inner product

Authors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

Synonyms

scalar product, dot product

Abstract

Many machine learning (ML) applications involve data points whose features form numeric arrays. Examples include the color intensities of image pixels, the amplitudes at regular intervals of a sensor signal, and the embeddings of tokens used by large language models (LLMs). These numeric arrays can be naturally represented by vectors in some vector space. Many ML methods rely on measuring the similarity between vectors that represent data points. Each neuron of an artificial neural network (ANN) matches its input, for example the color intensities of an image, against a learned template. The attention unit of an LLM compares the query vector of a token with the key vector of another token, both formed from the embeddings of those tokens. Both comparisons compute an inner product: a number assigned to a pair of vectors in a vector space that measures their similarity.

Definition

Data points are often characterized by features that form a numeric array: the color intensities of the pixels of an image, the amplitudes of a sensor signal at regular intervals, or the embedding of a token in a large language model (LLM). Such an array can be represented by a vector, i.e., an element of a vector space. Many machine learning (ML) methods compare two data points by measuring how similar the vectors that represent them are. One principled measure of this similarity is an inner product.

Consider a vector space $\mathcal{V}$ over the field of real numbers $\mathbb{R}$. An inner product in $\mathcal{V}$ is a function

$$\langle \cdot, \cdot \rangle : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$$

that satisfies the following properties for all vectors $\mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathcal{V}$ and all scalars $\beta \in \mathbb{R}$ (Axler, 2024, Def. 6.2):

- Symmetry: $\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle$;
- Linearity in the first argument: $\langle \beta \mathbf{u} + \mathbf{w}, \mathbf{v} \rangle = \beta \langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{w}, \mathbf{v} \rangle$;
- Positive-definiteness: $\langle \mathbf{u}, \mathbf{u} \rangle \geq 0$, with equality if and only if $\mathbf{u} = \mathbf{0}$.

For a vector space over the field of complex numbers $\mathbb{C}$, symmetry is replaced by conjugate symmetry, $\langle \mathbf{u}, \mathbf{v} \rangle = \overline{\langle \mathbf{v}, \mathbf{u} \rangle}$, while linearity in the first argument is retained; the inner product is then conjugate-linear in the second argument, $\langle \mathbf{u}, \beta\mathbf{v} + \mathbf{w} \rangle = \overline{\beta}\langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{u}, \mathbf{w} \rangle$ (Axler, 2024, Thm. 6.6). Positive-definiteness needs no change, because conjugate symmetry makes $\langle \mathbf{u}, \mathbf{u} \rangle$ real: applied to $\mathbf{v} = \mathbf{u}$, it gives $\langle \mathbf{u}, \mathbf{u} \rangle = \overline{\langle \mathbf{u}, \mathbf{u} \rangle}$, and a number equal to its own conjugate is real, so the inequality $\langle \mathbf{u}, \mathbf{u} \rangle \geq 0$ is meaningful in the complex case as well. Two vectors $\mathbf{u}, \mathbf{v} \in \mathcal{V}$ are called orthogonal if their inner product is zero, $\langle \mathbf{u}, \mathbf{v} \rangle = 0$.

The pair $(\mathcal{V}, \langle \cdot, \cdot \rangle)$ is called an inner product space. Each inner product induces a norm via $\|\mathbf{u}\| := \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle}$ for all $\mathbf{u} \in \mathcal{V}$, which in turn induces a metric via $d(\mathbf{u}, \mathbf{v}) := \|\mathbf{u} - \mathbf{v}\|$ for all $\mathbf{u}, \mathbf{v} \in \mathcal{V}$.

For the Euclidean space $\mathcal{V} = \mathbb{R}^d$, the standard inner product, also called the dot product, is $\langle \mathbf{x}, \mathbf{x}' \rangle = \mathbf{x}^\top \mathbf{x}' = \sum_{j=1}^{d} x_j x'_j$ (Strang, 2016, Sect. 1.2). Geometrically, the inner product determines orthogonal projections: the orthogonal projection of $\mathbf{x}'$ onto the direction of a nonzero vector $\mathbf{x}$ has the signed length $\langle \mathbf{x}, \mathbf{x}' \rangle / \|\mathbf{x}\|$ (Axler, 2024, Example 6.56) (see Fig. 1).

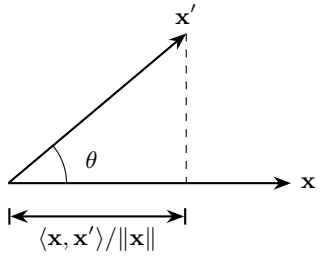

Figure 1: Geometry of the standard inner product in the Euclidean space $\mathbb{R}^2$. The dashed ruler drops from $\mathbf{x}'$ perpendicularly onto the direction of $\mathbf{x}$ and meets that direction at the orthogonal projection of $\mathbf{x}'$. This orthogonal projection lies at the signed distance $\langle \mathbf{x}, \mathbf{x}' \rangle / \|\mathbf{x}\|$ from the origin, marked by the double arrow. The angle θ between the two vectors is defined via $\cos\theta := \langle \mathbf{x}, \mathbf{x}' \rangle / \big(\|\mathbf{x}\| \|\mathbf{x}'\|\big)$. The inner product measures the alignment of the two vectors and vanishes when they are orthogonal

The inner product also defines the angle θ between two nonzero vectors $\mathbf{x}, \mathbf{x}' \in \mathcal{V}$, via $\cos\theta := \langle \mathbf{x}, \mathbf{x}' \rangle / \big(\|\mathbf{x}\| \|\mathbf{x}'\|\big)$. The Cauchy-Schwarz inequality $|\langle \mathbf{x}, \mathbf{x}' \rangle| \leq \|\mathbf{x}\| \|\mathbf{x}'\|$ ensures that this ratio lies in $[-1, 1]$ (Axler, 2024, Thm. 6.14). Here cos is the cosine function, which decreases strictly from $\cos 0 = 1$ to $\cos \pi = -1$ and therefore takes each value of $[-1, 1]$ exactly once on the interval $[0, \pi]$. For the Euclidean plane $\mathbb{R}^2$, as illustrated by Fig. 1, the angle defined in this way agrees with the elementary geometric angle between two vectors (Strang, 2016, Sect. 1.2).

For two vectors, the sign and magnitude of their inner product can be read off the angle between these vectors. The inner product is positive for $\theta < \pi/2$, zero for $\theta = \pi/2$, where the two vectors are orthogonal, and negative for $\theta > \pi/2$; and

$\cos\theta$ is the inner product divided by the two norms, so it grades alignment on $[-1, 1]$ without regard to length. The angle drawn in Fig. 1 is smaller than $\pi/2$: the two vectors have a positive inner product, and the orthogonal projection therefore falls on the same side of the origin as $\mathbf{x}$.

Fig. 1 projects one vector onto the direction of another, that is, onto the line it spans. The same characterization holds for a whole subspace, and it is what turns approximation problems, with respect to the induced norm, into systems of linear equations. Consider a subspace $\mathcal{U} \subseteq \mathcal{V}$ and a vector $\mathbf{v} \in \mathcal{V}$. In general, the vector of $\mathcal{U}$ closest to $\mathbf{v}$ with respect to some norm is called the projection of $\mathbf{v}$ onto $\mathcal{U}$. If the norm is induced by some inner product, the projection becomes an orthogonal projection: a vector $\widehat{\mathbf{v}} \in \mathcal{U}$ minimizes $\|\mathbf{v} - \mathbf{u}\|$ over all $\mathbf{u} \in \mathcal{U}$ if and only if the approximation error $\mathbf{v} - \widehat{\mathbf{v}}$ is orthogonal to every vector in $\mathcal{U}$,

$$\langle \mathbf{v} - \widehat{\mathbf{v}}, \mathbf{u} \rangle = 0 \quad \text{for all } \mathbf{u} \in \mathcal{U}. \tag{1}$$

The reason is the Pythagorean identity

$$\|\mathbf{v} - \mathbf{u}\|^2 = \|\mathbf{v} - \widehat{\mathbf{v}}\|^2 + \|\widehat{\mathbf{v}} - \mathbf{u}\|^2,$$

which holds for every $\mathbf{u} \in \mathcal{U}$ whenever (1) does. The second term is non-negative, so no vector of $\mathcal{U}$ is closer to $\mathbf{v}$ than $\widehat{\mathbf{v}}$ is: orthogonality of the error and being closest are the same condition (Axler, 2024, Thm. 6.61). The orthogonality condition (1) is linear in $\widehat{\mathbf{v}}$, so a minimizer can be computed by solving linear equations.

The characterization extends beyond subspaces. For a non-empty closed convex set $\mathcal{C} \subseteq \mathbb{R}^d$, a vector $\widehat{\mathbf{v}} \in \mathcal{C}$ is the projection of $\mathbf{v}$ onto $\mathcal{C}$ if and only if

$$\langle \mathbf{v} - \widehat{\mathbf{v}}, \mathbf{u} - \widehat{\mathbf{v}} \rangle \leq 0 \quad \text{for all } \mathbf{u} \in \mathcal{C}. \tag{2}$$

Such a projection exists and is unique (Bertsekas, 2009, Prop. 1.1.9).

One application of the subspace case (1) is a characterization of linear regression: the learned model parameters are those for which the prediction error is orthogonal to every column of the feature matrix (see normal equations), so the resulting predictions are the orthogonal projection of the label vector onto the column space of the feature matrix. The convex case (2) becomes relevant when linear regression is modified by adding a convex constraint on the model parameters.

The Euclidean space $\mathbb{R}^d$ carries other inner products besides the standard one (Axler, 2024, Example 6.3). As a case in point, each symmetric and positive definite (pd) matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ induces an inner product $\langle \mathbf{x}, \mathbf{x}' \rangle := \mathbf{x}^\top \mathbf{A} \mathbf{x}'$. Such a matrix weighs the coordinate directions against each other, and thereby fixes a different notion of similarity between data points. Take $\mathbf{A} = \operatorname{diag}(4, 1/4)$, which counts the first feature sixteen times as heavily as the second, and the query $\mathbf{x} = (1, 1)^\top$. Of the two candidates $(0.9, 1.2)^\top$ and $(1.3, 0.4)^\top$, the standard inner product prefers the first (2.1 against 1.7) and the weighted one prefers the second (5.3 against 3.9): the same query and the same candidates, but a different answer to which is more similar. Since an inner product also induces

a norm and a metric, the choice of $\mathbf{A}$ fixes which data points count as nearest neighbors as well.

An inner product also gives rise to the notion of an orthonormal basis: a basis $\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(d)}$ of a vector space $\mathcal{V}$ of dimension d is orthonormal if its elements are pairwise orthogonal and have unit norm, $\langle \mathbf{b}^{(j)}, \mathbf{b}^{(j')} \rangle = 0$ for $j \neq j'$ and $\langle \mathbf{b}^{(j)}, \mathbf{b}^{(j)} \rangle = 1$ (Axler, 2024, Def. 6.22). The coordinates of a vector with respect to an orthonormal basis are delivered by inner products: $\mathbf{u} = \sum_{j=1}^{d} \beta_j \mathbf{b}^{(j)}$ with expansion coefficients $\beta_j = \langle \mathbf{u}, \mathbf{b}^{(j)} \rangle$. Conversely, an inner product can be defined by declaring a basis orthonormal: given any basis $\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(d)}$ of $\mathcal{V}$, setting

$$\langle \mathbf{u}, \mathbf{v} \rangle := \sum_{j=1}^{d} \beta_j \beta'_j \quad \text{for } \mathbf{u} = \sum_{j=1}^{d} \beta_j \mathbf{b}^{(j)} \text{ and } \mathbf{v} = \sum_{j=1}^{d} \beta'_j \mathbf{b}^{(j)}$$

yields the unique inner product for which this basis is orthonormal. The standard inner product of the Euclidean space arises in this way from the standard basis given by the columns of the identity matrix $\mathbf{I}_d$.

Fig. 2 illustrates inner products between feature vectors in $\mathbb{R}^2$ that represent five European cities. Each city is characterized by its latitude and longitude as its two features.

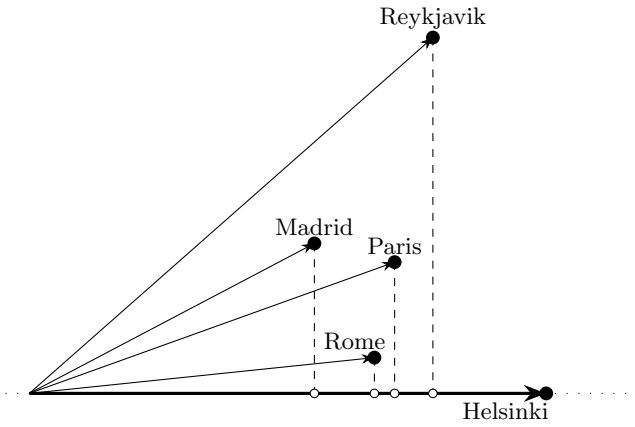

Figure 2: Five European cities, each represented by a feature vector $\mathbf{x} = (x_1, x_2)^\top \in \mathbb{R}^2$ with latitude x_1 and longitude x_2 as features. Each city is drawn at the vector $x_1\mathbf{u} + x_2\mathbf{v}$ (filled circles) for a fixed orthonormal pair $\mathbf{u}, \mathbf{v} \in \mathbb{R}^2$, chosen such that the vector of Helsinki (thick arrow) lies on the horizontal line and the other cities lie above it. The drawn arrows are therefore not the feature vectors $(x_1, x_2)^\top$ themselves. However, since $\mathbf{u}, \mathbf{v}$ are orthonormal, the inner product of two vectors equals the standard inner product of the corresponding feature vectors, $\langle x_1\mathbf{u} + x_2\mathbf{v}, x_1'\mathbf{u} + x_2'\mathbf{v}\rangle = x_1x_1' + x_2x_2'$. Each dashed ruler projects the vector of a city orthogonally onto the Helsinki vector. The open circles mark the resulting orthogonal projections. City coordinates retrieved from OpenStreetMap

The similarity measure of the opening paragraph is what several ML constructions are built on. An embedding maps a text block to a vector in a vector space. The alignment of two text blocks is measured by an inner product of their corresponding embeddings (Goodfellow et al., 2016, Sect. 15.4). Each neuron of an artificial neural network (ANN) matches its input vector against a vector of learned model parameters (Goodfellow et al., 2016, Ch. 6). Attention scores in an LLM are scaled inner products between query and key vectors (Vaswani et al., 2017).

Many ML methods access the feature vectors of data points only through inner products between feature vectors. A linear model uses hypotheses $h(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x}\rangle$. Consider regularized empirical risk minimization (RERM) over a training set $\mathcal{D}^{(\text{train})} = \big\{\big(\mathbf{x}^{(r)}, y^{(r)}\big)\big\}_{r=1}^{m}$, with a penalty that increases strictly with $\|\mathbf{w}\|$. It returns model parameters that are a linear combination of the feature vectors in the training set, $\widehat{\mathbf{w}} = \sum_{r=1}^{m} \beta_r \mathbf{x}^{(r)}$. This is the representer theorem, and the support-vector expansion of the support vector machine (SVM) is a special case of it (Schölkopf and Smola, 2002). The prediction for a new data

point then reads

$$\hat{h}(\mathbf{x}) = \sum_{r=1}^{m} \beta_r \langle \mathbf{x}^{(r)}, \mathbf{x} \rangle,$$

so training and prediction touch the feature vectors only through the inner products $\langle \mathbf{x}^{(r)}, \mathbf{x}^{(r')} \rangle$ within the training set and $\langle \mathbf{x}^{(r)}, \mathbf{x} \rangle$ between the training set and the new data point. This is exploited by kernel methods that use a kernel $k(\mathbf{x}, \mathbf{x}')$ to compute inner products without ever forming feature vectors (Schölkopf and Smola, 2002) (see kernel method).

Methods built on Euclidean distances access the feature vectors only through inner products as well, even when they are defined without reference to one. The squared Euclidean distance between two feature vectors is a sum of three inner products, $\|\mathbf{x} - \mathbf{x}'\|_2^2 = \langle \mathbf{x}, \mathbf{x} \rangle - 2\langle \mathbf{x}, \mathbf{x}' \rangle + \langle \mathbf{x}', \mathbf{x}' \rangle$. A method that uses the feature vectors only through their Euclidean distances therefore also uses them only through inner products. k-nearest neighbors (k-NN) is one example: it ranks the data points in the training set by the distance of their feature vectors to the feature vector of the new data point (Hastie et al., 2009, Sect. 2.3.2). The same kernel substitution therefore reaches it: by the identity above, each Euclidean distance is computed from three kernel evaluations, and no feature vector is ever formed.

Cross References

field, norm, vector, Hilbert space, orthogonality condition, metric space, Euclidean space, Cauchy-Schwarz inequality, orthogonal projection, kernel, kernel method, neuron, attention, PCA, k-NN

References

1. Axler (2024). Linear Algebra Done Right. 4th ed., Springer, Cham, Switzerland.

2. Bertsekas (2009). Convex Optimization Theory. Athena Scientific, Belmont, MA, USA.

3. Goodfellow et al. (2016). Deep Learning. MIT Press, Cambridge, MA, USA.

4. Schölkopf and Smola (2002). Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA, USA.

5. Strang (2016). Introduction to Linear Algebra. 5th ed., Wellesley-Cambridge Press, Wellesley, MA, USA.

6. Hastie et al. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed., Springer Science+Business Media, New York, NY, USA.

7. Vaswani et al. (2017). Attention is all you need. In: Adv. Neural Inf. Process. Syst., pp. 5998–6008.