|.$$

Both the euclidean and the manhattan distance satisfy the following mathematical properties.

* Non-negativity : $d(i,j) \geq 0$; distance is non-negative number.
* Identity of indiscernibles : $d(i,i) = 0$; The distance of an object to itself is 0.
* Symmetry $d(i,j) = d(j,i)$; Distance is a symmetric function
* Triangle inequality :- $d(i,j) \leq d(i,k) + d(k,j)$: Going directly from object i to object j in space is no more than making a detour over any other object k.

Euclidean distance and Manhattan distance:→

Let $x_1 = (1,2)$ and $x_2 = (3,5)$ represent two objects as shown in fig. The Euclidean distance between the two is $\sqrt{2^2+3^2} = 3.61$

The Manhattan distance between the two is $2+3=5$

Minkowski Distance is a generalization of the Euclidean and Manhattan distances.

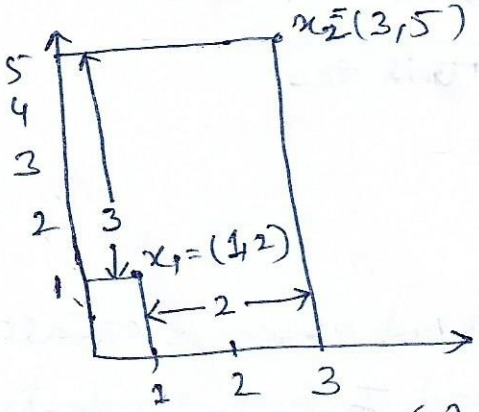
it is defined as

$$d(i,j) = \sqrt[h]{|x_{i1}-x_{j1}|^h + |x_{i2}-x_{j2}|^h + \cdots + |x_{ip}-x_{jp}|^h}$$

where h is a real number such that $h \geq 1$. It represents the Manhattan distance when $h=1$ and Euclidean distance when $h=2$.

* The supremum distance (also called Chebyshev distance) is a generalization of the Minkowski distance for $h \to \infty$. To compute it, we find the attribute f that gives the maximum difference in values b/w the two objects. This difference is supremum distance, defined more formally as:

$$d(i,j) = \lim_{h\to\infty} \left(\sum_{f=1}^{p} |x_{if}-x_{jf}|^h\right)^{1/h} = \max_{f}^{p} |x_{if}-x_{jf}|.$$



Euclidean Distance $\Rightarrow (2^2+3^2)^{1/2} = 3.61$

Manhattan Distance $\Rightarrow (2+3) = 5$

* for the same example, where $x_1 = (1,2)$ and $x_2 = (3,5)$, as in figure, the second attribute gives the greatest difference b/w values for the objects, which is $5-2=3$. this is the supremum distance b/w objects.

Example of Proximity Measures:→

Similarity Measures for Binary Data:→

Similarity measures between objects that contain only binary attributes are called similarity coefficient, and typically have value b/w 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 ~~relationship~~ 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 comparision of two such objects i.e. two binary vector, 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 1 and y is 1

Simple matching coefficient :– one commonly used similarity coefficient is the simple matching (SMC), which is defined as

$$SMC = \frac{\text{number of matching attribute values}}{\text{number of attributes}}$$

$$= \frac{f_{11} + f_{00}}{f_{01} + f_{10} + f_{11} + f_{00}}$$

This measure counts both presences and absences equally. Consequently, the SMC could be used to find students who had answered questions similarly on a test that consisted only of true/false questions.

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.

Since 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 transaction 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

$$J = \frac{\text{number of matching presences}}{\text{number of attributes not involved in 00 matches}}$$

$$= \frac{f_{11}}{f_{01} + f_{10} + f_{11}}$$

Ex:- To illustrate the difference b/w these two similarity measures, we calculate SMC and J for the following two binary vectors

$x = (1, 0, 0, 0, 0, 0, 0, 0, 0, 0)$

$y = (0, 0, 0, 0, 0, 0, 1, 0, 0, 1)$

$f_{01} = 2$; $f_{10} = 1$, $f_{00} = 7$, $f_{11} = 0$

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

$$J = \frac{f_{11}}{f_{01} + f_{10} + f_{11}} = \frac{0}{2+1+0} = 0$$

Cosine Similarity :→ Documents are often represented as vector, where each attribute represents the frequency with which a particular term (word) occurs in the documents

Even though documents ~~is~~ ~~sparse~~ have thousands or tens of thousands of attributes (terms), each document is sparse since it ~~has~~ ~~few~~ relatively few non-zero attributes.

For Ex :- Two term-frequency vectors may have many 0 values in common, meaning that the corresponding documents do not share many words, but this does not make them similar.

We need a measure that will focus on the words that the two documents do have in common, and the occurance frequency of such words. In other words, we need a measure for numeric data that ignores zero-matches.

Cosine-Similarity :- it is a measure of similarity that can be used to compare documents or, say, give a ranking of documents with respect to a given vector of query words. Let x and y be two vectors for comparison. using the cosine measure as a similarity function, we have

$$Sim(x,y) = \frac{x \cdot y}{\|x\| \|y\|}$$

where $\|x\|$ is the euclidean norm of vector $x = (x_1, x_2 \cdots x_p)$ defined as $\sqrt{x_1^2 + x_2^2 + \cdots + x_p^2}$. Conceptually, it is the length of the vector. Similarly, $\|y\|$ is the Euclidean norm of vector y.

The measure computes the cosine of the angle between vector x and y. A cosine value of 0 means that the two vectors are at 90 degrees to each other (orthogonal) and have no match. The closer the cosine value to 1, the smaller the angle(θ) and greater the match b/w vectors.

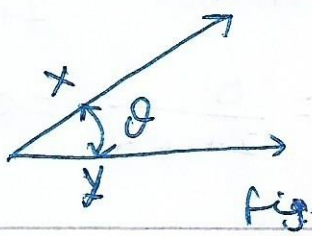


Fig- Geometric illustration of the cosine measure

Cosine similarity between Two term frequency vectors :-

Suppose that x and y are the first Two term-frequency vectors. that is $x = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)$ and

$y = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$. How similar are x and y? using eq to compute the Cosine Similarity b/w the Two vectors, we get:

$$x \cdot y = (5\times3 + 0\times0 + 3\times2 + 0\times0 + 2\times1 + 0\times1 + 0\times0 + 2\times1 + 0\times0 + 0\times1)$$

$$\Rightarrow (15 + 6 + 2 + 2) = 25$$

$$||x|| = \sqrt{5^2+0^2+3^2+0^2+2^2+0^2+0^2+2^2+0^2+0^2} \Rightarrow 6.48$$

$$||y|| = \sqrt{3^2+2^2+1^2+1^2+1^2+1^2} \Rightarrow 4.12$$

$$Sim(x,y) = \frac{25}{6.48 \times 4.12} \Rightarrow \underline{0.94} \quad \underline{\text{Ans}}$$

Therefore, if we were using the Cosine similarity measure to compare these documents, they would be considered quite similar.

Correlation:→ The correlation b/w two data objects that have binary or continuous variables is a measure of linear relationship between the attributes of the objects. (The calculation of correlation between attributes, which is more common, can be defined similarly). More precisely, Pearson's correlation coefficient between two data objects, x and y is defined by the following equation:

$$\text{Corr}(x,y) = r_{x,y} = \frac{\text{Covariance}(x,y)}{\text{Standard\_deviation}(x) * \text{Standard\_dev}(y)}$$

$$\rightarrow \frac{S_{xy}}{S_x \cdot S_y}$$

where we are using the following standard statistical notation and definitions:

$$\text{Cov}(x,y) = S_{x,y} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})(y_i - \bar{y})$$

$$\text{Std\_dev}(x) = S_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\text{Std\_dev}(y) = S_y = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2}$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \text{ is the mean of } x$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \text{ is the mean of } y.$$

Example (Perfect correlation):- Correlation is always in the range -1 to 1. A correlation of 1 means that x and y have a Perfect Positive linear relationship. A correlation of -1 means that x and y have a perfect negative linear relationship; that is $x_i = ay_i + b$, where a and b are constants.

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

In the first case, the means of x and y were chosen to be 0, for simplicity

(1) $x = (-3, 6, 0, 3, -6)$

$y = (1, -2, 0, -1, 2)$

$\bar{x} = 0, \quad \bar{y} = 0, \quad S_x = \sqrt{\frac{1}{4}\left[(-3)^2 + (6)^2 + (3)^2 + (-6)^2\right]}$

$\rightarrow \sqrt{\frac{1}{4}(9+36+9+36)} \rightarrow \sqrt{\frac{1}{4} \times 90}$

$S_y = \sqrt{\frac{1}{4}(1+4+1+4)} \rightarrow \sqrt{\frac{1}{4} \times 10}$

$$Cov(x,y) = \frac{1}{4}\left[\begin{array}{l}(-3-0)(1-0) + (6-0)(-2-0) + (3-0)(0-0) \\ + (3-0)(-1-0) + (-6-0)(2-0)\end{array}\right]$$

$$\rightarrow \frac{1}{4}\left[-3 - 12 - 3 - 12\right] \rightarrow \frac{1}{4}\left[-30\right]$$

$$Corr(x,y) = \frac{1/4 \times -30}{\sqrt{\frac{1}{4} \times 90}\sqrt{\frac{1}{4} \times 10}} \rightarrow \frac{-30}{\sqrt{90 \times 10}} \rightarrow \frac{-30}{30} \rightarrow -1$$

$x = (3, 6, 0, 3, -6)$

$y = (1, 2, 0, 1, 2)$

$Corr(x,y) = 1$

Ex: (Non-linear relationship):-

if the correlation is 0, then there is no linear relationship between the attributes of the two data objects. However, non-linear relationship may still exists. In the following example, $x_k = y_k^2$, but their correlation is 0

$$X = (-3,-2,-1,0,1,2,3)$$
$$Y = (9,4,1,0,1,4,9)$$

Sol":-

$$\bar{x} = 0 \; ; \; \bar{y} = 0$$

$$S_x = \sqrt{\frac{1}{5}(9+4+1+1+4+9)} = \sqrt{\frac{1}{5} \times 28}$$

$$S_y = \sqrt{\frac{1}{5}(81+16+1+0+1+16+81} = \sqrt{\frac{1}{5} \times 196}$$

$$Cov(x,y) = \frac{1}{5}\Big[(-3-0)(9-0) + (-2-0)(4-0) + (-1-0)(1-0) + (0-0)(0-0) + (1-0)(1-0) + (2-0)(4-0) + (3-0)(9-0)\Big]$$

$$= \frac{1}{5}\Big[-27-8-1+0+1+8+27\Big] = 0$$

so that $Cov(x,y) = r_{x,y} = 0 / \sqrt{1/5 \times 28} \cdot \sqrt{1/5 \times 196}$

$Corr(x,y) = 0$

Ans

Table:- A Sample Data Table contains Attributes of mixed Type

object identifier	test-1 (nominal)	test-2 (ordinal)	test-3 (numeric)	test-4 (binary)	test-5 (continuous)
1	code A	excellent	45	0	1.193
2	code B	fair	22	1	2.013
3	code C	good	64	1	1.231
4	code A	excellent	28	0	0.219

General characteristics of Data set:→ We discuss three characteristics that apply to many data set and have a significant impact on the data mining techniques that are used:

* Dimensionality * Sparsity, and * Resolution.

*Dimensionality:- The Dimensionality of a data set is the number of attributes that the objects in the data set possess. Data with a small number of dimensions tends to be qualitatively different than moderate or high dimensional data.

Indeed, the difficulties associated with analyzing high Dimensional data are sometimes referred to as the curse of Dimensionality. Because of this, an important motivation in preprocessing of data is dimensionality reduction.

* Sparsity:→ For some data sets, such as those with asymmetric features, most attributes of an object have values of 0; in many cases, fewer than 1% of the entries are non-zero.

In Practical terms, sparsity is an advantage because usually only the non-zero values need to be stored and manipulated. this results in significant savings with respect to computation time and storage.

Furthermore, some data mining algorithms work well only for sparse data.

Resolution: →

It is frequently possible to obtain data at different level of resolution, and often the properties of the data are different at different resolutions.

For instance, the surface of the earth seems very uneven at a resolution of few meters, but is relatively smooth at a resolution of tens of kilometers. The patterns in the data also depend on the level of resolution. If the resolution is too fine, a pattern may not be visible or may be buried in noise; If the resolution is too coarse (मुरदुरा), the pattern may disappear.

For Ex:- variations in atmospheric pressure on a scale of hours reflect the movement of storms and other weather systems. On a scale of months, such phenomena are not detectable.