Cosine similarity

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 angle in the interval $(0, \pi]$ radians. It is thus a judgment of orientation and not magnitude: two vectors with the same orientation have a cosine similarity of 1, two vectors oriented at 90° relative to each other have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude. The cosine similarity is particularly used in positive space, where the outcome is neatly bounded in $[\mathbf{0}, \mathbf{1}]$. The name derives from the term "direction cosine": in this case, 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.

These bounds apply for any number of dimensions, and the 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 in each dimension corresponds to the number of times the 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]

The term cosine distance is 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 advantage of cosine similarity is its low-complexity, especially for sparse vectors: only the non-zero dimensions need to be considered.

Other names of cosine similarity are *Orchini* similarity and the *Tucker* coefficient of congruence; *Ochiai* similarity (see below) is cosine similarity applied to binary data.

Contents

Definition

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

$$\mathbf{A} \cdot \mathbf{B} = \|\mathbf{A}\| \, \|\mathbf{B}\| \cos\theta$$

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

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

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 or decorrelation, while in-between values indicate intermediate similarity or dissimilarity.

For text matching, the attribute vectors *A* and *B* are usually the term frequency vectors of the documents. 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 (using 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 the centered cosine similarity and is equivalent to the Pearson correlation coefficient. For an example of centering,

$$\text{if } A = [A_1, A_2]^T, \text{ then } \bar{A} = \left[\frac{(A_1 + A_2)}{2}, \frac{(A_1 + A_2)}{2}\right]^T, \text{ so } A - \bar{A} = \left[\frac{(A_1 - A_2)}{2}, \frac{(-A_1 + A_2)}{2}\right]^T.$$

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:

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

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

Or, if the vector elements are always positive:

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

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

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 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.

Otsuka-Ochiai coefficient

In biology, there is a similar concept known as the Otsuka-Ochiai coefficient named after Yanosuke Otsuka (also spelled as Ōtsuka, Ootsuka or Otuka,[3] Japanese: 大塚 弥之助)[4] and Akira Ochiai (Japanese: 落合 明),[5] also known as the Ochiai-Barkman[6] or Ochiai coefficient,[7] which can be represented as:

$$K = \frac{|A \cap B|}{\sqrt{|A| \times |B|}}$$

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

In a recent book,[8] the coefficient is misattributed to another Japanese researcher with the family name Otsuka. The confusion arises because in 1957 Akira Ochiai attributes the coefficient only to Otsuka (no first name mentioned)[5] by citing an article by Ikuso Hamai (Japanese: 浜井 生三),[9] who in turn cites the original 1936 article by Yanosuke Otsuka.[4]

Properties

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

$$\|A - B\|^2 = (A - B)^\mathsf{T}(A - B) = \|A\|^2 + \|B\|^2 - 2A^\mathsf{T}B$$

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

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

The Euclidean distance is called the *chord distance* (because it is the length of the chord on the unit circle) and it is the Euclidean distance between the vectors which were normalized to unit sum of squared values within them.

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.[10][11] Other types of data such as bitstreams, which only take the values 0 or 1, the null distribution takes a different form and may have a nonzero mean.[12]

Soft cosine measure

A soft cosine or ("soft" similarity) between two vectors considers similarities between pairs of features.[13] 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 help generalize the concept of cosine (and soft cosine) as well as the idea of (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[14] 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 mapped to different points in VSM; yet they are semantically related. 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 calculating soft cosine, the matrix **s** is used to indicate similarity between features. It can be calculated through Levenshtein distance, WordNet similarity, or other similarity measures. Then we just multiply by this matrix.

Given two N-dimension vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, the soft cosine similarity is calculated as follows:

$$\text{soft\_cosine}_1(a, b) = \frac{\sum_{i,j}^{N} s_{ij} a_i b_j}{\sqrt{\sum_{i,j}^{N} s_{ij} a_i a_j} \sqrt{\sum_{i,j}^{N} s_{ij} b_i b_j}},$$

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

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 time complexity of this measure is quadratic, which makes it applicable to real-world tasks. Note that the complexity can be reduced to subquadratic.[15]

See also

- Sørensen–Dice coefficient
- Hamming distance
- Correlation
- Jaccard index
- SimRank
- Information retrieval

References

1. Singhal, Amit (2001). "Modern Information Retrieval: A Brief Overview (http://singhal.info/ieee2001.pdf)". *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. Omori, Masae (2004). "Geological idea of Yanosuke Otuka, who built the foundation of neotectonics (geoscientist)" (https://www.jstage.jst.go.jp/article/agcjchikyukagaku/58/4/58_KJ00004410060/_pdf/-char/en). *Earth Science*. **58** (4): 256–259. doi:10.15080/agcjchikyukagaku.58.4_256 (https://doi.org/10.15080%2Fagcjchikyukagaku.58.4_256).
4. Otsuka, Yanosuke (1936). "The faunal character of the Japanese Pleistocene marine Mollusca, as evidence of the climate having become colder during the Pleistocene in Japan". *Bulletin of the Biogeographical Society of Japan*. **6** (16): 165–170.
5. Ochiai, Akira (1957). "Zoogeographical studies on the soleoid fishes found in Japan and its neighhouring regions-II" (https://www.jstage.jst.go.jp/article/suisan1932/22/9/22_9_526/_pdf/-char/en). *Bulletin of the Japanese Society of Scientific Fisheries*. **22** (9): 526–530. doi:10.2331/suisan.22.526 (https://doi.org/10.2331%2Fsuisan.22.526).
6. Barkman, Jan J. (1958). *Phytosociology and Ecology of Cryptogamic Epiphytes: Including a Taxonomic Survey and Description of Their Vegetation Units in Europe*. Assen: Van Gorcum.
7. H. Charles Romesburg (1984). *Cluster Analysis for Researchers* (https://books.google.com/books?id=ZuIPv7OKm10C&pg=PA149). Belmont, California: Lifetime Learning Publications. p. 149.
8. Howarth, Richard J. (2017). *Dictionary of Mathematical Geosciences: With Historical Notes* (https://books.google.com/books?id=MNwlDwAAQBAJ&pg=PA421). Cham, Switzerland: Springer. p. 421. doi:10.1007/978-3-319-57315-1 (https://doi.org/10.1007%2F978-3-319-57315-1). ISBN 978-3-319-57314-4.
9. Hamai, Ikuso (1955). "Stratification of community by means of "community coefficient" (continued)" (https://www.jstage.jst.go.jp/article/seitai/5/1/5_KJ00002869450/_pdf/-char/en). *Japanese Journal of Ecology*. **5** (1): 41–45. doi:10.18960/seitai.5.1_41 (https://doi.org/10.18960%2Fseitai.5.1_41).
10. 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).
11. "Distribution of dot products between two random unit vectors in RD" (https://stats.stackexchange.com/q/85916). *CrossValidated*.
12. 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).
13. Sidorov, Grigori; Gelbukh, Alexander; Gómez-Adorno, Helena; Pinto, David (29 September 2014). "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.
14. Sidorov, Grigori; Velasquez, Francisco; Stamatatos, Efstathios; Gelbukh, Alexander; Chanona-Hernández, Liliana (2013). *Advances in Computational Intelligence*. Lecture Notes in Computer Science. **7630**. LNAI 7630. pp. 1–11. doi:10.1007/978-3-642-37798-3_1 (https://doi.org/10.1007%2F978-3-642-37798-3_1). ISBN 978-3-642-37798-3.
15. Novotný, Vít (2018). *Implementation Notes for the Soft Cosine Measure*. The 27th ACM International Conference on Information and Knowledge Management. Torun, Italy: Association for Computing Machinery. pp. 1639–1642. arXiv:1808.09407 (https://arxiv.org/abs/1808.09407). doi:10.1145/3269206.3269317 (https://doi.org/10.1145%2F3269206.3269317). ISBN 978-1-4503-6014-2.

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)