Merkmalsraum

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

Der Merkmalsraum einer machine learning (ML)-Anwendung ist die Menge aller Werte, die der Merkmalsvektor eines Datenpunkt annehmen kann. Er ist der Wertebereich der Merkmalstransformation, die einen Datenpunkt auf seinen Merkmalsvektor abbildet. Für Datenpunkte mit d numerischen Merkmale ist der Merkmalsraum üblicherweise der Euclidean space $\mathbb{R}^d$, dessen euklidischer Abstand und Skalarprodukt von Nächste-Nachbar-Methoden und lineare Modelle verwendet werden. Dieselben Datenpunkte, die in zwei Merkmalsräumen platziert sind, haben unterschiedliche nächste Nachbarn und ergeben unterschiedliche Hypothesen, daher ist die Wahl des Merkmalsraums Teil des Modell, und das Weglassen eines Merkmal ist model pruning. Ein Merkmalsraum kann auch eine endliche Menge sein, wie die Monate eines Jahres, oder eine Menge von Zeichenketten oder Graphen, die ein Kernel zugänglich macht. Ein Kernel bringt einen zweiten Merkmalsraum mit sich, einen Hilbertraum von möglicherweise unendlicher Dimension, in dem der Kernel das Skalarprodukt ist und den eine Kernel-Methode verwendet, ohne jemals seine Elemente zu bilden.

Definition

Ein Tag in Krems an der Donau (Österreich) kann durch eine einzige Zahl beschrieben werden, seine Lufttemperatur um 06:00 (UTC), oder durch zwei Zahlen, diesen Wert und den 15:00 (UTC)-Wert des Vortages (GeoSphere Austria, 2026). Die erste Beschreibung verwandelt jeden Tag in einen Punkt der reellen Achse $\mathbb{R}$, die zweite in einen Punkt der Ebene $\mathbb{R}^2$. Die Menge aller Werte, die der Merkmalsvektor eines Datenpunkt annehmen kann, ist der Merkmalsraum $\mathcal{X}$ einer machine learning (ML)-Anwendung. Eine Merkmalstransformation

$$\mathbf{\Phi} : \mathcal{Z} \to \mathcal{X}, \quad \mathbf{x} = \mathbf{\Phi}(\mathbf{z}), \tag{1}$$

bildet einen Datenpunkt $\mathbf{z}$ aus der Menge $\mathcal{Z}$ aller Datenpunkte der Anwendung auf seinen Merkmalsvektor ab, und der Merkmalsraum ist der Wertebereich dieser Abbildung (Shalev-Shwartz and Ben-David, 2014, Sect. 16.1). Die beiden Beschreibungen eines Tages sind zwei Merkmalstransformationen (1) derselben Datenpunkte, mit den Merkmalsräumen $\mathbb{R}$ und $\mathbb{R}^2$.

Für Datenpunkte mit d numerischen Merkmale ist der Standard-Merkmalsraum der Euclidean space $\mathbb{R}^d$ (Axler, 2024, Sect. 1A). Das Wort *Raum*, anstatt *Menge*, besagt, dass $\mathcal{X}$ eine Struktur trägt, die eine ML-Methode verwendet: Die

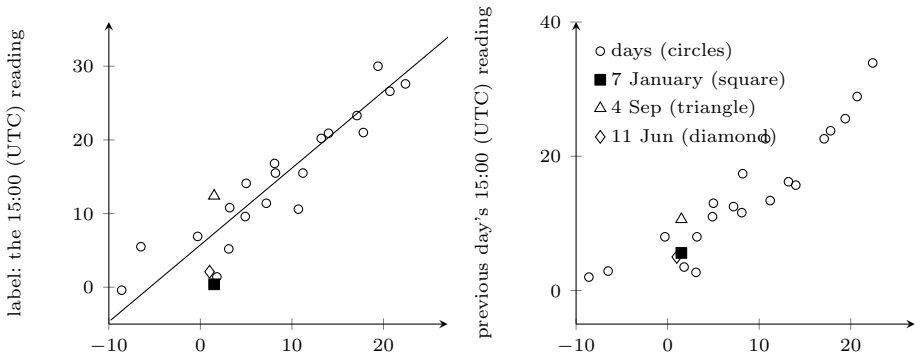
euklidischer Abstand $\|\mathbf{x} - \mathbf{x}'\|_2$ macht $\mathbb{R}^d$ zu einem metric space, und das Skalarprodukt $\mathbf{x}^\top \mathbf{x}'$ macht ihn zu einem Hilbertraum. Die k-nearest neighbors (k-NN)-Methode verwendet die Abstand: der nächste Nachbar eines Datenpunkt mit Merkmalsvektor $\mathbf{x}$ in einem Trainingsdatensatz mit Merkmalsvektoren $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$ ist $\mathbf{x}^{(r^\star)}$ mit

$$r^\star := \operatorname*{arg\,min}_{r=1,\ldots,m} \left\| \mathbf{x} - \mathbf{x}^{(r)} \right\|_2, \tag{2}$$

und das Label dieses Nachbar ist die Vorhersage für $\mathbf{x}$ (Cover and Hart, 1967; Hastie et al., 2009, Sect. 2.3.2). Ein lineares Modell verwendet das Skalarprodukt: seine Hypothesen sind die Abbildungen $h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$ mit Modellparameter $\mathbf{w} \in \mathbb{R}^d$ und $b \in \mathbb{R}$ (Hastie et al., 2009, Sect. 2.3.1).

In welchem Merkmalsraum die Tage des Jahres 2024 in Krems platziert werden, ändert, was beide Methoden liefern. Das Label eines Tages ist sein 15:00-Wert, und ein Tag gelangt nur dann in das Datensatz, wenn auch sein Vorgänger erfasst ist, was 363 Tage übrig lässt. Am 7. Januar 2024 betrug der 06:00-Wert 1.5 Grad Celsius, der 15:00-Wert des Vortages 5.6, und der 15:00-Wert des Tages selbst 0.4. Im Merkmalsraum $\mathbb{R}$ ist der nächste Nachbar (2) dieses Tages unter den anderen 362 Tagen der 27. Februar, mit demselben 06:00-Wert von 1.5, aber einem 15:00-Wert von 12.4. In $\mathbb{R}^2$ ist es der 1. Dezember, mit dem Merkmalsvektor $(1.0, 5.0)^\top$ und einem 15:00-Wert von 2.1 (Abb. 1). Die Vorhersage durch den nächsten Nachbar verfehlt das Label um 12.0 Grad im ersten Merkmalsraum und um 1.7 Grad im zweiten, und 339 der 363 Tage haben einen anderen nächsten Nachbar in $\mathbb{R}$ als in $\mathbb{R}^2$. Ein lineares Modell, das mittels least squares aus allen 363 Tagen gelernt wurde, hat einen Training mean squared error (MSE) von 15.3 in $\mathbb{R}$ und von 10.4 in $\mathbb{R}^2$.

Die Wahl des Merkmalsraums ist daher Teil des Modell. Ein Hypothesenraum ist eine Menge von Abbildungen $h : \mathcal{X} \to \mathcal{Y}$ in den Labelraum $\mathcal{Y}$, sodass eine Änderung von $\mathcal{X}$ ändert, welche Abbildungen er enthält. Die Hypothesen $h(x) = wx + b$ auf $\mathbb{R}$ sind, als Abbildungen des Paares $(x_1, x_2)^\top$ gelesen, die Hypothesen $w_1 x_1 + w_2 x_2 + b$ auf $\mathbb{R}^2$ mit $w_2 = 0$. Das Weglassen eines Merkmal verkleinert somit den Hypothesenraum, was model pruning ist, eine der drei Routen, die die Regularisierung unterscheidet: Der Trainingsfehler von empirische Risikominimierung (ERM) kann nicht fallen (15.3 gegenüber 10.4), während der kleinere Hypothesenraum besser generalisieren kann. Repräsentationslernen-Methoden und feature standardization ändern den Merkmalsraum in die andere Richtung, indem sie die Merkmalstransformation (1) lernen oder neu skalieren. Das Neuskalieren eines Merkmal ändert die Abstände in der Nächste-Nachbar-Regel (2) und damit die nächsten Nachbarn.



(a) the feature space $\mathbb{R}$ is the horizontal axis

(b) the feature space $\mathbb{R}^2$ is the plane

Abbildung 1: Dieselben Tage des Jahres 2024 in Krems in zwei Merkmalsräumen: jeder 18. der 363 Tage (Kreise), 7. Januar (gefülltes Quadrat), sein nächster Nachbar in $\mathbb{R}$, 27. Februar (Dreieck), und sein nächster Nachbar in $\mathbb{R}^2$, 1. Dezember (Raute), insgesamt 23 Tage. (a) Der Merkmalsraum $\mathbb{R}$ des 06:00 (UTC)-Wertes, gezeichnet gegen das Label; die Abstand zwischen zwei Tagen ist der horizontale Abstand zwischen ihren Markierungen, und die durchgezogene Linie ist die lineares Modell Hypothese, die mittels least squares gelernt wurde. (b) Der Merkmalsraum $\mathbb{R}^2$ des 06:00-Wertes und des 15:00-Wertes des Vortages; das gefüllte Quadrat und die Raute liegen eng beieinander, während das Dreieck weit von beiden entfernt ist. Daten generiert von `pythondemos/featurespace.py`

Ein Merkmalsraum muss kein Euclidean space sein. Die Beschreibung eines Tages durch seinen Monat ergibt den endlichen Merkmalsraum $\mathcal{X} = \{1, \ldots, 12\}$. Eine Hypothese darauf ist eine Tabelle von 12 Werten, und die Tabelle, die den Training MSE minimiert, enthält die monatlichen Mittelwerte des 15:00-Wertes, mit einem MSE von 18.2, über den 15.3, die der einzelne 06:00-Wert erreicht: Ein größerer Merkmalsraum ist nicht dasselbe wie ein nützlicherer. Die One-Hot-Kodierung (siehe Merkmalsvektor) bildet die 12 Monate auf die 12 Einheits-Vektoren von $\mathbb{R}^{12}$ ab, und auf diesen ist jede Tabelle ein lineares Modell: least squares lernt dieselben 12 Mittelwerte. Unter der diskreten Metrik, die 0 für zwei gleiche und 1 für zwei verschiedene Monate ist, ist jeder der 27 anderen erfassten Januartage ein nächster Nachbar des 7. Januar. Andere Merkmalsräume ohne Arithmetik sind die Zeichenketten über einem Alphabet, für Textdokumente und DNA-Sequenzen, und die Graphen, für Moleküle. Eine Methode greift dann über einen Kernel auf den Merkmalsraum zu, der die Ähnlichkeit zweier seiner Elemente misst (Schölkopf and Smola, 2002, Sect. 13.2; Vishwanathan et al., 2010). Die Stunden eines Tages, platziert auf einem Zyklus-Graph, sind der Merkmalsraum der Kernel-Ridge-Regression.

Ein Kernel bringt einen zweiten Merkmalsraum mit sich. Ein Kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ ist ein Skalarprodukt transformierter Merkmalsvektoren $\phi(\mathbf{x})$ in einem Hilbertraum $\mathcal{H}$, $k\big(\mathbf{x}, \mathbf{x}'\big) = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle_{\mathcal{H}}$ (Bishop, 2006, Gl. (6.1)), und $\mathcal{H}$

wird auch der Merkmalsraum des Kernel genannt. Für den polynomialen Kernel zweiten Grades auf $\mathbb{R}^2$ ist dieser Raum explizit,

$$k(\mathbf{x}, \mathbf{x}') = \left(1 + \mathbf{x}^\top \mathbf{x}'\right)^2 = \phi(\mathbf{x})^\top \phi(\mathbf{x}'), \tag{3}$$

mit der Merkmalstransformation $\phi(\mathbf{x}) = \left(1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, \sqrt{2}x_1x_2, x_2^2\right)^\top \in \mathbb{R}^6$, sodass ein lineares Modell auf den transformierten Merkmalsvektoren eine quadratische Hypothese auf dem ursprünglichen Merkmalsraum ist (Shalev-Shwartz and Ben-David, 2014, Example 16.1). Die 363×363 Matrix der polynomialen Kernel-Werte (3) der Tage in Krems hat den rank 6, die Dimension von $\mathbb{R}^6$. Der Merkmalsraum des Gaußschen Kernel $\exp\left(-\|\mathbf{x} - \mathbf{x}'\|_2^2/(2\sigma^2)\right)$ hat unendliche Dimension (Shalev-Shwartz and Ben-David, 2014, Example 16.2): für $\sigma = 1$ Grad hat die Matrix seiner Werte für die 363 Tage den rank 361, einen für jeden Punkt von $\mathbb{R}^2$, den die Tage einnehmen, da zwei Tagepaare einen Merkmalsvektor teilen; kein Merkmalsraum kleinerer Dimension erzeugt diese Werte. Eine Kernel-Methode bildet niemals die transformierten Merkmalsvektoren; sie verwendet den Merkmalsraum $\mathcal{H}$ nur über Kernel-Werte, was einen Raum unendlicher Dimension nutzbar macht (Bishop, 2006, Sect. 6.2; Shalev-Shwartz and Ben-David, 2014, Sect. 16.2). Abb. 2 zeigt drei weitere Formen eines Merkmalsraums: die reelle Achse, eine beschränkte konvex Teilmenge der Ebene und die Knoten eines undirected graph. Der zweite ist der Merkmalsraum eines Sensors mit einem begrenzten Bereich, und im dritten ist die Abstand zwischen zwei Merkmalsvektoren die Länge des kürzesten Pfades zwischen ihren Knoten.

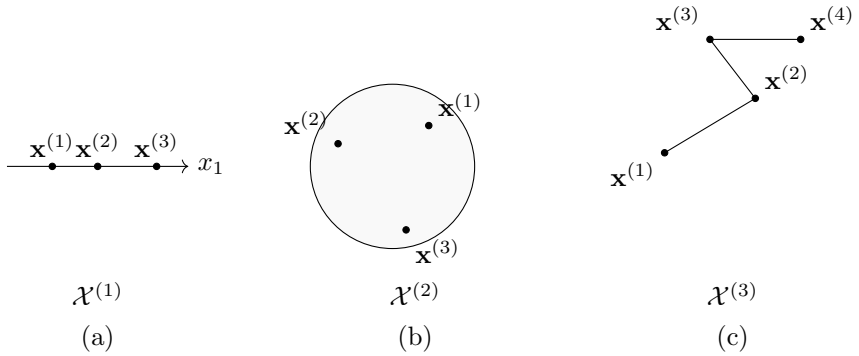


Abbildung 2: Drei Merkmalsräume. (a) Die reelle Achse $\mathcal{X}^{(1)} = \mathbb{R}$. (b) Eine beschränkte konvex Menge $\mathcal{X}^{(2)} \subseteq \mathbb{R}^2$. (c) Ein endlicher Merkmalsraum $\mathcal{X}^{(3)}$, dessen Elemente die Knoten eines undirected graph sind

Synonyme: Eingaberaum, Instanzraum.

Siehe auch: Merkmalsvektor, Merkmal, Merkmalstransformation, Datenpunkt, Labelraum, Euclidean space, euklidischer Abstand, metric space, Hilbertraum, Skalarprodukt, k-NN, lineares Modell, Hypothesenraum, model pruning, Repräsentationslernen, feature standardization, Kernel, Kernel-

Methode, Kernel-Ridge-Regression, reproducing kernel Hilbert space (RKHS), Graph.

Querverweise

Literatur

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

2. Bishop (2006). Pattern Recognition and Machine Learning. Springer Science+Business Media, New York, NY, USA.

3. Cover and Hart (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), pp. 21–27.

4. GeoSphere Austria (2026). GeoSphere Austria climate station archive. Data hub, station historical resource klima-v2-1d. URL: https://dataset.api.hub.geosphere.at/v1/station/hist v2-1d.

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

6. Shalev-Shwartz and Ben-David (2014). Understanding Machine Learning: From Theory to Algorithms. Cambridge Univ. Press, New York, NY, USA.

7. Vishwanathan et al. (2010). Graph Kernels. Journal of Machine Learning Research, 11, pp. 1201–1242.

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