A Tutorial on Distance and Similarity

Abstract – This is an introductory tutorial on distance and similarity measures. In information retrieval, a distance is a metric that denotes dissimilarity or lack of resemblance while similarity is a measure of resemblance.

Keywords: distance, similarity, distance metric, triangular inequality, symmetry, reflexivity, distance-similarity transformations, binary data sets

Published: 02-13-2015; Updated: 09-11-2016

Introduction

This tutorial was written as a companion for two of our tools (Garcia, 2015a; 2015b). These were developed to simplify distance (*D*) and similarity (*S*) calculations.

In data mining and information retrieval, these are considered association measures where distance is lack of similarity and similarity is resemblance. Some authors prefer to use the term 'dissimilarity' instead of distance.

What is distance?

Distance is a metric. A function f is said to be a metric if it exhibits reflexivity, symmetry, and triangular inequality. Consider three points a, b, and c describing a triangle in a two-dimensional space.

- **Reflexivity** means that the distance from a point to itself is zero; e.g., f(a, a) = f(b, b) = f(c, c) = 0.
- **Symmetry** refers to the fact that the distance between any two points, measured from either one, is the same; e.g., f(a, b) = f(b, a).
- **Triangular inequality** means that the distance between any two points is equal or less than the distance between these measured through a third point; e.g., f(a, b) + f(b, c) => f(a, c).

If these conditions are not met, the function in question is not a metric. See Figure 1.

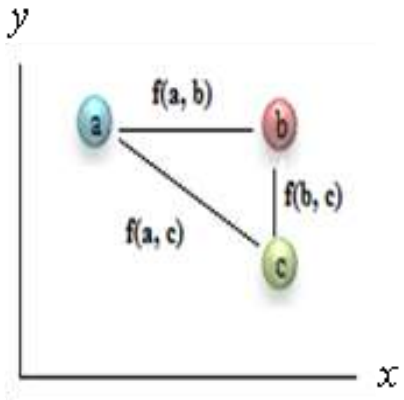

Figure 1. Three points in a two-dimensional space showing the properties of reflexivity, symmetry, and triangular inequality.

Note from the figure that distances cannot be negative and are not naturally upper bounded. Finally, we can arithmetically average, add, or subtract distances to compute new distances.

What is Similarity?

Similarity is a measure of the resemblance between data sets; i.e., how similar or alike the sets are. Although similarities are symmetric, they are not metrics. They can be negative and upper or lower bounded. For example, Hamann, Yule, and Pearson's Phi adopt values between -1 and +1. In addition, the similarity of a point or data set to itself is 1, $S_{ii} = 1$.

Similarities cannot be arithmetically averaged, added, or subtracted to compute new similarities. However, they can be rescaled to improve comparisons. For instance, some binary similarity measures can adopt values outside the [0, 1] range. These can be rescaled to said range and then transformed into distances using the procedure described by Todeschini, et al. (2012).

In Information Retrieval, the best known similarity measure is the Cosine Similarity. This similarity measure is computed by representing data sets as vectors in a term space and then comparing the angles formed by the vectors. Figure 2 shows that as the angle between any two vectors decreases their cosine similarity increases.

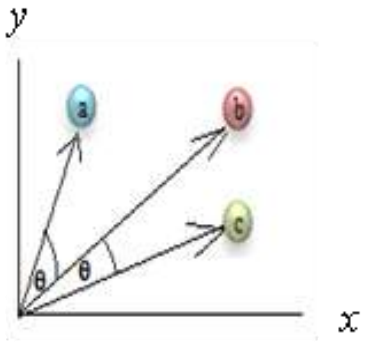

Figure 2. Cosine similarities given by the angles between vectors representing data sets. The axes represent two arbitrary term dimensions.

For unit vectors, those whose lengths have been normalized to 1, it can be easily demonstrated that their dot product is the same as the cosine of the angle between them.

For mean-centered paired data normalized by their standard deviations (paired z-scores), their cosine similarity and Pearson's correlation coefficient are the same thing. Because cosines are not additive, we must conclude that we cannot arithmetically add correlation coefficients, or similarity measures, and compute arithmetic averages from these.

Tools for Computing Distances and Similarities

As mentioned early, we have developed two tools for solving distance and similarity problems. Both tools work by accepting any two binary data sets of same size. Consider the following data sets.

A = {1,0,1,1,0}

B = {1,1,0,1,1}

We may express their distance or similarity using one or several different frameworks, depending on the meaning of the 0's and 1's or the problem at hand. Our tools do this by generating 2x2 contingency tables consisting of the following (*i*,*j*) counts:

- (1,1) counts, meaning 'positive matches'.
- (1,0) counts, meaning '*i* absence mismatches'.
- (0,1) counts, meaning '*j* absence mismatches'.
- (0,0) counts, meaning 'negative matches'.

Those familiar with 2x2 contingency tables know that for binary data the diagonal from (1,1) to (0,0) represents the total number of matches or "correct answers" between i and j. By contrast, the diagonal from (0,1) to (1,0) represents the total number of mismatches or "incorrect answers" between i and j. This is the so-called Hamming Distance.

The Hamming Distance is a measure that only allows substitutions and applies to sets of same size. For binary sets of same size, the Hamming, Manhattan (City-Block), and Squared Euclidean distances are all the same thing. For same-size sets, Hamming Distance is an upper bound on the Levenshtein Distance.

Distance-Similarity Transformations

As noted by Lin (1998), the definition of similarity depends on the model or knowledge domain under inspection and is tied to a specific problem. Arbitrarily transforming distances into similarities and vice versa compounds many of the problems described by Lin.

Sometimes such transformations are done using the following “tricks of the trade” (du Toit, Steyn, & Stumpf, 1986; Lin, 1998; Tolechini, et. al., 2012).

$$D = 1 - S \tag{1}$$

$$D = \frac{1-S}{S} \tag{2}$$

$$D = \sqrt{(1 - S)} \tag{3}$$

$$D = \sqrt{2(1 - S)} \tag{4}$$

$$D = \arccos(S) \tag{5}$$

$$D = -\ln(S) \tag{6}$$

where (1) is typically used to transform Jaccard, Dice, Sokal-Michener, Rogers-Tanimoto, and Rusell-Rao similarities into distances (Wolfram, 2015).

In general, the similarity-distance transformations to be used depend on the problem to be solved. As stressed by du Toit, et al. (1986), while a distance can be transformed into a similarity, the reverse process is not so obvious because of the triangular inequality which must be satisfied by a distance metric.

So given a similarity matrix **S** populated with S_{jj} values: How could we compute the corresponding distance matrix **D**? Well, assuming that the similarity matrix **S** is positive semi-definite

$$D_{i,j} = \sqrt{S_{i,i} - 2S_{i,j} + S_{j,j}} \tag{7}$$

is the standard transformation from **S** to **D**. For the particular case of $S_{ii} = S_{jj} = 1$ and

$$D_{i,j} = \sqrt{2(1 - S_{i,j})} \tag{8}$$

which can be used to compute $D_{i,j}$ and matrix **D**.

Conclusion

Distances and similarities have symmetry. However, distances are always positive while similarities can be negative.

Distances are metrics, while similarities are not. We can arithmetically average, add, or subtract distances to compute new distances, but we cannot do the same with similarities. If similarity is resemblance and distance is lack of resemblance, then “similarity distance” (Cilibrasi & Vitányi, 2007) is an oxymoron.

References

Cilibrasi, R. L. and Vitányi, P. M. B. (2007). The Google Similarity Distance. IEEE Transactions On Knowledge And Data Engineering, Vol. 19, No 3, pp 370–383. Retrieved from http://arxiv.org/pdf/cs/0412098v3.pdf

du Toit, S. H. C., Steyn, A. G. W., and Stumpf, R. H. (1986). Graphical Exploratory Data Analysis, Chapter 5: Cluster Analysis, p. 79, Springer-Verlag.

Garcia, E. (2015a). Binary Distance Calculator. Retrieved from http://www.minerazzi.com/tools/distance/binary-distance-calculator.php

Garcia, E. (2015b). Binary Similarity Calculator. Retrieved from http://www.minerazzi.com/tools/similarity/binary-similarity-calculator.php

Lin, D. (1998). An Information-Theoretic Definition of Similarity. ICML '98 Proceedings of the Fifteenth International Conference on Machine Learning. pp. 296-304. Retrieved from http://www.cs.ualberta.ca/~lindek/papers/sim.pdf

Todeschini, R., Consonni, V., Xiang, H., Holliday, J., Buscema, M., and Willet, P. (2012). Similarity Coefficients for Binary Chemoinformatics Data: Overview and Extended Comparison Using Simulated and Real Data Sets. J. Chem. Inf. Model. 52 (11). Retrieved from http://anothersample.net/similarity-coefficients-for-binary-chemoinformatics-data-overview-and-extended-comparison-using-simulated-and-real-data-sets

Wolfram (2015). Language Guide. Distance and Similarity Measures. Retrieved from http://reference.wolfram.com/language/guide/DistanceAndSimilarityMeasures.html