Cosine Similarity Tutorial

Abstract – This is a tutorial on the cosine similarity measure. Its meaning in the context of uncorrelated and orthogonal variables is examined.

Keywords: cosine similarity, tutorial, dot products, vectors, orthogonal, uncorrelated
Published: 04-10-2015; Updated: 09-15-2018
 admin@minerazzi.com

Introduction

In previous tutorials we discussed the difference between distance and similarity measures and the risks of arbitrarily transforming or averaging these (Garcia, 2015a; 2015b; 2015c; 2015d).

We mentioned that a *Pearson's Correlation Coefficient* (*r*) computed from mean-centered variables, or from *z*-scores, is a cosine similarity. In this case, *r* can be computed from the regression curve slope.

We also mentioned that like correlations, slopes, and cosines, cosine similarity measures are not additive so we cannot compute arithmetic averages from any of these measures. The same can be said about standard deviations, rates, and dissimilar ratios. It should be noted that a mean value is not an estimate of central tendency when a distribution is either skewed or Cauchy. Furthermore, the Law of Large Numbers does not apply to a Cauchy Distribution.

Conventions used in this tutorial

This tutorial was written as a companion for a Cosine Similarity Calculator (Garcia, 2015a), and might serve as a basic tutorial for students and those starting in data mining and information retrieval. For the sake of clarity, we adopt the following conventions:

- $\boldsymbol{a}$ and $\boldsymbol{b}$ are the $\vec{a}$ and $\vec{b}$ vectors.
- dp_{ab} is the $\vec{a} \cdot \vec{b}$ dot product between $\boldsymbol{a}$ and $\boldsymbol{b}$.
- dp_{aa} is the dot product of $\boldsymbol{a}$ with itself.
- dp_{bb} is the dot product of $\boldsymbol{b}$ with itself.
- l_a and l_b are the $\|\vec{a}\|$ and $\left\|\vec{b}\right\|$ vector lengths.
- $\hat{\boldsymbol{a}}$ and $\hat{\boldsymbol{b}}$ are unit vectors; i.e., $l_a = l_b = 1$.

Instead of just saying that the cosine similarity between two vectors is given by the expression

$$\cos\theta = \frac{\vec{a}\cdot\vec{b}}{\|\vec{a}\|\|\vec{b}\|} \quad (1)$$

we want to explain **what is actually scored** with (1). In the Appendix section, we have include a light discussion of (1) using a legacy material that inspired this tutorial.

Similarity is an interesting measure as there are many ways of computing it. Indeed, we built a tool that computes over 70 different similarity measures (Garcia, 2016). Since there are so many ways of expressing similarity, what kind of resemblance a cosine similarity actually scores? This is the question that this tutorial pretends to address.

Generally speaking, similarity is a measure of resemblance; i.e., how similar or alike things being compared are. One way of computing similarity is through the use of vectors.

Representing data sets as vectors

A *matrix* is just a table filled with values. Suppose that said table consists of *r* number of rows and *c* number of columns. We may refer to these as *vectors*. Thus for a square matrix, one with same number of rows and columns, row vectors are data sets of size $n = c$ and column vectors are data sets of size $n = r$.

A vector is a quantity or phenomenon with two independent properties: direction and magnitude. For $n = 2$ and $n = 3$, we can visualize a vector in its *n*-dimensional space as a line segment ending in an arrow. The orientation of the line segment is its direction, and the length is its magnitude. For $n > 3$ we cannot visualize vectors, but we can still compute them.

Computing Dot Products

We may multiply a vector by *itself or another* vector and compute a quantity called the dot product (*dp*). This is done by multiplying vector elements and taking summations. To illustrate, suppose that ***a*** and ***b*** are vectors such that

- ***a*** = [1, 2, 3]
- ***b*** = [4, -5, 6]

In this case, the dot product of

- $\boldsymbol{a}$ with $\boldsymbol{b}$ is dp_{ab} = 1*4 + 2*-5 + 3*6 = 12
- $\boldsymbol{a}$ with itself is dp_{aa} = 1*1 + 2*2 + 3*3 = 14
- $\boldsymbol{b}$ with itself is dp_{bb} = 4*4 + -5*-5 + 6*6 = 77

We may now ask: What kind of information is stored in dp_{ab}, dp_{aa}, and dp_{bb}?

What information is stored in dp_{ab}?

Good question: dp_{ab} holds information about the *direction* of the vectors. To be precise, if

- $dp_{ab} > 0$, $\boldsymbol{a}$ and $\boldsymbol{b}$ form an angle less than 90^o.
- $dp_{ab} = 0$, $\boldsymbol{a}$ and $\boldsymbol{b}$ form an angle that is exactly 90^o.
- $dp_{ab} < 0$, $\boldsymbol{a}$ and $\boldsymbol{b}$ form an angle greater than 90^o.

Table 1 shows different types of angles.

Table 1. Possible types of angles.

Type	acute	right	obtuse	straight	reflex	perigon
Angle	$\theta < 90^o$	$\theta = 90^o$	$180 > \theta > 90^o$	$\theta = 180^o$	$360^o > \theta > 180^o$	$\theta = 360^o$

An angle of 0^o means that $cos\ \theta = 1$ so the vectors point to identical directions. An angle of 90^o means that $cos\ \theta = 0$ so the vectors point to perpendicular directions or are *orthogonal*.

What information is stored in dp_{aa} and dp_{bb}?

Taking square roots, it is clear that dp_{aa} and dp_{bb} hold *length information*. So

- $l_a = (dp_{aa})^{½} = (14)^{½} = 3.74$; i.e., the length of $\boldsymbol{a}$.
- $l_b = (dp_{bb})^{½} = (77)^{½} = 8.77$; i.e., the length of $\boldsymbol{b}$.
- $l_a * l_b = (dp_{aa})^{½} * (dp_{bb})^{½} = 32.83$; i.e., the length product (lp_{ab}) of $\boldsymbol{a}$ and $\boldsymbol{b}$.

In other words, (1) is a dot product/length product ratio

$$\cos\theta = \frac{dp_{ab}}{lp_{ab}} = \frac{dp_{ab}}{\sqrt{dp_{aa}}\sqrt{dp_{bb}}} = \frac{12}{32.84} = 0.37 \quad (2)$$

Therefore, when we compute a cosine similarity we are measuring the direction-length resemblance between data sets represented as vectors.

Comparing Vectors of Different Lengths

To compare vectors of different lengths, these can be recomputed as *unit vectors*. A unit vector is computed by dividing its elements by its length. In other words, we write the previous vectors as

$\hat{\boldsymbol{a}}$ = [1/3.74, 2/3.74, 3/3.74]
$\hat{\boldsymbol{b}}$ = [4/8.77, -5/8.77, 6/8.77]

where the hat (^) denotes a unit vector. Since the new lengths are equal to 1, the cosine similarity between $\hat{\boldsymbol{a}}$ and $\hat{\boldsymbol{b}}$ is their dot product; hence

$$\cos\theta = dp_{\hat{a}\hat{b}} = 0.37 \quad (3)$$

Expressions (1), (2), and (3) return the same result, confirming that a cosine is a judgment of the orientation of the vectors, independent of their lengths (Wikipedia, 2015a).

What does $\cos\theta = 0$ mean?

As mentioned before, an angle of 90^o means that $\cos\theta = 0$ so the vectors are perpendicular or orthogonal. This does not necessarily mean that the variables are uncorrelated.

According to Rodgers, Nicewander, & Toothaker (1984) when referring to variables, orthogonal denotes that *raw* variables are perpendicular while uncorrelated means that *centered* variables are perpendicular. Centered variables are those with their mean removed, so they have zero mean.

It turns out that unlike vector length normalization, subtracting the mean from raw variables can change the angle between the vectors. The following can happen to the raw variables:

- If they are perpendicular, can become not perpendicular so they are orthogonal, but not uncorrelated.
- If they are not perpendicular, can become perpendicular so they are uncorrelated, but not orthogonal.
- If they are perpendicular and remain perpendicular, they are orthogonal and uncorrelated.
- If they are not perpendicular and remain not perpendicular, they are neither orthogonal nor uncorrelated—although their angle can change.

The following figure, adapted from Rodgers *et. al* (1984), illustrates these relationships. Notice that not all uncorrelated variables are orthogonal and *vice versa* (Wikipedia, 2015b). The figure also shows that while all uncorrelated (or orthogonal) variables are independent, the reverse is not true.

Therefore, a textbook referring to orthogonal variables as uncorrelated is probably referring to paired random variables with zero mean; i.e., to the overlapping region shown in the figure.

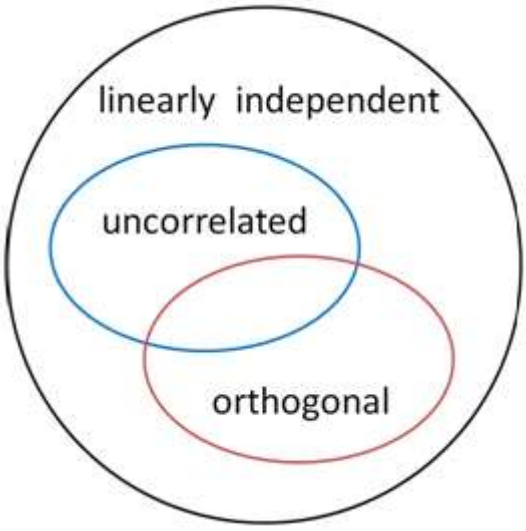

Figure 1. Venn Diagram for linearly independent, uncorrelated, and orthogonal variables.

These results are not surprising. It can be shown for centered variables that

$$r = \cos\theta \tag{4}$$

Because r is the covariance of paired variables normalized by their standard deviation,

$$r = \frac{cov_{ab}}{s_a s_b} \tag{5}$$

so for *centered* variables if $\cos\theta = 0$, then $r = 0$, $cov_{ab} = 0$, and the variables must be uncorrelated.

The relationship $r = \cos\theta$ is also true for standardized variables, also known as *z*-scores. These are centered variables normalized by their standard deviation. Transforming variables into *z*-scores is, though not always, useful and recommended.

Once in a *z*-score format, we can run other types of tests on the variables like a quantile-quantile analysis (Garcia, 2015e), or even compute r, and therefore $\cos\theta$, from the slope of the regression curve of *z*-scores. For those interested, we have developed another tool called the Standardizer (Garcia, 2015f). This tool was initially thought of as an *x-to-z*-score standardizer, hence its name. It is now a versatile statistical tool for univariate analysis.

Conclusion

Unlike other similarity measures, a cosine similarity is a measure of the direction-length resemblance between vectors.

An angle of 0^o means that $\cos\theta = 1$ and that the vectors are oriented in identical directions; i.e., that the corresponding data sets are completely similar to one another. An angle of 90^o means that $\cos\theta = 0$ and that the corresponding variables are perpendicular, but not necessarily that are uncorrelated unless these are also mean-centered.

Computing $\cos\theta$ from raw and center variables are two different things. This fact can be used to examine relationships between paired variables. Regardless of the method used for calculating cosine similarities, and for the sake of transparency, it is always a good idea to state whether raw or centered variables were used and why.

Appendix

Provided below is the legacy material that inspired this tutorial, extracted from the article *Cosine Similarity and Term Weight Tutorial* which I published circa 2006 at http://www.miislita.com, now a search engine site. After many thoughts I am adding this material with its content heavily edited.

Let us first define a reference point $C(x_0, y_0)$ at the origin of the x-y plane so by default $x_0 = 0, y_0 = 0$. Similarly, let's refer to any two points, A and B, on this plane as $A(x_1, y_1)$ and $B(x_2, y_2)$. See Figure A1.

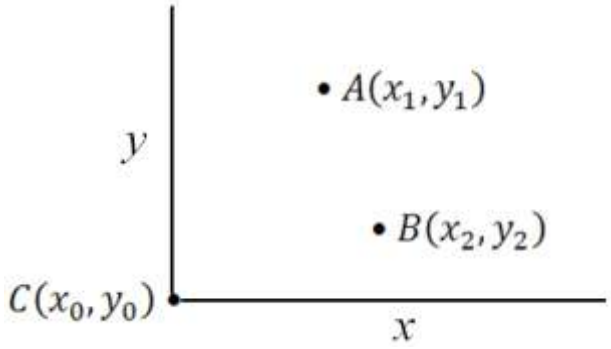

Figure A1. Coordinates of points, *A*, *B*, and *C* in a two-dimensional plane.

If we multiply the coordinates of A and B and add the products together we get the "mythical" dot product.

$$A \bullet B = x_1 * x_2 + y_1 * y_2 \qquad \text{(a1)}$$

The little bullet in "$A \bullet B$" indicates -you guess right- the dot product between A and B. If these points are defined in three dimensions, their coordinates are (x_1, y_1, z_1) and (x_2, y_2, z_2), and can be referred to as $A(x_1, y_1, z_1)$ and $B(x_2, y_2, z_2)$. The $A \bullet B$ dot product is given now by

$$A \bullet B = x_1 * x_2 + y_1 * y_2 + z_1 * z_2 \qquad \text{(a2)}$$

For additional dimensions, we just keep adding product terms to (a2). It cannot get any easier than this.

Now, to define a straight line we need at least two points. So if we draw a straight line from *C* to either *A* or *B*, we can define the distance between the points. This is called the Euclidean Distance *d* which is computed in three easy steps:

1. Take the difference between point coordinates.
2. Square all differences and add them together.
3. Square root the result.

Since we have defined $x_0 = 0$ and $y_0 = 0$, then to find out how far *A* and *B* are from *C*, we define the Euclidean Distances

$$d_{AC} = ((x_1 - x_0)^2 + (y_1 - y_0)^2)^{1/2} = (x_1^2 + y_1^2)^{1/2} \quad \text{(a3)}$$

$$d_{BC} = ((x_2 - x_0)^2 + (y_2 - y_0)^2)^{1/2} = (x_2^2 + y_2^2)^{1/2} \quad \text{(a4)}$$

Figure A2 depicts these distances as straight lines connecting the points.

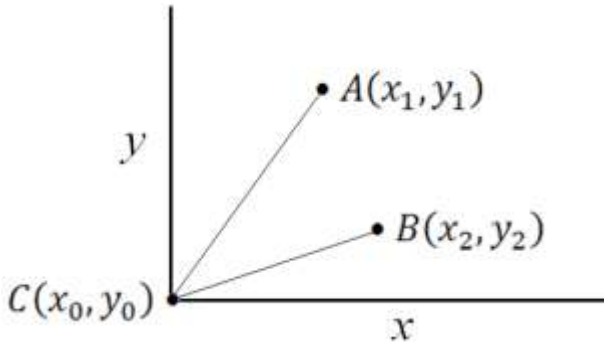

Figure A2. Straight lines representing Euclidean Distances between points *A* and *B*, with *C*.

The straight lines in Figure A2 can be replaced by arrows representing vectors. As previously discussed, a vector is a quantity with direction and magnitude. The head and angle of the arrow indicates the direction of the vector, while its magnitude is defined by the usual Euclidean Distance.

Since in this example $x_0 = 0$ and $y_0 = 0$, we can simplify and express the magnitudes of the *A* and *B* vectors as $d_{AC} = |\boldsymbol{A}|$ and $d_{BC} = |\boldsymbol{B}|$. Again, the pipe symbol indicates that we are dealing with absolute magnitudes.

Thus, the lengths of the arrows represent vector magnitudes. The angle described by the vectors represents their orientation in a two-dimensional space. See Figure A3.

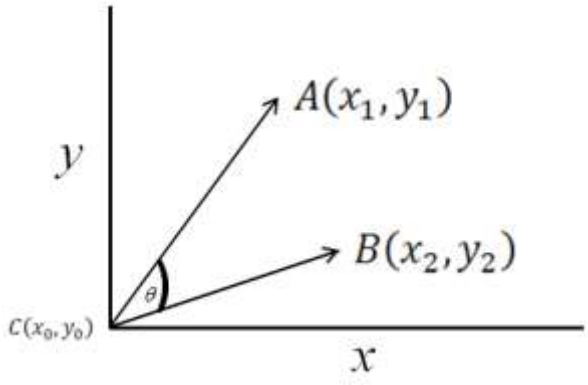

Figure A3. *A* and *B* Vectors.

To normalize the $A \bullet B$ dot product we divide it by the Euclidean Distance between *A* and *B*; i.e., $A \bullet B/(|A||B|)$. This ratio defines the cosine of the angle between vectors, commonly known as the vectors *cosine similarity (cosim)* and denoted below as *Sim(A, B)*.

$$Sim(A,B) = \cos\theta = \frac{A \bullet B}{|A||B|} = \frac{x_1 * x_2 + y_1 * y_2}{\sqrt{(x_1^2 + y_1^2)(x_2^2 + y_2^2)}} \qquad \text{(a5)}$$

Some Applications to Information Retrieval

As the angle between the vectors shortens, $\cos\theta$ approaches 1, meaning that the two vectors are getting closer so the similarity of whatever is represented by the vectors increases. This is a convenient way of ranking searchable documents. For instance, let say that $A(x_1, y_1)$ represents a query *q* and points $B(x_2, y_2)$, $D(x_3, y_3)$, $E(x_4, y_4)$, $F(x_5, y_5)$,… represent documents.

We should be able to compute the cosine of the angle between *q* and each document *D* and sort the documents in decreasing order of cosine similarites. This treatment can be extended to entire collection of documents.

To do that we need to construct a term space. The term space is defined by an index or list of unique terms. These terms are extracted from the collection of documents to be queried. The coordinates of the points representing documents and queries are now defined according to a weighting scheme.

$$Sim(q, D_i) = \cos\theta = \frac{\sum_i w_{q,j} w_{i,j}}{\sqrt{\sum_j w_j^2}\sqrt{\sum_j w_{i,j}^2}} \quad \text{(a6)}$$

where the sigma symbol (Σ) means "the sum of" and *w* are weights assigned to query and document terms. The following is a list of some of the weighting schemes utilized by early web search engines for retrieving and ranking documents:

- *w = tf*
- *w = tf/tfmax*
- *w = IDF = log(N/n)*
- *w = tf*IDF = tf*log(N/n)*
- *w = tf*IDF = tf*log((N – n)/n)*

where

- *tf* stands for term frequency or how many times a term is mentioned in a document.
- *tfmax* is the frequency of the term that is mentioned the most.
- *N* is the size of the collection of documents queried.
- *n* is the number of documents mentioning a query term.
- *IDF* stands for Inverse Document Frequency.

Modern search engines use algorithms less dependent on *tf* and that cannot be gamed by artificially repeating terms. Some of these are the family of algorithms known as Best Match (BM), Latent Semantic Indexing (LSI), and similar algorithms.

References

Garcia, E. (2016). Binary Similarity Calculator. Retrieved from http://www.minerazzi.com/tools/similarity/binary-similarity-calculator.php

Garcia, E. (2015a). Cosine Similarity Calculator. Retrieved from http://www.minerazzi.com/tools/cosine-similarity/cosine-similarity-calculator.php

Garcia, E. (2015b). A Tutorial on Distance and Similarity. Retrieved from http://www.minerazzi.com/tutorials/distance-similarity-tutorial.pdf

Garcia, E. (2015c). The Self-Weighting Model Tutorial: Part 1 Retrieved from http://www.minerazzi.com/tutorials/self-weighting-model-tutorial-part-1.pdf

Garcia, E. (2015d). The Self-Weighting Model Tutorial: Part 2 Retrieved from http://www.minerazzi.com/tutorials/self-weighting-model-tutorial-part-2.pdf

Garcia, E. (2015e). A Quantile-Quantile Plot tutorial. Retrieved from http://www.minerazzi.com/tutorials/quantile-quantile-tutorial.pdf

Garcia, E. (2015f). The Standardizer. Retrieved from http://www.minerazzi.com/tools/standardizer/z-scores.php

Rodgers, J. L., Nicewander, W. A., Toothaker, L. (1984). Linearly Independent, Orthogonal, and Uncorrelated Variables. *The American Statistician*, Vol. 38, No. 2. Pp 133-134. Retrieved from https://pdfs.semanticscholar.org/177d/c3b719fdb7b2c01d6c341e0bd58a09a47a83.pdf

Wikipedia (2015a). Cosine Similarity. Retrieved from http://en.wikipedia.org/wiki/Cosine_similarity

Wikipedia (2015b). Uncorrelated. Retrieved from http://en.wikipedia.org/wiki/Uncorrelated