Skalarprodukt

Autor und Herausgeber

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland (ORCID 0000-0001-7538-0990)
Editors: Konstantina Olioumtsevits (ORCID 0000-0003-2697-8659), Ekkehard Schnoor (ORCID 0000-0001-8252-1499), Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

Zusammenfassung

Das Skalarprodukt zweier Vektoren in einem Vektorraum ist eine Zahl, die deren Ähnlichkeit misst. Viele Anwendungen des maschinellen Lernens stellen einen Datenpunkt durch ein numerisches Array von Merkmale dar, wie etwa die Farbintensitäten eines Bildes, die Amplituden eines Sensor-Signals oder die Einbettungen der von großen Sprachmodellen verwendeten Tokens. Jedes solche Array ist ein Vektor in einem Vektorraum. Viele Methoden des maschinellen Lernens beruhen darauf, die Ähnlichkeit zwischen den Vektoren zu messen, die Datenpunkte repräsentieren. Jedes Neuron eines künstlichen neuronalen Netzes gleicht seine Eingabe mit einer gelernten Vorlage ab, und die Attention-Einheit eines großen Sprachmodells vergleicht den Query-Vektor eines Tokens mit dem Key-Vektor eines anderen. Beide Vergleiche berechnen ein Skalarprodukt.

Definition

Die Farbintensitäten der Pixel eines Bildes, die Amplituden eines Sensor-Signals in regelmäßigen Abständen und die Einbettung eines Tokens in einem großen Sprachmodell sind numerische Arrays, und Datenpunkte werden oft durch Merkmale charakterisiert, die ein solches Array bilden. Ein solches Array kann durch einen Vektor, d. h. ein Element eines Vektorraums, dargestellt werden. Viele Methoden des maschinellen Lernens vergleichen zwei Datenpunkte, indem sie messen, wie ähnlich sich die Vektoren sind, die sie repräsentieren. Ein solches Ähnlichkeitsmaß ist ein Skalarprodukt.

Betrachtet sei ein Vektorraum $\mathcal{V}$ über dem Körper der reellen Zahlen $\mathbb{R}$. Ein Skalarprodukt in $\mathcal{V}$ ist eine Funktion

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

die für alle Vektoren $\mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathcal{V}$ und alle Skalare $\beta \in \mathbb{R}$ die folgenden Eigenschaften erfüllt (Axler, 2024, Def. 6.2):

- Symmetrie: $\langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathbf{v}, \mathbf{u} \rangle$;
- Linearität im ersten Argument: $\langle \beta \mathbf{u} + \mathbf{w}, \mathbf{v} \rangle = \beta \langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{w}, \mathbf{v} \rangle$;
- Positive Definitheit: $\langle \mathbf{u}, \mathbf{u} \rangle \geq 0$, mit Gleichheit genau dann, wenn $\mathbf{u} = \mathbf{0}$.

Über dem Körper der komplexen Zahlen wird die Symmetrie zur hermiteschen Symmetrie, $\langle \mathbf{u}, \mathbf{v} \rangle = \overline{\langle \mathbf{v}, \mathbf{u} \rangle}$, und das Skalarprodukt ist semilinear im zweiten Argument (Axler, 2024, Thm. 6.6); im Folgenden werden nur reelle Vektorräume betrachtet. Zwei Vektoren $\mathbf{u}, \mathbf{v} \in \mathcal{V}$ heißen orthogonal, wenn ihr Skalarprodukt null ist, $\langle \mathbf{u}, \mathbf{v} \rangle = 0$.

Das Paar $(\mathcal{V}, \langle \cdot, \cdot \rangle)$ wird als Innenproduktraum (oder Prä-Hilbertraum) bezeichnet. Jedes Skalarprodukt induziert eine Norm über $\|\mathbf{u}\| := \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle}$ für alle $\mathbf{u} \in \mathcal{V}$ (Axler, 2024, Thm. 6.17), die wiederum eine Metrik über $d(\mathbf{u}, \mathbf{v}) := \|\mathbf{u} - \mathbf{v}\|$ für alle $\mathbf{u}, \mathbf{v} \in \mathcal{V}$ induziert.

Für den euklidischen Raum $\mathcal{V} = \mathbb{R}^d$ ist das Standardskalarprodukt, auch Punktprodukt genannt, gegeben durch $\langle \mathbf{x}, \mathbf{x}' \rangle = \mathbf{x}^\top \mathbf{x}' = \sum_{j=1}^{d} x_j x'_j$ (Strang, 2016, Sect. 1.2). Geometrisch bestimmt das Skalarprodukt orthogonale Projektionen: Die orthogonale Projektion von $\mathbf{x}'$ auf die Richtung eines von null verschiedenen Vektors $\mathbf{x}$ hat die vorzeichenbehaftete Länge $\langle \mathbf{x}, \mathbf{x}' \rangle / \|\mathbf{x}\|$ (Axler, 2024, Example 6.56) (siehe Abb. 1).

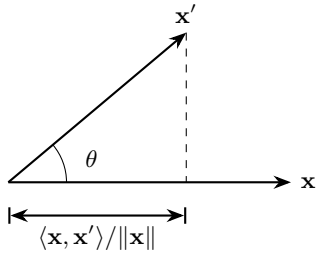

Abbildung 1: Geometrie des Standardskalarprodukts im euklidischen Raum $\mathbb{R}^2$. Das gestrichelte Lot fällt von $\mathbf{x}'$ senkrecht auf die Richtung von $\mathbf{x}$ und trifft diese Richtung an der orthogonalen Projektion von $\mathbf{x}'$. Diese orthogonale Projektion liegt in der vorzeichenbehafteten Länge $\langle \mathbf{x}, \mathbf{x}' \rangle / \|\mathbf{x}\|$ vom Ursprung entfernt, markiert durch den Doppelpfeil. Der Winkel θ zwischen den beiden Vektoren erfüllt $\cos\theta = \langle \mathbf{x}, \mathbf{x}' \rangle / \big(\|\mathbf{x}\| \|\mathbf{x}'\|\big)$ (unten definiert). Das Skalarprodukt misst die Ausrichtung der beiden Vektoren zueinander und verschwindet, wenn sie orthogonal sind

Das Skalarprodukt definiert auch den Winkel θ zwischen zwei von null verschiedenen Vektoren $\mathbf{x}, \mathbf{x}' \in \mathcal{V}$ über $\cos\theta := \langle \mathbf{x}, \mathbf{x}' \rangle / \big(\|\mathbf{x}\| \|\mathbf{x}'\|\big)$. Die Cauchy-Schwarz-Ungleichung $|\langle \mathbf{x}, \mathbf{x}' \rangle| \leq \|\mathbf{x}\| \|\mathbf{x}'\|$ stellt sicher, dass dieses Verhältnis in $[-1, 1]$ liegt (Axler, 2024, Thm. 6.14). Hierbei ist cos die Kosinus-Funktion, die streng monoton von $\cos 0 = 1$ bis $\cos \pi = -1$ fällt, da ihre Ableitung $-\sin$ auf $(0, \pi)$ negativ ist, und daher jeden Wert aus $[-1, 1]$ auf dem Intervall $[0, \pi]$ genau einmal annimmt. Für die euklidische Ebene $\mathbb{R}^2$, wie in Abb. 1 dargestellt, stimmt der auf diese Weise definierte Winkel mit dem elementargeometrischen Winkel zwischen zwei Vektoren überein (Strang, 2016, Sect. 1.2).

Für zwei Vektoren mit gegebenen Normen werden Vorzeichen und Betrag ihres Skalarprodukts durch den Winkel θ zwischen ihnen bestimmt. Das Skalar-

produkt ist positiv für $\theta < \pi/2$, null für $\theta = \pi/2$, wobei die beiden Vektoren orthogonal sind, und negativ für $\theta > \pi/2$. Der in Abb. 1 eingezeichnete Winkel ist kleiner als $\pi/2$: Die beiden Vektoren haben ein positives Skalarprodukt, und die orthogonale Projektion fällt daher auf dieselbe Seite des Ursprungs wie $\mathbf{x}$.

Abb. 1 projiziert einen Vektor auf die Richtung eines anderen, d. h. auf die von ihm aufgespannte Gerade. Dieselbe Charakterisierung gilt für einen ganzen Unterraum, und sie ist es, die Approximationsprobleme bezüglich der induzierten Norm in Systeme linearer Gleichungen überführt. Betrachtet seien ein Unterraum $\mathcal{U} \subseteq \mathcal{V}$ und ein Vektor $\mathbf{v} \in \mathcal{V}$. Im Allgemeinen wird der Vektor aus $\mathcal{U}$, der $\mathbf{v}$ bezüglich einer Norm am nächsten liegt, als die Projektion von $\mathbf{v}$ auf $\mathcal{U}$ bezeichnet. Wird die Norm durch ein Skalarprodukt induziert, so wird die Projektion zu einer orthogonale Projektion. Ein Vektor $\widehat{\mathbf{v}} \in \mathcal{U}$ minimiert $\|\mathbf{v} - \mathbf{u}\|$ über alle $\mathbf{u} \in \mathcal{U}$ genau dann, wenn der Approximationsfehler $\mathbf{v} - \widehat{\mathbf{v}}$ orthogonal zu jedem Vektor in $\mathcal{U}$ ist,

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

Der Grund dafür ist der Satz des Pythagoras

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

der für jedes $\mathbf{u} \in \mathcal{U}$ gilt, wann immer (1) erfüllt ist. Der zweite Term ist nichtnegativ, sodass kein Vektor aus $\mathcal{U}$ näher an $\mathbf{v}$ liegt als $\widehat{\mathbf{v}}$: Orthogonalität des Fehlers und minimale Distanz sind dieselbe Bedingung (Axler, 2024, Thm. 6.61). Die Orthogonalitätsbedingung (1) ist linear in $\widehat{\mathbf{v}}$, sodass ein Minimierer durch das Lösen linearer Gleichungen berechnet werden kann.

Abb. 2 veranschaulicht Skalarprodukte zwischen Merkmalsvektoren in $\mathbb{R}^2$, die fünf europäische Städte repräsentieren. Jede Stadt ist durch ihren Breiten- und Längengrad als ihre beiden Merkmale charakterisiert und am Vektor $x_1\mathbf{u} + x_2\mathbf{v}$ für ein festes orthonormales Paar $\mathbf{u}, \mathbf{v} \in \mathbb{R}^2$ eingezeichnet. Die orthogonalen Projektionen der anderen Städte auf die Richtung von Helsinki ordnen diese nach ihrer Ausrichtung zu Helsinki.

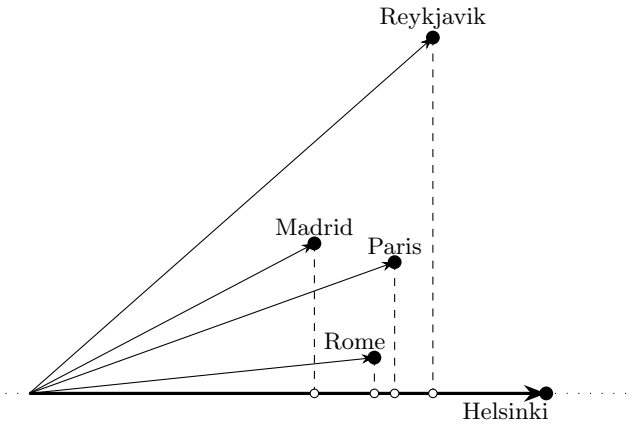

Abbildung 2: Fünf europäische Städte, jeweils dargestellt durch einen Merkmalsvektor $\mathbf{x} = (x_1, x_2)^\top \in \mathbb{R}^2$ mit Breitengrad x_1 und Längengrad x_2 als Merkmale. Jede Stadt ist am Vektor $x_1\mathbf{u} + x_2\mathbf{v}$ (ausgefüllte Kreise) für ein festes orthonormales Paar $\mathbf{u}, \mathbf{v} \in \mathbb{R}^2$ eingezeichnet, das so gewählt ist, dass der Vektor von Helsinki (dicker Pfeil) auf der horizontalen Linie liegt und die anderen Städte darüber liegen. Die eingezeichneten Pfeile sind daher nicht die Merkmalsvektoren $(x_1, x_2)^\top$ selbst. Da jedoch $\mathbf{u}, \mathbf{v}$ orthonormal sind, stimmt das Skalarprodukt zweier Vektoren mit dem Standardskalarprodukt der zugehörigen Merkmalsvektoren überein, $\langle x_1\mathbf{u}+x_2\mathbf{v}, x_1'\mathbf{u}+x_2'\mathbf{v}\rangle = x_1x_1' + x_2x_2'$. Jedes gestrichelte Lot projiziert den Vektor einer Stadt orthogonal auf den Helsinki-Vektor. Die offenen Kreise markieren die resultierenden orthogonale Projektionen. Stadtkoordinaten abgerufen von OpenStreetMap

Die Charakterisierung lässt sich über Unterräume hinaus verallgemeinern. Für eine nichtleere abgeschlossene konvexe Menge $\mathcal{C} \subseteq \mathbb{R}^d$ ist ein Vektor $\widehat{\mathbf{v}} \in \mathcal{C}$ genau dann die Projektion von $\mathbf{v}$ auf $\mathcal{C}$, wenn

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

Eine solche Projektion existiert und ist eindeutig (Bertsekas, 2009, Prop. 1.1.9).

Eine Anwendung des Unterraum-Falls (1) ist eine Charakterisierung der linearen Regression. Die gelernten Modellparameter sind diejenigen, für die der Vorhersagefehler orthogonal zu jeder Spalte der Merkmalsmatrix ist (siehe Normalgleichungen). Die resultierenden Vorhersagen sind daher die orthogonale Projektion des Labelvektors auf den Spaltenraum der Merkmalsmatrix (Hastie et al., 2009, Sect. 3.2). Der konvexe Fall (2) wird relevant, wenn die lineare Regression durch Hinzufügen einer konvexen Nebenbedingung an die Modellparameter modifiziert wird.

Der euklidische Raum $\mathbb{R}^d$ trägt neben dem Standardskalarprodukt noch weitere Skalarprodukte (Axler, 2024, Beispiel 6.3). Beispielsweise induziert jede symmetrische und positiv definite Matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$, d. h. eine mit $\mathbf{x}^\top\mathbf{A}\mathbf{x} > 0$

für jeden von null verschiedenen $\mathbf{x}$, ein Skalarprodukt $\langle \mathbf{x}, \mathbf{x}' \rangle := \mathbf{x}^\top \mathbf{A} \mathbf{x}'$. Eine solche Matrix gewichtet die Koordinatenrichtungen gegeneinander und legt dadurch einen anderen Ähnlichkeitsbegriff zwischen Datenpunkte fest. Betrachtet sei $\mathbf{A} = \mathrm{diag}(4, 1/4)$, wodurch das erste Merkmal sechzehnmal so stark wie das zweite gewichtet wird, sowie die Anfrage $\mathbf{x} = (1, 1)^\top$. Von den beiden Kandidaten $(0.9, 1.2)^\top$ und $(1.3, 0.4)^\top$ ist das Standardskalarprodukt für den ersten größer (2.1 gegenüber 1.7) und das gewichtete für den zweiten (5.3 gegenüber 3.9): dieselbe Anfrage und dieselben Kandidaten, aber eine andere Antwort darauf, welcher ähnlicher ist. Da ein Skalarprodukt auch eine Norm und eine Metrik induziert, legt die Wahl von $\mathbf{A}$ ebenso fest, welche Datenpunkte als nächste Nachbarn gelten.

Ein Skalarprodukt führt außerdem zum Begriff einer orthonormalen Basis. Betrachtet sei eine Basis $\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(d)}$ eines Vektorraums $\mathcal{V}$ der Dimension d. Sie ist orthonormal, wenn ihre Elemente paarweise orthogonal sind und die Norm eins haben, $\langle \mathbf{b}^{(j)}, \mathbf{b}^{(j')} \rangle = 0$ für $j \neq j'$ und $\langle \mathbf{b}^{(j)}, \mathbf{b}^{(j)} \rangle = 1$ (Axler, 2024, Def. 6.22). Die Koordinaten eines Vektors bezüglich einer orthonormalen Basis werden durch Skalarprodukte geliefert: $\mathbf{u} = \sum_{j=1}^{d} \beta_j \mathbf{b}^{(j)}$ mit Entwicklungskoeffizienten $\beta_j = \langle \mathbf{u}, \mathbf{b}^{(j)} \rangle$ (Axler, 2024, Satz 6.30). Umgekehrt lässt sich ein Skalarprodukt definieren, indem eine Basis als orthonormal erklärt wird. Sei $\mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(d)}$ eine beliebige Basis von $\mathcal{V}$. Dann liefert die Festlegung

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

das eindeutige Skalarprodukt, für das diese Basis orthonormal ist. Das Standardskalarprodukt des euklidischen Raums entsteht auf diese Weise aus der durch die Spalten der Einheits-Matrix $\mathbf{I}_d$ gegebenen Standard-Basis.

Mehrere Konstruktionen des maschinellen Lernens nutzen das Ähnlichkeitsmaß aus dem einleitenden Absatz. Eine Einbettung bildet einen Textblock auf einen Vektor in einem Vektorraum ab. Die Übereinstimmung zweier Textblöcke wird durch ein Skalarprodukt ihrer zugehörigen Einbettungen gemessen (Goodfellow et al., 2016, Abschn. 12.4.2). Jedes Neuron eines künstlichen neuronalen Netzes gleicht seinen Eingabevektor mit einem Vektor gelernter Modellparameter ab (Goodfellow et al., 2016, Kap. 6). Attention-Werte in einem großen Sprachmodell sind skalierte Skalarprodukte zwischen Anfrage- und Schlüssel-Vektoren (Vaswani et al., 2017).

Viele Methoden des maschinellen Lernens greifen auf die Merkmalsvektoren von Datenpunkte nur über Skalarprodukte zwischen Merkmalsvektoren zu. Ein lineares Modell verwendet Hypothesen $h(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle$. Betrachtet sei die regularisierte empirische Risikominimierung über einer Trainingsdatensatz $\mathcal{D}^{(\mathrm{train})} = \left\{ \left(\mathbf{x}^{(r)}, y^{(r)} \right) \right\}_{r=1}^{m}$, mit einem Verlust, der von $\mathbf{w}$ nur über die Vorhersagen $\langle \mathbf{w}, \mathbf{x}^{(r)} \rangle$ abhängt, und einem Strafterm, der streng mit $\|\mathbf{w}\|$ wächst. Jeder Minimierer ist eine Linearkombination der Merkmalsvektoren in der Trainingsdatensatz, $\widehat{\mathbf{w}} = \sum_{r=1}^{m} \beta_r \mathbf{x}^{(r)}$. Diese Entwicklung ist der Representer-Satz (Schölkopf and Smola, 2002, Satz 4.2), und die Stütz-Vektor-Entwicklung der Stützvektormaschine (SVM) ist ein Spezialfall davon (Lampert, 2009, Satz 2.3).

Die Vorhersage für einen neuen Datenpunkt lautet dann

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

Training und Vorhersage nutzen die Merkmalsvektoren daher ausschließlich über Skalarprodukte. Innerhalb der Trainingsdatensatz sind dies $\langle \mathbf{x}^{(r)}, \mathbf{x}^{(r')} \rangle$, und zwischen der Trainingsdatensatz und dem neuen Datenpunkt sind es $\langle \mathbf{x}^{(r)}, \mathbf{x} \rangle$. Kernel-Methoden verwenden einen Kernel $k(\mathbf{x}, \mathbf{x}')$, um diese Skalarprodukte zu berechnen, ohne jemals Merkmalsvektoren zu bilden (Schölkopf and Smola, 2002).

Methoden, die auf euklidischen Abständen beruhen, greifen ebenfalls nur über Skalarprodukte auf die Merkmalsvektoren zu, selbst wenn sie ohne Bezugnahme auf ein solches definiert sind. Der quadrierte euklidische Abstand zwischen zwei Merkmalsvektoren ist eine Summe aus drei Skalarprodukten, $\|\mathbf{x} - \mathbf{x}'\|_2^2 = \langle \mathbf{x}, \mathbf{x} \rangle - 2\langle \mathbf{x}, \mathbf{x}' \rangle + \langle \mathbf{x}', \mathbf{x}' \rangle$. Eine Methode, die die Merkmalsvektoren nur über ihre euklidischen Abstände verwendet, nutzt sie daher ebenfalls nur über Skalarprodukte. K-nächste Nachbarn (k-NN) ist ein Beispiel dafür: Es ordnet die Datenpunkte in der Trainingsdatensatz nach dem Abstand ihrer Merkmalsvektoren zum Merkmalsvektor des neuen Datenpunkts (Hastie et al., 2009, Abschn. 2.3.2). Dieselbe Kernel-Ersetzung ist daher auch hier anwendbar: Nach der obigen Identität wird jeder euklidische Abstand aus drei Kernel-Auswertungen berechnet, und es wird niemals ein Merkmalsvektor gebildet.

Synonyme: inneres Produkt, Punktprodukt.

Siehe auch: Norm, Vektor, Hilbertraum, Orthogonalitätsbedingung, euklidischer Raum, Cauchy-Schwarz-Ungleichung, orthogonale Projektion, Kernel.

Querverweise

Literatur

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. Lampert (2009). Kernel Methods in Computer Vision. Found. Trends Comput. Graph. Vis., 4(3), pp. 193–285.

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

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

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

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