Similarity Search

Estimated Reading Time: 30 minutes

🔍 What is Similarity Search?

Similarity search is the process of finding items in a dataset that are most similar to a given query item. It's widely used in:

- **Recommendation systems** (for example, suggesting similar movies)
- **Image and video retrieval**
- **Natural language processing** (for example, finding similar documents or sentences)
- **Biometrics** (for example, face recognition)

At the core of similarity search is a **distance or similarity metric** that quantifies how alike two data points are. The choice of metric depends on the nature of the data and the application.

Some background Math

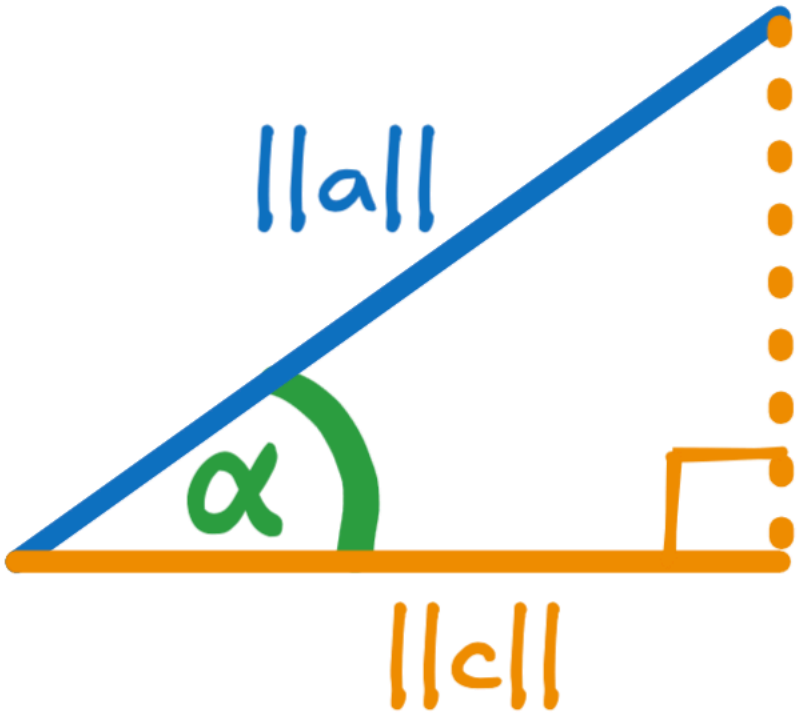
In this section, we'll go over some basic math concepts that will help explain how different distance metrics work. Don't worry if the details feel a bit complex, we won't be testing you on the math. These explanations are just here to give you a better sense of how and why the metrics differ from one another.

What is the cosine of an angle?

Recall that the cosine of an angle is defined as the ratio of the length of the adjacent side to the length of the hypotenuse:

$$cos(\alpha) = \frac{adjacent}{hypotenuse}\cos(\alpha) = \text{hypotenuseadjacent}$$

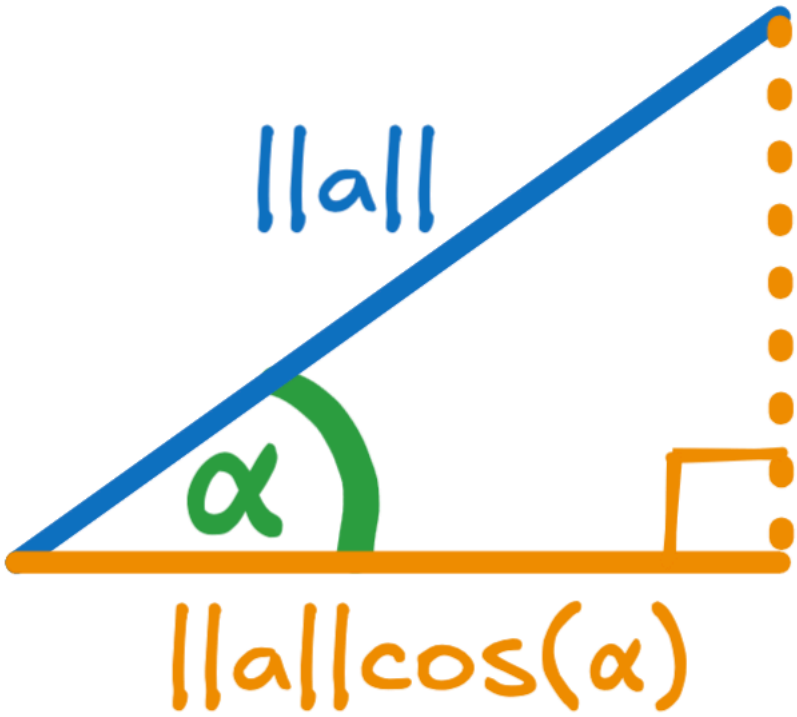
Diagramatically, this can be represented as follows:



where the length of the adjacent side is$||c||||c||$and the length of they hypotenuse is$||a||||a||$. Why we chose$||c||||c||$and$||a||||a||$to represent the lengths of the sides of this right-angle triangle will be apparent later on.

For the above triangle, the cosine of angle$\alpha\alpha$is given by the following formula:$cos(\alpha) = \frac{||c||}{||a||}\cos(\alpha) = ||a||||c||$.

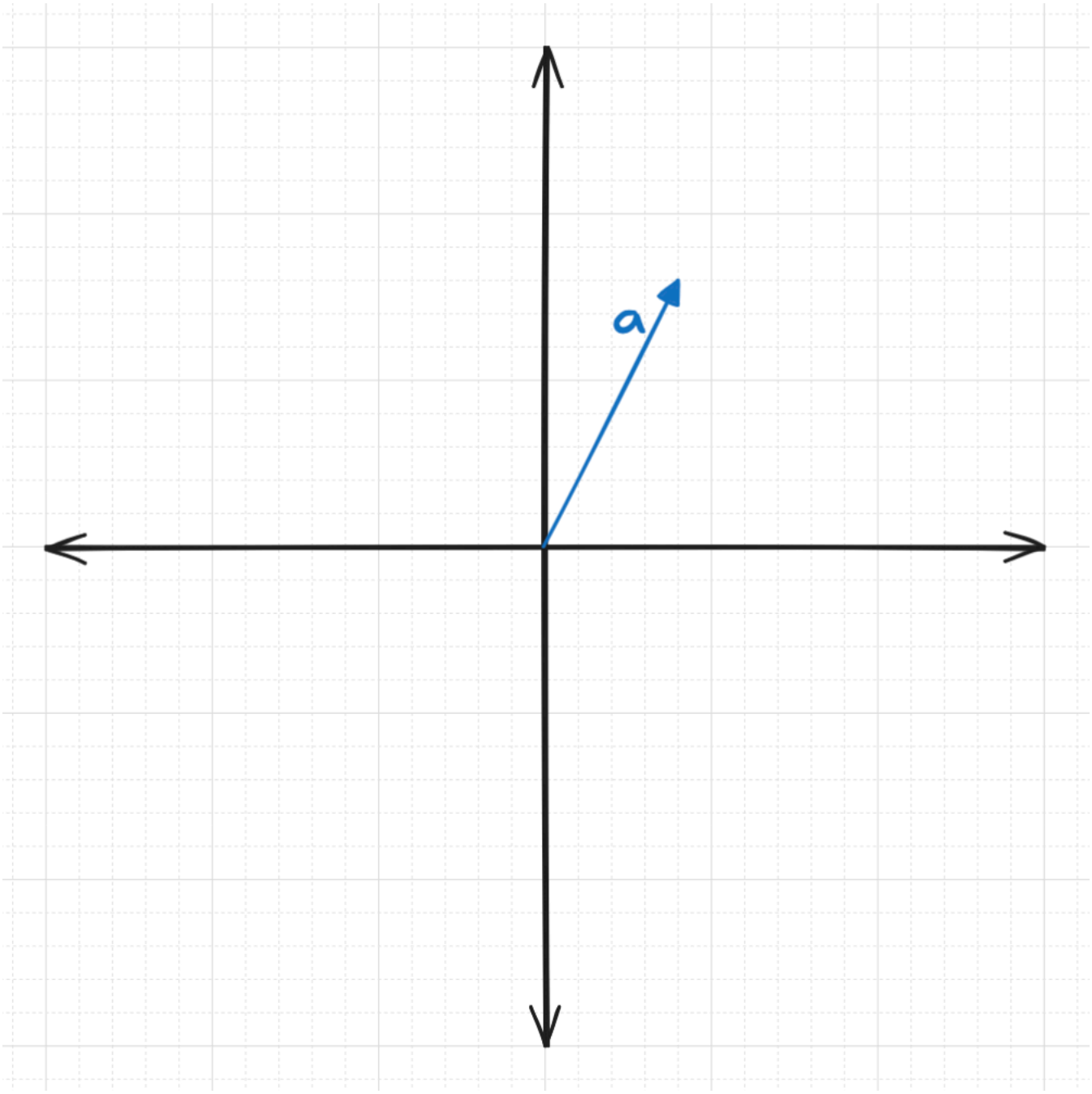
Of course, it is possible to rearrange the above formula to solve for the length of the adjacent side$||c||||c||$ in order to express it in terms of the length of the hypotenuse$||a||||a||$and the cosine of the angle$\alpha\alpha$: $||c|| = ||a||cos(\alpha)||c|| = ||a||\cos(\alpha)$. Replacing$||c||||c||$with$||a||cos(\alpha)$ $||a||\cos(\alpha)$results in the following diagram:



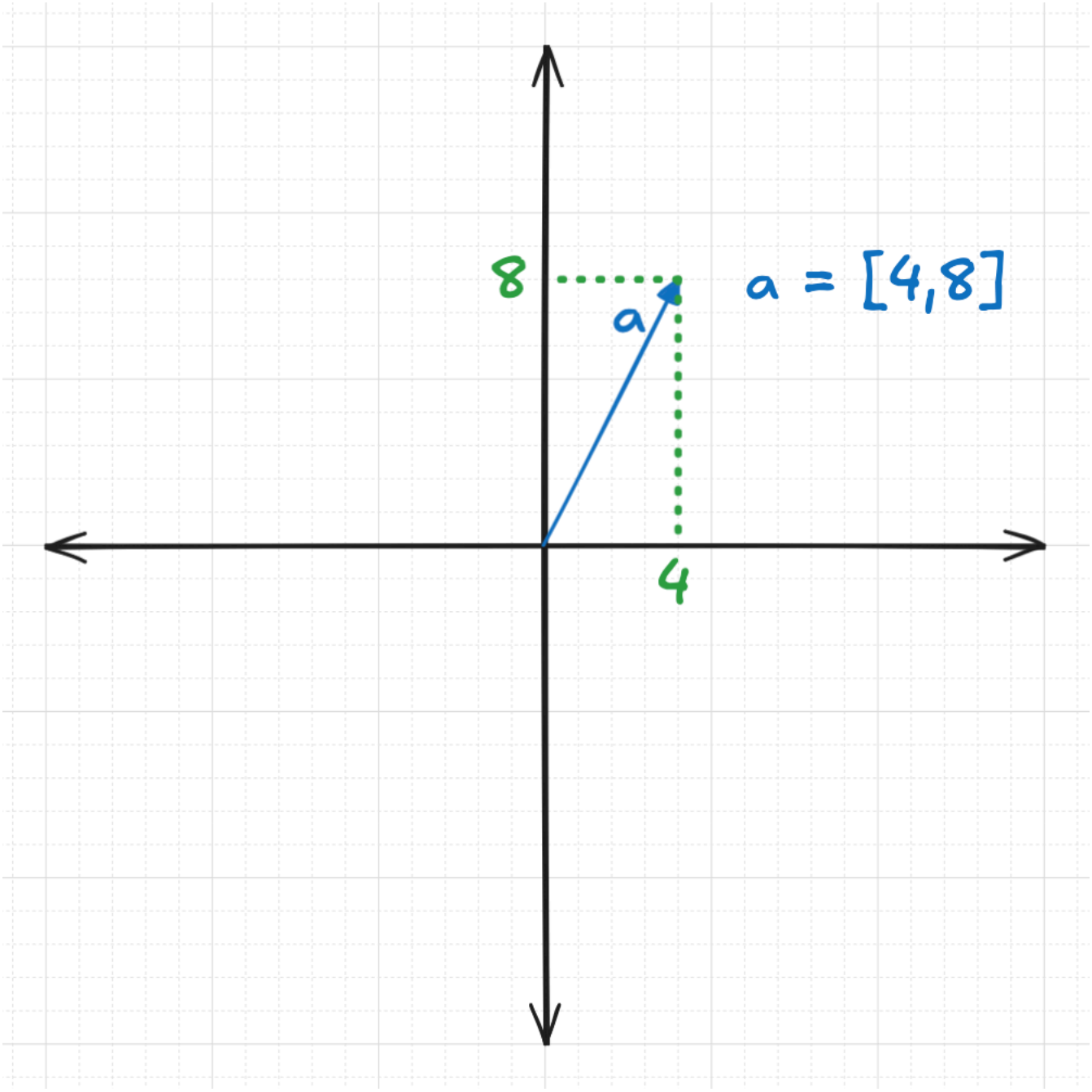
What is a vector?

A vector is a geometric object that has a length and a direction. It can be represented on a Cartesian plane that has the same number of dimensions as the number of components in the vector. For instance a vector with two components can be represented by a line from the origin on a 2-dimensional Cartesian plane, and a vector with three components can be represented by a line from the origin on a 3-dimensional Cartesian plane.

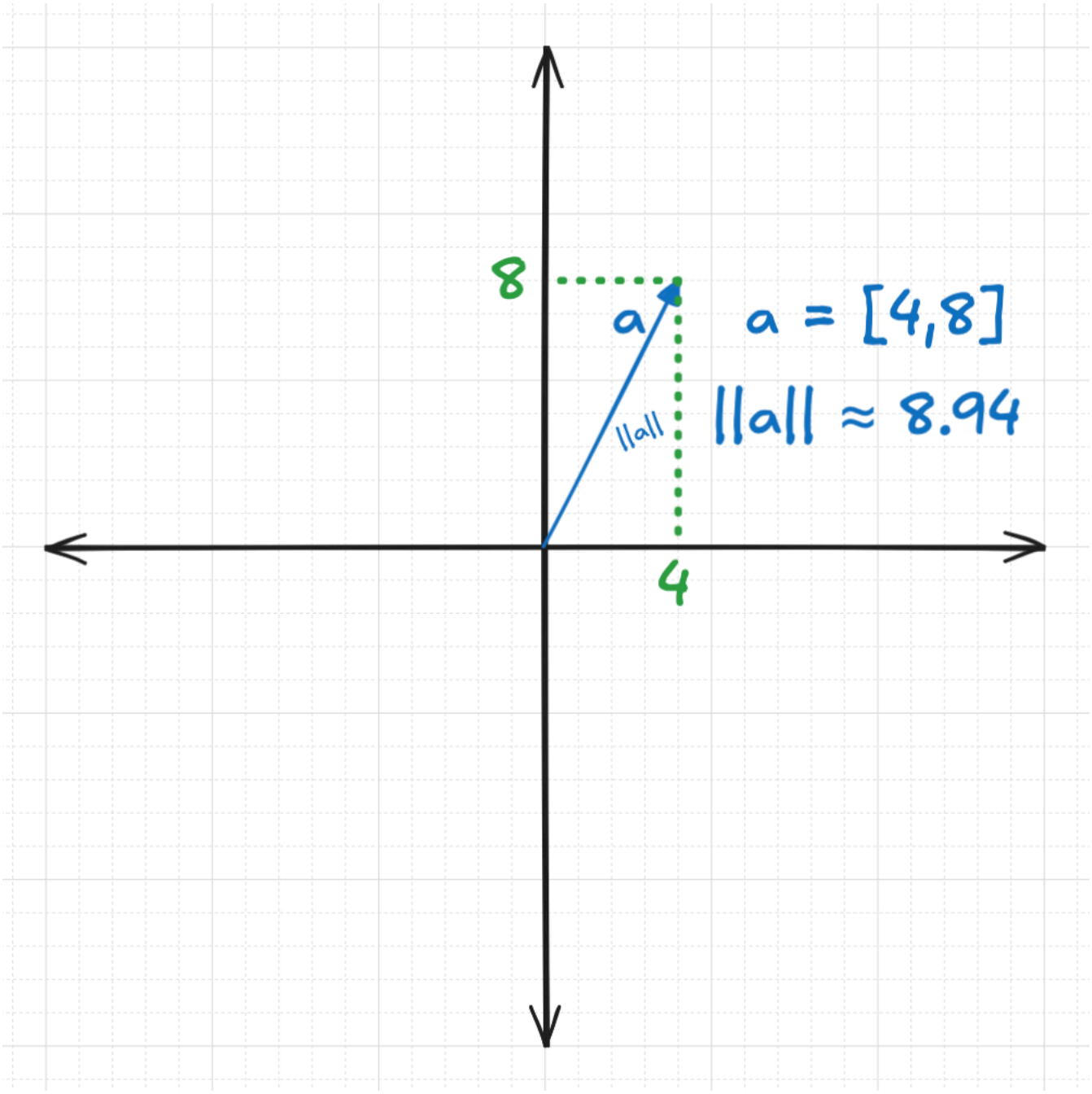
Consider the vector$a = [4, 8]a = [4, 8]$. This vector has two components, and can thus be represented on the 2 dimensional Cartesian plane using the following diagram:



It is important to remember that $\boldsymbol{a}$ is a *vector* not a *length* or magnitude. For clarity, the following diagram clearly indicates the vector components:



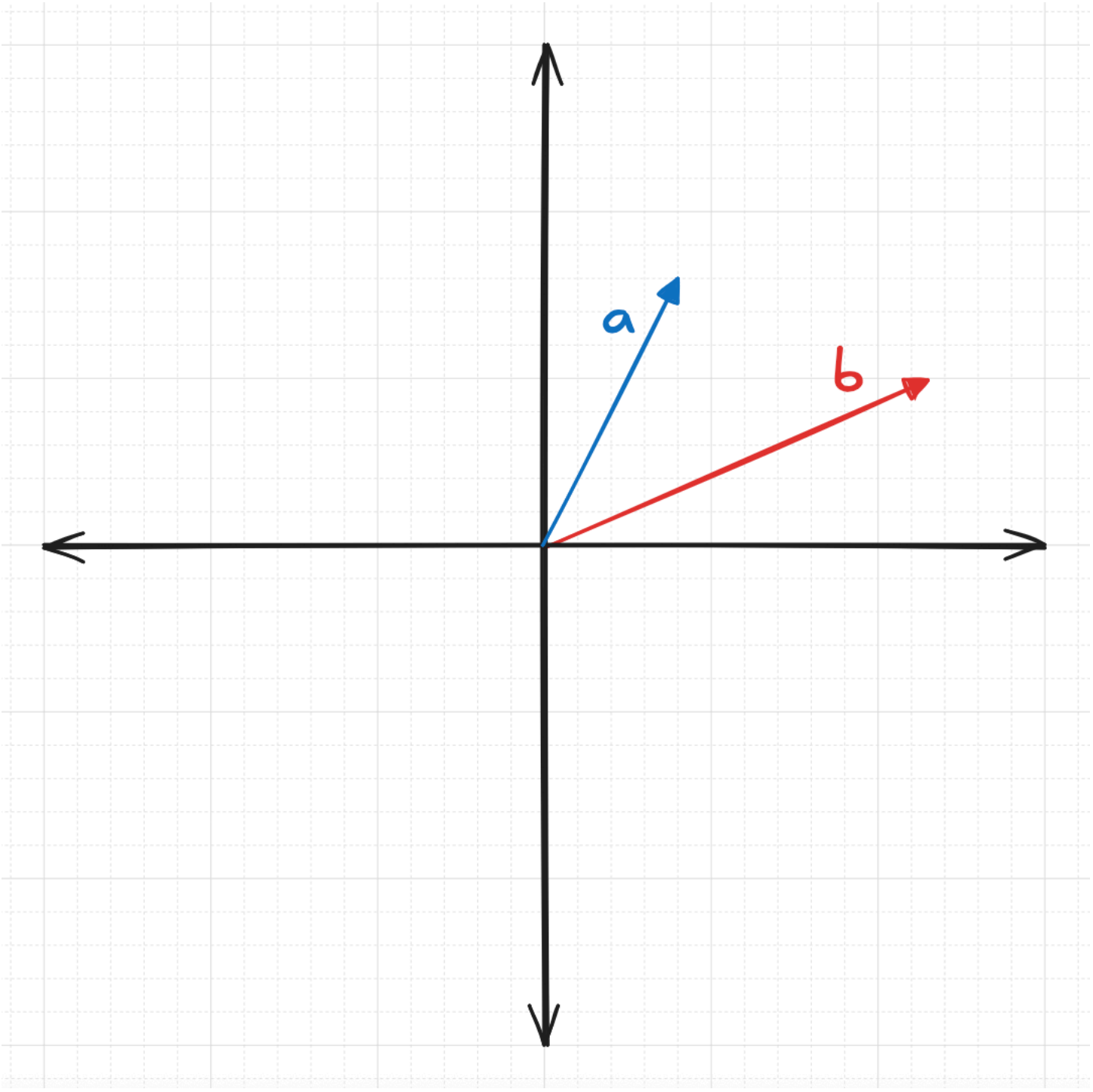
The magnitude of vector$a$acan be calculated using the L2 norm, which is also known as the Euclidean norm:$||a|| = \sqrt{\sum_{k=1}^{n} a_k^2}$||a|| = ∑k=1n ak2
√
, where$a_k$ak represents the k'th component of vector$a$a, and$n$nrepresents the total number of components. For a vector with just two components,$x$xand$y$y, the magnitude of vector is given by the Pythagorean theorem:$||a|| = \sqrt{x^2 + y^2}$||a|| = x2 + y2
√
. For instance, for the vector$a = [4, 8]$a = [4, 8]the magnitude is$||a|| = \sqrt{4^2 + 8^2} \approx 8.94$||a|| = 42 + 82
√
≈ 8.94:



When vectors are used as embeddings, the direction of the vector typically encodes the semantic meaning or topic, while the magnitude (or length) of the vector can sometimes reflect the intensity, confidence, or salience of that meaning—such as how popular a product is or how authoritative a source might be.

Multiple vectors

Multiple vectors can be plotted on the same Cartesian plane. The following diagram shows two vectors on a 2-dimensional plane: vector$a = [4, 8]$a = [4, 8]and vector$b = [11.5, 5]$b = [11.5, 5]:



The objective of a distance function is to provide a number that indicates how far apart these vectors are, with greater magnitudes indicating that the vectors are farther apart. In contrast, the objective of a similarity function is to provide a number that reflects how similar these vectors are, with greater numbers indicating that the vectors are more similar. In many instances, a similarity metric can be converted to a distance metric (and vice versa) using a very simple formula. Let's analyze some of the more common distance and similarity metrics.

📏 Common Distance and Similarity Metrics

1. L2 Distance (Euclidean Distance)

Definition:

The L2 distance between two vectors$a$aand$b$bis the square root of the sum of the squared differences between corresponding elements:

$$\text{L2}(a, b) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2} \text{L2(a, b)} = \quad \sum \text{i=1 n(ai} - \text{bi)2}$$

$$\sqrt{}$$

Example 1:

Let's see how we can calculate the L2 distance between vectors $a = [4, 8]$ a = [4, 8] and vector $b = [11.5, 5]$ b = [11.5, 5]:

$$L2(a, b) = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2} \quad \text{L2(a, b)} = \sqrt{\sum \text{i=1 n(ai − bi)2}}$$

$$L2(a, b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2} \quad \text{L2(a, b)} = \sqrt{\text{(a1 − b1)2 + (a2 − b2)2}}$$

$$L2(a, b) = \sqrt{(4 - 11.5)^2 + (8 - 5)^2} \quad \text{L2(a, b)} = \sqrt{\text{(4 − 11.5)2 + (8 − 5)2}}$$

$$L2(a, b) \approx 8.08 \quad \text{L2(a, b)} \approx 8.08$$

Example 2:

For higher dimensions we proceed analogously. For instance, to calculate the L2 distance between vectors $q = [1, 2, 3]$ q = [1, 2, 3] and $r = [4, 5, 6]$ r = [4, 5, 6]:

$$L2(q, r) = \sqrt{\sum_{i=1}^{n} (q_i - r_i)^2} \quad \text{L2(q, r)} = \sqrt{\sum \text{i=1 n(qi − ri)2}}$$

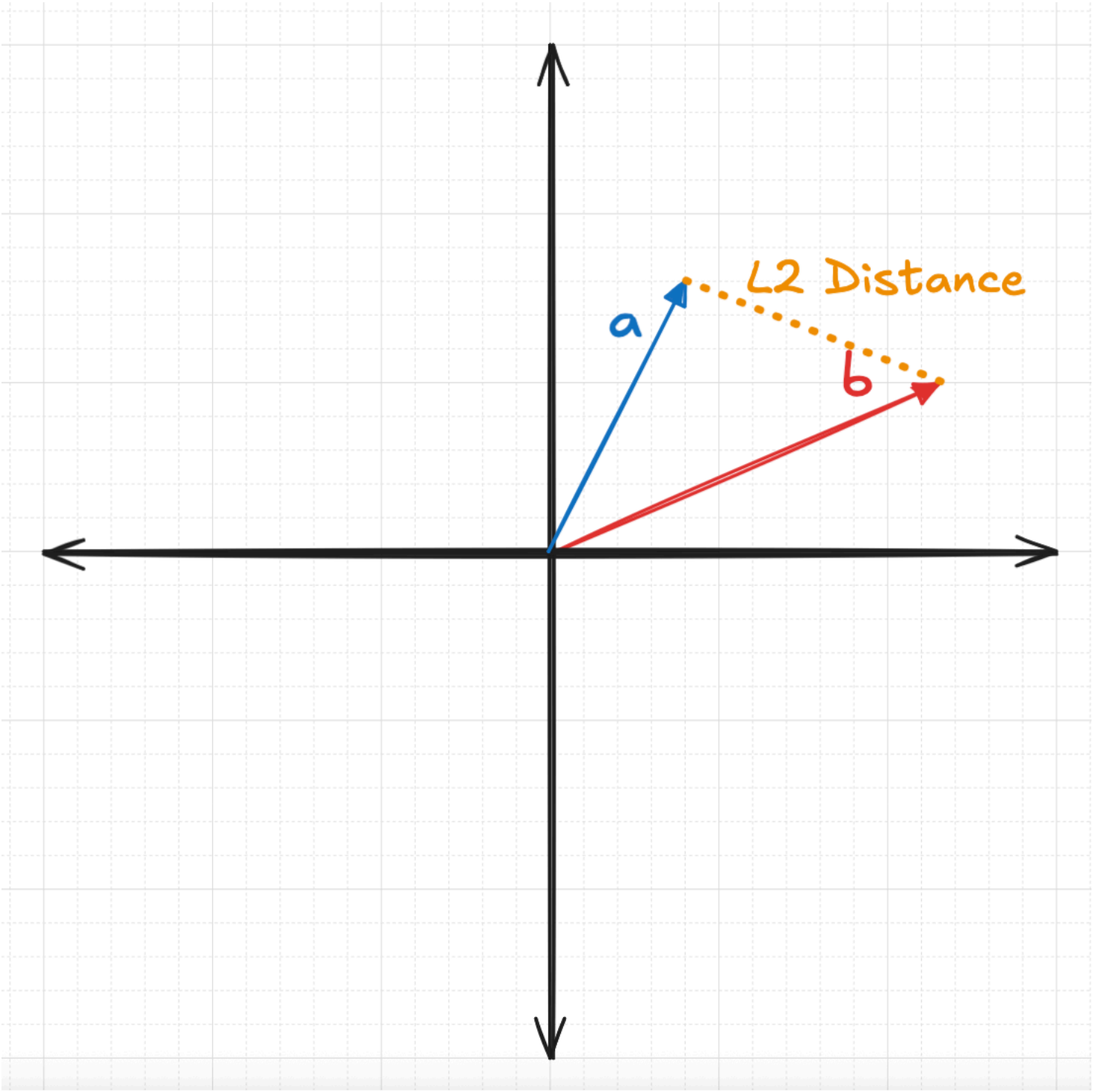
$$L2(q, r) = \sqrt{(q_1 - r_1)^2 + (q_2 - r_2)^2 + (q_3 - r_3)^2} \quad \text{L2(q, r)} = \sqrt{\text{(q1 − r1)2 + (q2 − r2)2 + (q3 − r3)2}}$$

$$L2(q, r) = \sqrt{(1 - 4)^2 + (2 - 5)^2 + (3 - 6)^2} \quad \text{L2(q, r)} = \sqrt{\text{(1 − 4)2 + (2 − 5)2 + (3 − 6)2}}$$

$$L2(q, r) = \approx 5.20 \quad \text{L2(q, r)} = \approx 5.20$$

Diagram:

Diagramatically, the L2 distance can be represented as the straight-line distance between the tips of the arrows representing vectors in the Cartesian plane:



Properties:

- The L2 distance measures the straight-line distance between two points in Euclidean space, following the principles of the Pythagorean theorem.
- It is sensitive to both the magnitude and direction of the vectors, meaning that changes in either can significantly affect the calculated distance.
- This distance metric is commonly used in applications involving spatial or geometric data, such as image analysis, computer vision, and geographic mapping.

Use Case:

- A common use case for L2 distance is determining the closest point to a given location in two-dimensional (2D) or three-dimensional (3D) space. This approach is frequently used in computer vision tasks, where spatial proximity between features or objects plays a critical role.

---

2. Dot Product (Inner Product) Similarity and Distance

Definition:
The dot product of two vectors is:

$a \cdot b = \sum_{i=1}^{n} a_i b_i$a · b = ∑i=1n aibi

Example 1:

Let's see how we can calculate the dot product between vectors$a = [4, 8]$a = [4, 8]and vector$b = [11.5, 5]$b = [11.5, 5]:

$a \cdot b = \sum_{i=1}^{n} a_i b_i$a · b = ∑i=1n aibi
$a \cdot b = a_1 b_1 + a_2 b_2$a · b = a1b1 + a2b2
$a \cdot b = 4 \times 11.5 + 8 \times 5$a · b = 4 × 11.5 + 8 × 5
$a \cdot b = 86$a · b = 86

Example 2:

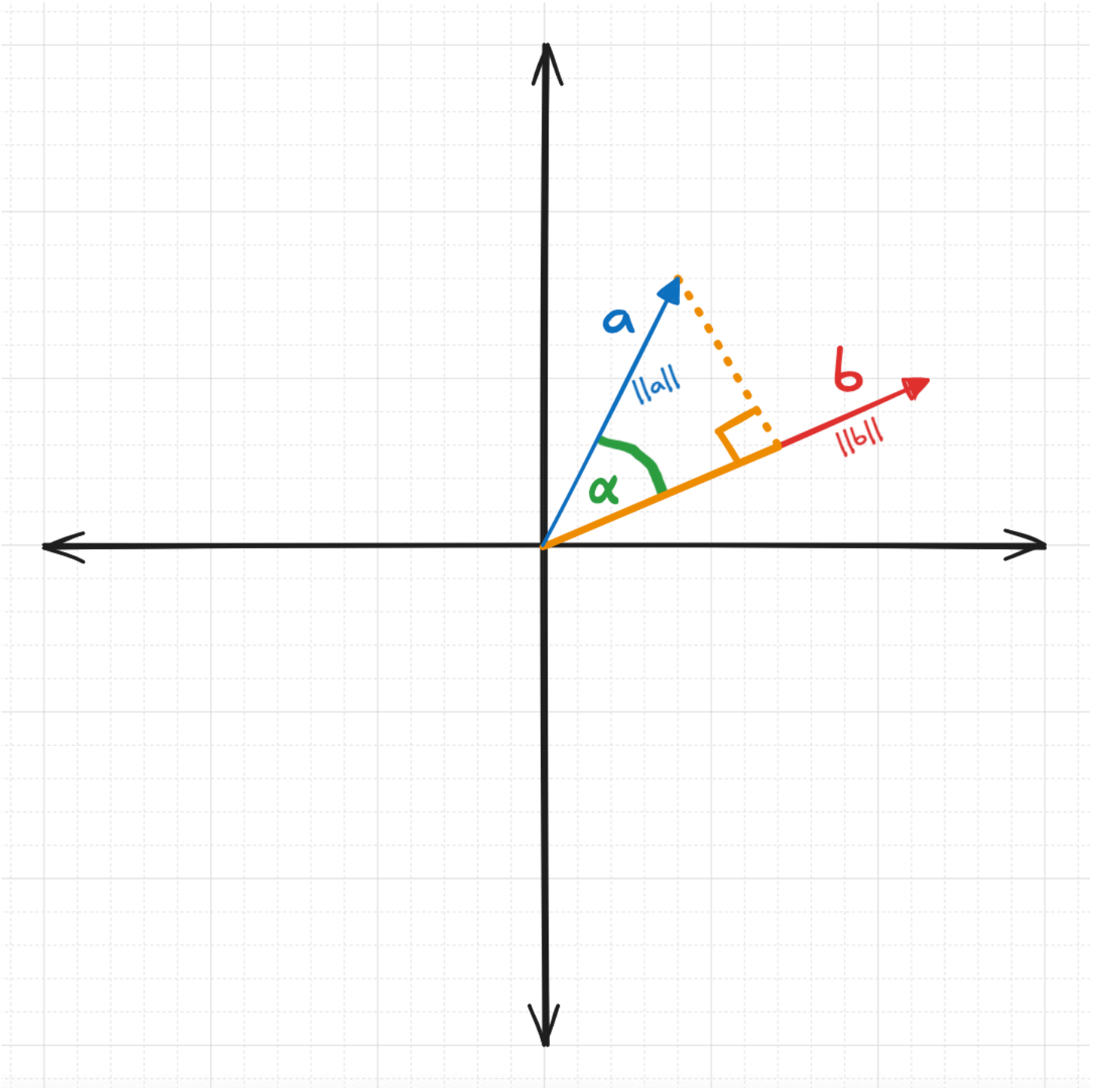
For higher dimensions we proceed analogously. For instance, to calculate the dot product between vectors$q = [1, 2, 3]$q = [1, 2, 3]and$r = [4, 5, 6]$r = [4, 5, 6]:

$q \cdot r = \sum_{i=1}^{n} q_i r_i$q · r = ∑i=1n qiri
$q \cdot r = q_1 r_1 + q_2 r_2 + q_3 r_3$q · r = q1r1 + q2r2 + q3r3
$q \cdot r = 1 \times 4 + 2 \times 5 + 3 \times 6$q · r = 1 × 4 + 2 × 5 + 3 × 6
$q \cdot r = 32$q · r = 32

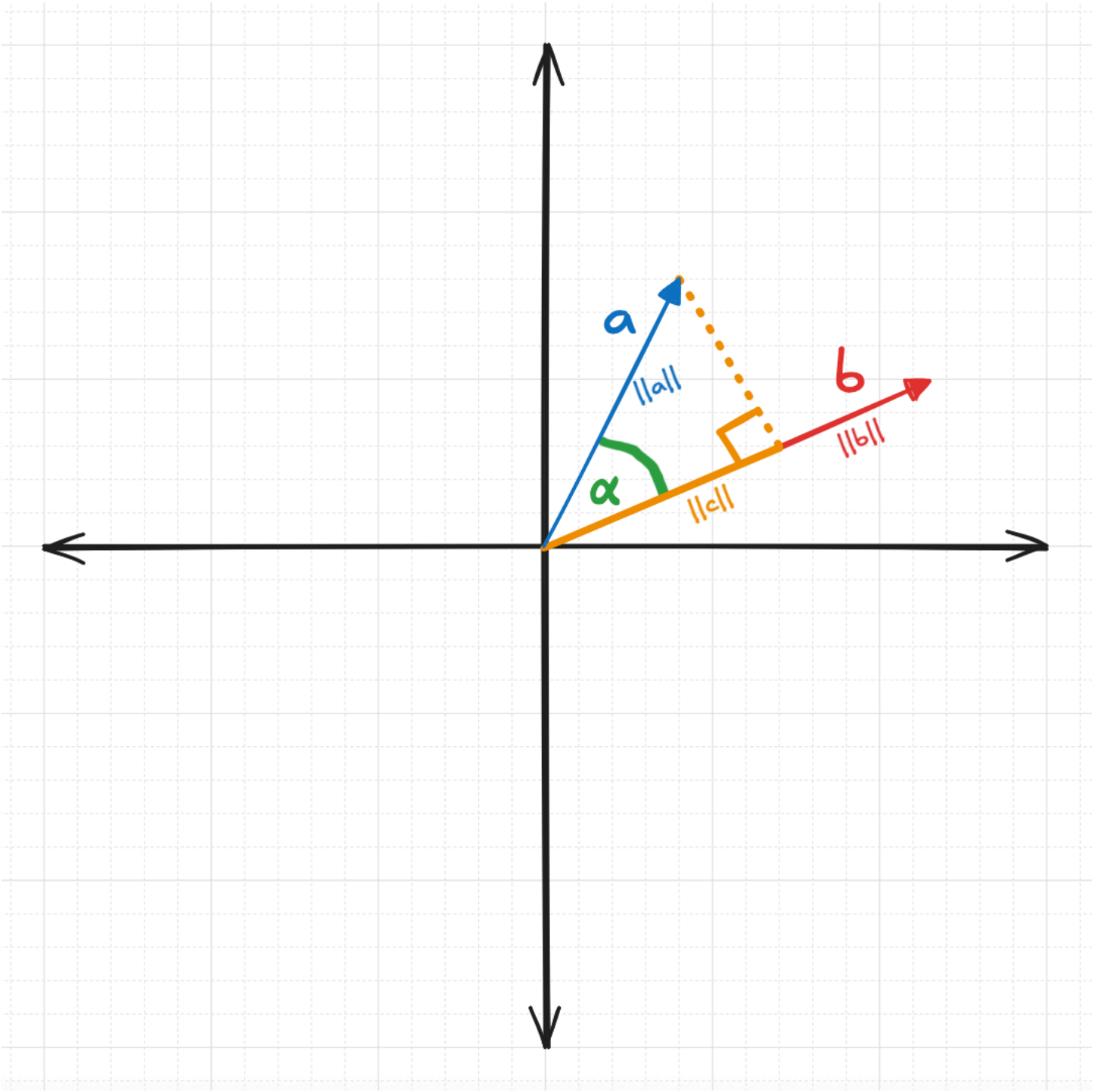
Note: Dot product is a similarity metric: the greater the dot product, the more similar the vectors are. In order to convert dot product to a distance, one calculates the negative of the dot product: the greater the negative dot product, the greater the distance between two vectors.

Alternative Calculation:

We want to compute the dot product using only the magnitudes of the vectors. Ideally, this would involve simply multiplying the magnitudes together. However, since the vectors may point in different directions, we must account for the angle between them. To do this, we can project one vector onto the direction of the other before performing the multiplication. Consider the following diagram that projects the vector$a = [4, 8]$a = [4, 8]onto vector$b = [11.5, 5]$b = [11.5, 5]:



On the above image vector$a$a's projection onto vector$b$bresults in a vector pointing in the same direction as$b$b represented by the orange chord that overlays part of vector$b$b. Let's define the length of this projection vector as$||c||$||c||:



It turns out that we can calculate the dot product $a \cdot b$ a · b by multiplying the length of vector b b($||b||$ ||b||) by the length of the projection of a a onto b b($||c||$ ||c||):

$$a \cdot b = ||b||\ ||c||$$ a · b = ||b|| ||c||

Now, let α α represent the angle between vectors a a and b b as shown in the above diagram. Note that the projection of a a onto b b forms a right-angle triangle with length $||c||$ ||c|| forming an adjacent side for angle α α and the length $||a||$ ||a|| forming the hypotenuse. This allows us to use the following equation for the cosine of an angle:

$$cos(\alpha) = \frac{adjacent}{hypotenuse}$$ cos(α) = hypotenuseadjacent

$$cos(\alpha) = \frac{||c||}{||a||}$$ cos(α) = ||a|| ||c||

Rearranging the above, we can solve for$||c||||c||$:

$$||c|| = ||a||cos(\alpha)||c|| = ||a||\cos(\alpha)$$

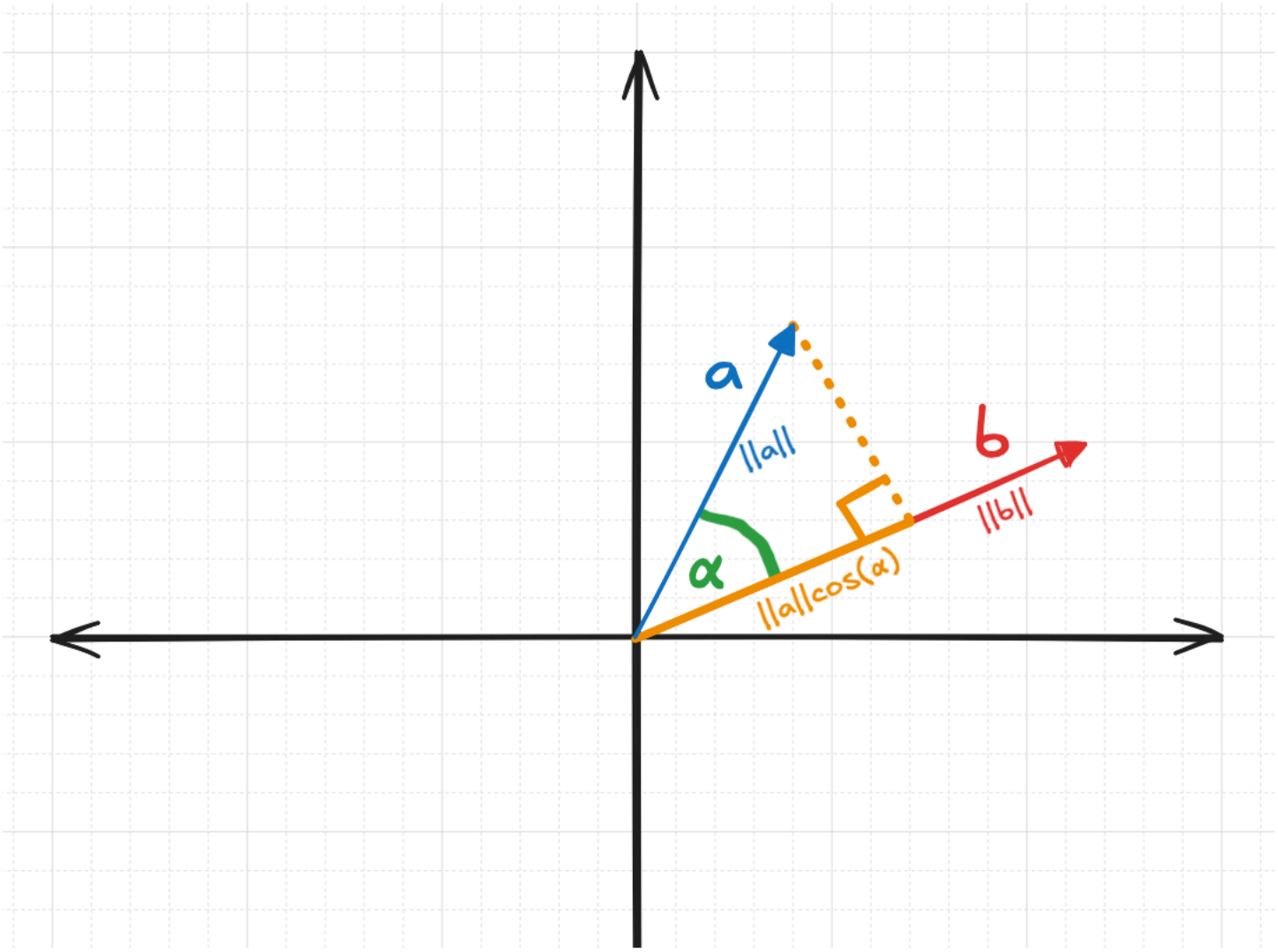
Finally, replacing$||c||||c||$with the$||a||cos(\alpha)||a||\cos(\alpha)$ in the calculation of the dot product:

$$a \cdot b = ||b||\ ||c||a \cdot b = ||b||\ ||c||$$

$$a \cdot b = ||b||\ ||a||cos(\alpha)a \cdot b = ||b||\ ||a||\cos(\alpha)$$

$$a \cdot b = ||a||\ ||b||cos(\alpha)a \cdot b = ||a||\ ||b||\cos(\alpha)$$

In other words, the dot product$a \cdot ba \cdot b$can be calculated by multiplying the length of$aa$by the length ofbb and the cosine of angle$\alpha\alpha$betweenaaandbb. Intuitively, the dot product is the multiplication of the lengths of two vectors after one of the vectors has been projected onto the other using the cosine of the angle in order to account for the fact that the vectors point in different directions. This concept can be partially portrayed using the following diagram:



In the above image, the length of the projection of$a$aonto$b$bis depicted as$||a||cos(\alpha)$||a||cos(α), and the length of vector$b$b is indicated as$||b||$ ||b||.The dot product$a \cdot b$a · bis just the product of the lengths$||b||$||b||and$||a||cos(\alpha)$||a||cos(α), or, when rearranged,$||a||\ ||b||cos(\alpha)$ ||a|| ||b||cos(α).

Example of alternative calculation:

Let's see how we can calculate the dot product between vectors$a = [4, 8]$a = [4, 8], vector$b = [11.5, 5]$b = [11.5, 5], and angle $\alpha = 39.94°$α = 39.94∘:

$$||a|| = \sqrt{\sum_{k=1}^{n} a_k^2}||a|| = \quad \sqrt{\sum k=1n\ ak2}$$

$$||a|| = \sqrt{4^2 + 8^2}||a|| = \quad \sqrt{42 + 82}$$

$$||a|| \approx 8.94||a|| \approx 8.94$$

$$||b|| = \sqrt{\sum_{k=1}^{n} b_k^2}||b|| = \quad \sqrt{\sum k=1n\ bk2}$$

$$||b|| = \sqrt{11.5^2 + 5^2}||b|| = \quad \sqrt{11.52 + 52}$$

$$||b|| \approx 12.54||b|| \approx 12.54$$

$$cos(39.94) \approx 0.767\cos(39.94) \approx 0.767$$

$$a \cdot b = ||a||\ ||b||cos(\alpha)a \cdot b = ||a||\ ||b||\cos(\alpha)$$

$$a \cdot b \approx 8.94 \times 12.54 \times 0.767a \cdot b \approx 8.94 \times 12.54 \times 0.767$$

$$a \cdot b \approx 85.99a \cdot b \approx 85.99$$

Which, after accounting for rounding errors, is equivalent to the length 86 we calculated before using the formula$a \cdot b = \sum_{i=1}^{n} a_i b_i$a · b = ∑i=1n ai bi.

Properties:

- The dot product of two vectors can be positive, negative, or zero, depending on the angle between them.
- Larger dot product values typically indicate a higher degree of similarity between the vectors, especially when they are pointing in similar directions.
  - If a **distance-like metric** is needed, the **negative of the dot product** can be used. In this case, larger (more negative) values correspond to greater dissimilarity or distance.
- The dot product is sensitive to both the magnitude and direction of the vectors, meaning that changes in either will affect the result.
- It is frequently used in machine learning models, such as in neural networks for computing activations or in matrix factorization for recommendation systems.

Use Case: When the length of the vector (representing something such as relevance, confidence, or quantity) is meaningful. For instance, in recommendation systems, the direction of vectors typically indicates two products are about the same topic, but larger magnitudes might indicate that a product is more popular. In this case using dot product similarity makes sense, because one would want to recommend the products that are not only about the same topic, but also popular.

---

3. Cosine Similarity and Distance

Definition:
Cosine similarity measures the cosine of the angle between two vectors:

$$\text{cosine\_similarity}(a, b) = \frac{a \cdot b}{||a||\ ||b||}\text{cosine\_similarity}(a, b) = ||a||\ ||b||a \cdot b$$

Note: Cosine similarity is a similarity metric: the greater the cosine similarity, the more similar the vectors are. In order to convert cosine similarity to cosine distance, use one minus cosine similarity:
$\text{cosine\_distance}(a, b) = 1 - \text{cosine\_similarity}(a, b)\text{cosine\_distance}(a, b) = 1 - \text{cosine\_similarity}(a, b)$. The greater the cosine distance, the less similar the vectors are.

Alternative Calculation:
Cosine similarity can be calculated even more efficiently if the vectors are normalized. In order to normalize a vector, divide it by its L2 norm:

$norm(a) = \frac{a}{||a||}$norm(a) = ||a||a

For instance, for the vector$a = [4, 8]$a = [4, 8]the normalized vector can be calculated as follows:

$||a|| = \sqrt{\sum_{k=1}^{n} a_k^2}||a|| = \quad \sqrt{\sum k=1 n\ ak 2}$

$||a|| = \sqrt{4^2 + 8^2}||a|| = \quad \sqrt{42 + 82}$

$||a|| \approx 8.94||a|| \approx 8.94$

$norm(a) = \frac{a}{||a||}$norm(a) = ||a||a

$norm(a) \approx \frac{[4, 8]}{8.94}$norm(a) ≈ 8.94[4,8]

$norm(a) \approx \left[\frac{4}{8.94}, \frac{8}{8.94}\right]$norm(a) ≈ [8.944, 8.948]

$norm(a) \approx [0.448, 0.895]$norm(a) ≈ [0.448, 0.895]

A normalized vector has the property that the sum of the squared components sums to one:

$\sum_{i=1}^{n} norm(a)_i^2 = 1\sum i=1 n\ norm(a)i2 = 1$

For instance, for the normalized vector$norm(a) = [0.448, 0.895]$norm(a) = [0.448, 0.895]:

$\sum_{i=1}^{n} norm(a)_i^2 = 0.448^2 + 0.895^2 \approx 1\sum i=1 n\ norm(a)i2 = 0.4482 + 0.8952 \approx 1$

Calculating cosine similarity between two normalized vectors is as easy as calculating the dot product between them:

$\text{cosine_similarity}(a, b) = \frac{a \cdot b}{||a||\ ||b||}$cosine_similarity(a, b) = ||a|| ||b||a·b

$\text{cosine_similarity}(a, b) = \frac{a}{||a||} \cdot \frac{b}{||b||}$cosine_similarity(a, b) = ||a||a · ||b||b

$\text{cosine_similarity}(a, b) = norm(a) \cdot norm(b)$cosine_similarity(a, b) = norm(a) · norm(b)

Note that many embedding models normalize vectors by default, as this allows cosine similarity or distance to be efficiently computed using the dot product. Since cosine-based comparisons are common and the dot product is faster to compute, storing normalized vectors is a practical choice when only cosine-based comparisons are needed.

Properties:

- Cosine similarity focuses on the orientation of vectors rather than their magnitude, measuring the cosine of the angle between them.
- It is particularly well-suited for high-dimensional, sparse data, such as text embeddings or term-frequency vectors in natural language processing.
- This metric is invariant to vector length, meaning that scaling a vector up or down does not affect the similarity score.

Use Case:

- A common use case for cosine similarity is measuring document similarity in natural language processing (NLP), where it helps identify texts with similar content regardless of their length.

🧠 Choosing the Right Metric

Metric	Sensitive to Magnitude	Normalized	Best For
L2 Distance	✅ Yes	❌ No	Spatial data, clustering
Cosine Distance	❌ No	✅ Yes	Text, embeddings, NLP

Metric	Sensitive to Magnitude	Normalized	Best For
Dot Product	✅ Yes	❌ No	Neural networks, recommender systems

🧪 Practical Considerations

- **Normalization**
  - Normalize vectors if you expect to only need to calculate cosine similarity. For many natural language processing tasks and text embedding models, this is the default option.
- **High-Dimensional Data**
  - L2 distance can suffer from the 'curse of dimensionality.' Consider using a different metric or using a dimensionality reduction technique if the data exhibits high dimensionality.
  - Cosine distance often performs better in high dimensions, and is the default choice in many natural language processing and text-based tasks.
  - Dot product can be computed efficiently using matrix operations, and can be used to calculate cosine similarity if the vectors are normalized.

Conclusion

Choosing the right distance metric is crucial for effective similarity search. L2 distance works well for continuous, lower-dimensional data where magnitude matters. Cosine distance excels with high-dimensional, sparse data where direction is more important than magnitude. Dot product offers computational efficiency and is useful when both magnitude and direction contribute to similarity. Understanding these trade-offs helps in selecting the most appropriate metric for your specific use case.

Author(s)

Wojciech "Victor" Fulmyk