Similarity Metrics

In Milvus, distance metrics are used to measure similarities among vectors. Choosing a good distance metric helps improve the classification and clustering performance significantly.

The following table shows how these widely used distance metrics fit with various input data forms and Milvus indexes.

Floating point embeddings | Binary embeddi

Distance Metrics	Index Types
• Jaccard • Tanimoto • Hamming	• FLAT • IVF_FLAT
• Superstructure • Substructure	FLAT

Euclidean distance (L2)

Essentially, Euclidean distance measures the length of a segment that connects 2 points.

The formula for Euclidean distance is as follows:

$$d(\boldsymbol{a}, \boldsymbol{b}) = d(\boldsymbol{b}, \boldsymbol{a}) = \sqrt{\sum_{i=1}^{n} (b_i - a_i)^2}$$

where **a** = (a1, a2,..., an) and **b** = (b1, b2,..., bn) are two points in n-dimensional Euclidean space

It's the most commonly used distance metric, and is very useful when the data is continuous.

Edit this page
Report a bug
Request doc changes

On this page

Similarity Metrics
FAQ

Inner product (IP)

The IP distance between two embeddings are defined as follows:

$$p(A,B) = A \cdot B = \sum_{i=1}^{n} a_i \times b_i$$

where A and B are embeddings, `||A||` and `||B||` are the norms of A and B.

IP is more useful if you are more interested in measuring the orientation but not the magnitude of the vectors.

> **NOTE**
>
> If you use IP to calculate embeddings similarities, you must normalize your embeddings. After normalization, inner product equals cosine similarity.

Suppose X' is normalized from embedding X:

$$X' = (x'_1, x'_2, \ldots, x'_n), X' \in^n$$

The correlation between the two embeddings is as follows:

$$x'_i = \frac{x_i}{\| X \|} = \frac{x_i}{\sqrt{\sum_{i=1}^{n} (x_i)^2}}$$

Jaccard distance

Jaccard similarity coefficient measures the similarity between two sample sets, and is defined as the cardinality of the intersection of the defined sets divided by the cardinality of the union of them. It can only be applied to finite sample sets.

$$J(A,B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

Jaccard distance measures the dissimilarity between data sets, and is obtained by subtracting the Jaccard similarity coefficient from 1. For binary variables, Jaccard distance is equivalent to Tanimoto coefficient.

$$d_j(A,B) = 1 - J(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

Tanimoto distance

For binary variables, the Tanimoto coefficient is equivalent to Jaccard distance:

$$T(A,B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}$$

In Milvus, the Tanimoto coefficient is only applicable for a binary variable, and for binary variables the Tanimoto coefficient ranges from 0 to +1 (where +1 is the highest similarity).

For binary variables, the formula of Tanimoto distance is:

$$d_t = \frac{A \cdot B}{|A|^2 + |B|^2 - A \cdot B}$$

The value ranges from 0 to +infinity.

Hamming distance

Hamming distance measures binary data strings. The distance between two strings of equal length is the number of bit positions at which the bits are different.

For example, suppose there are two strings 1101 1001 and 1001 1101.

11011001 ⊕ 10011101 = 01000100. Since, this contains two 1s, the Hamming distance, d (11011001, 10011101) = 2.

Superstructure

Superstructure is used to measure the similarity of a chemical structure and its superstructure. The less the value, the more similar the structure is to its superstructure. Only the vectors whose distance equals to 0 can be found now.

Superstructure similarity can be measured by:

$$1 - \frac{N_{A\&B}}{N_A}$$

Where

- B is the superstructure of A

- NA specifies the number of bits in the fingerprint of molecular A.
- NB specifies the number of bits in the fingerprint of molecular B.
- NAB specifies the number of shared bits in the fingerprint of molecular A and B.

Substructure

Substructure is used to measure the similarity of a chemical structure and its substructure. The less the value, the more similar the structure is to its substructure. Only the vectors whose distance equals to 0 can be found now.

Substructure similarity can be measured by:

$$1 - \frac{N_{A\&B}}{N_B}$$

Where

- B is the substructure of A
- NA specifies the number of bits in the fingerprint of molecular A.
- NB specifies the number of bits in the fingerprint of molecular B.
- NAB specifies the number of shared bits in the fingerprint of molecular A and B.

FAQ

▸ Why is the top1 result of a vector search not the search vector itself, if the metric type is inner product?

▸ What is normalization? Why is normalization needed?

▸ Why do I get different results using Euclidean distance (L2) and inner product (IP) as the distance metric?

> metric.md was last updated at 2021-06-29 11:48:04: 1.1.0 addhome (#270)