Cosine similarity

From Wikipedia, the free encyclopedia

Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them. The cosine of 0° is 1, and it is less than 1 for any other angle. It is thus a judgment of orientation and not magnitude: two vectors with the same orientation have a cosine similarity of 1, two vectors at 90° have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude. Cosine similarity is particularly used in positive space, where the outcome is neatly bounded in [0,1]. The name derives from the term "direction cosine": in this case, note that unit vectors are maximally "similar" if they're parallel and maximally "dissimilar" if they're orthogonal (perpendicular). This is analogous to the cosine, which is unity (maximum value) when the segments subtend a zero angle and zero (uncorrelated) when the segments are perpendicular.

Note that these bounds apply for any number of dimensions, and cosine similarity is most commonly used in high-dimensional positive spaces. For example, in information retrieval and text mining, each term is notionally assigned a different dimension and a document is characterised by a vector where the value of each dimension corresponds to the number of times that term appears in the document. Cosine similarity then gives a useful measure of how similar two documents are likely to be in terms of their subject matter.[1]

The technique is also used to measure cohesion within clusters in the field of data mining.[2]

Cosine distance is a term often used for the complement in positive space, that is: $D_C(A,B) = 1 - S_C(A,B)$, where D_C is the cosine distance and S_C is the cosine similarity. It is important to note, however, that this is not a proper distance metric as it does not have the triangle inequality property—or, more formally, the Schwarz inequality—and it violates the coincidence axiom; to repair the triangle inequality property while maintaining the same ordering, it is necessary to convert to angular distance (see below.)

One of the reasons for the popularity of cosine similarity is that it is very efficient to evaluate, especially for sparse vectors, as only the non-zero dimensions need to be considered.

Contents

Definition

The cosine of two non-zero vectors can be derived by using the Euclidean dot product formula:

$$\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\|_2 \|\mathbf{b}\|_2 \cos\theta$$

Given two vectors of attributes, *A* and *B*, the cosine similarity, *cos(θ)*, is represented using a dot product and magnitude as

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|_2\|\mathbf{B}\|_2} = \frac{\sum\limits_{i=1}^{n} A_i B_i}{\sqrt{\sum\limits_{i=1}^{n} A_i^2}\sqrt{\sum\limits_{i=1}^{n} B_i^2}}$$

, where A_i and B_i are components of vector A and B respectively.

The resulting similarity ranges from −1 meaning exactly opposite, to 1 meaning exactly the same, with 0 indicating orthogonality (decorrelation), and in-between values indicating intermediate similarity or dissimilarity.

For text matching, the attribute vectors *A* and *B* are usually the term frequency vectors of the documents. The cosine similarity can be seen as a method of normalizing document length during comparison.

In the case of information retrieval, the cosine similarity of two documents will range from 0 to 1, since the term frequencies (tf-idf weights) cannot be negative. The angle between two term frequency vectors cannot be greater than 90°.

If the attribute vectors are normalized by subtracting the vector means (e.g., $A - \bar{A}$), the measure is called centered cosine similarity and is equivalent to the Pearson correlation coefficient.

Angular distance and similarity

The term "cosine similarity" is sometimes used to refer to a different definition of similarity provided below. However the most common use of "cosine similarity" is as defined above and the similarity and distance metrics defined below are referred to as "angular similarity" and "angular distance" respectively. The normalized angle between the vectors is a formal distance metric and can be calculated from the similarity score defined above. This angular distance metric can then be used to compute a similarity function bounded between 0 and 1, inclusive.

When the vector elements may be positive or negative:

$$\text{distance} = \frac{\cos^{-1}(\text{similarity})}{\pi}$$

$$\text{similarity} = 1 - \text{distance}$$

Or, if the vector elements are always positive:

$$\text{distance} = \frac{2\cdot\cos^{-1}(\text{similarity})}{\pi}$$

$$\text{similarity} = 1 - \text{distance}$$

Although the term "cosine similarity" has been used for this angular distance, the term is used as the cosine of the angle only as a convenient mechanism for calculating the angle itself and is no part of the meaning. The advantage of the angular similarity coefficient is that, when used as a difference coefficient (by subtracting it from 1) the resulting function is a proper distance metric, which is not the case for the first meaning. However, for most uses this is not an important property. For any use where only the relative ordering of similarity or distance within a set of vectors is important, then which function is used is immaterial as the resulting order will be unaffected by the choice.

Confusion with "Tanimoto" coefficient

The cosine similarity may be easily confused with the Tanimoto metric - a specialised form of a similarity coefficient with a similar algebraic form:

$$T(A,B) = \frac{A \cdot B}{\|A\|^2 + \|B\|^2 - A \cdot B}$$

In fact, this algebraic form was first defined by Tanimoto as a mechanism for calculating the Jaccard coefficient in the case where the sets being compared are represented as bit vectors. While the formula extends to vectors in general, it has quite different properties from cosine similarity and bears little relation other than its superficial appearance.

Ochiai coefficient

This coefficient is also known in biology as Ochiai coefficient, or Ochiai-Barkman coefficient, or Otsuka-Ochiai coefficient:[3][4]

$$K = \frac{n(A \cap B)}{\sqrt{n(A) \times n(B)}}$$

Here, A and B are sets, and $n(A)$ is the number of elements in A. If sets are represented as bit vectors, the Ochiai coefficient can be seen to be the same as the cosine similarity.

Properties

Cosine similarity is related to Euclidean distance as follows. Denote Euclidean distance by the usual $\|A - B\|$, and observe that

$$\|A - B\|^2 = (A - B)^\top (A - B) = \|A\|^2 + \|B\|^2 - 2A^\top B$$

by expansion. When A and B are normalized to unit length, $\|A\|^2 = \|B\|^2 = 1$ so the previous is equal to

$$2(1 - \cos(A, B))$$

Null distribution: For data which can be negative as well as positive, the null distribution for cosine similarity is the distribution of the dot product of two independent random unit vectors. This distribution has a mean of zero and a variance of $1/n$ (where n is the number of dimensions), and although the distribution is bounded between -1 and +1, as n grows large the distribution is increasingly well-approximated by the normal distribution.[5][6] For other types of data, such as bitstreams (taking values of 0 or 1 only), the null distribution will take a different form, and may have a nonzero mean.[7]

Soft cosine measure

The soft cosine measure is a measure of “soft” similarity between two vectors, i.e., the measure that considers similarity of pairs of features.[8] The traditional cosine similarity considers the vector space model (VSM) features as independent or completely different, while the soft cosine measure proposes considering the similarity of features in VSM, which allows generalization of the concepts of cosine measure and also the idea of similarity (soft similarity).

For example, in the field of natural language processing (NLP) the similarity among features is quite intuitive. Features such as words, n-grams or syntactic n-grams[9] can be quite similar, though formally they are considered as different features in the VSM. For example, words “play” and “game” are different words and thus are mapped to different dimensions in VSM; yet it is obvious that they are related semantically. In case of n-grams or syntactic n-grams, Levenshtein distance can be applied (in fact, Levenshtein distance can be applied to words as well).

For calculation of the soft cosine measure, the matrix **s** of similarity between features is introduced. It can be calculated using Levenshtein distance or other similarity measures, e.g., various WordNet similarity measures. Then we just multiply by this matrix.

Given two N-dimension vectors a and b, the soft cosine similarity is calculated as follows:

$$\mathbf{soft\_cosine}_1(a,b)=\frac{\sum_{i,j}^{N} s_{ij}a_ib_j}{\sqrt{\sum_{i,j}^{N} s_{ij}a_ia_j}\sqrt{\sum_{i,j}^{N} s_{ij}b_ib_j}},$$

where s_{ij} = similarity(feature$_i$, feature$_j$).

If there is no similarity between features ($s_{ii} = 1, s_{ij} = 0$ for $i \neq j$), the given equation is equivalent to the conventional cosine similarity formula.

The complexity of this measure is quadratic, which makes it perfectly applicable to real world tasks. The complexity can be transformed to subquadratic.

See also

- Sørensen's quotient of similarity
- Hamming distance
- Correlation
- Dice's coefficient
- Jaccard index
- SimRank
- Information retrieval

References

1. Singhal, Amit (2001). "Modern Information Retrieval: A Brief Overview". Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 24 (4): 35–43.
2. P.-N. Tan, M. Steinbach & V. Kumar, "Introduction to Data Mining", , Addison-Wesley (2005), ISBN 0-321-32136-7, chapter 8; page 500.
3. *Ochiai A.* Zoogeographical studies on the soleoid fishes found Japan and its neighboring regions. II // Bull. Jap. Soc. sci. Fish. 1957. V. 22. № 9. P. 526-530.
4. *Barkman J.J.* Phytosociology and ecology of cryptogamic epiphytes, including a taxonomic survey and description of their vegetation units in Europe. – Assen. Van Gorcum. 1958. 628 p.
5. Spruill, Marcus C (2007). "Asymptotic distribution of coordinates on high dimensional spheres". *Electronic communications in probability*. **12**: 234–247. doi:10.1214/ECP.v12-1294 (https://doi.org/10.1214%2FECP.v12-1294).
6. CrossValidated: Distribution of dot products between two random unit vectors in RD (http://stats.stackexchange.com/questions/85916/distribution-of-dot-products-between-two-random-unit-vectors-in-mathbbrd)
7. Graham L. Giller (2012). "The Statistical Properties of Random Bitstreams and the Sampling Distribution of Cosine Similarity". *Giller Investments Research Notes* (20121024/1). doi:10.2139/ssrn.2167044 (https://doi.

org/10.2139%2Fssrn.2167044).

8. Sidorov, Grigori; Gelbukh, Alexander; Gómez-Adorno, Helena; Pinto, David. "Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model" (http://cys.cic.ipn.mx/ojs/index.php/CyS/article/view/2043). *Computación y Sistemas*. **18** (3): 491–504. doi:10.13053/CyS-18-3-2043 (https://doi.org/10.13053%2FCyS-18-3-2043). Retrieved 7 October 2014.
9. Sidorov, Grigori; Velasquez, Francisco; Stamatatos, Efstathios; Gelbukh, Alexander; Chanona-Hernández, Liliana. *Syntactic Dependency-based N-grams as Classification Features* (http://link.springer.com/chapter/10.1007%2F978-3-642-37798-3_1). LNAI 7630. pp. 1–11. ISBN 978-3-642-37798-3. Retrieved 7 October 2014.

External links

- Weighted cosine measure (http://mathforum.org/kb/message.jspa?messageID=5658016&tstart=0)
- A tutorial on cosine similarity using Python (http://blog.christianperone.com/?p=2497)
- Web API to Compute Cosine, Jaccard and Dice for Text in Any Language (http://www.rxnlp.com/api-reference/text-similarity-api-reference/)

Retrieved from "https://en.wikipedia.org/w/index.php?title=Cosine_similarity&oldid=793441851"