Merkmalsvektor

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

Ein Merkmalsvektor ist eine Liste von Zahlen, die einen Datenpunkt beschreibt, mit je einem Eintrag für jedes seiner Merkmale. Ein Merkmal ist ein Attribut des Datenpunkts, das ohne nennenswerte menschliche Aufsicht gemessen oder berechnet wird. Die genaue Festlegung dieser Messungen definiert eine Merkmalstransformation, die Funktion, die den Datenpunkt auf seinen Merkmalsvektor abbildet, der zu einem zugrunde liegenden Merkmalsraum gehört. Der Merkmalsraum ist üblicherweise ein euklidischer Raum, dessen Dimension die Anzahl der Merkmale ist. Nichtnumerische Merkmale werden als Zahlen kodiert, ein binäres Attribut als 0 oder 1 und ein kategoriales durch One-Hot-Kodierung. Die Definitionsbereich der Merkmalstransformation unterscheidet sich je nach Anwendung: Ein Datenpunkt kann ein Datensatz von Attributen, ein Text, ein digitales Bild, eine Audioaufnahme oder ein Knoten eines Graphen sein, dessen Merkmalsvektor Einträge der Eigenvektoren der Laplace-Matrix zusammenfasst. Allgemeiner kann der Merkmalsraum ein beliebiger Hilbertraum sein, und eine Kernel-Methode verwendet solche Merkmalsvektoren, ohne sie jemals explizit zu bilden, allein über ihre Skalarprodukte.

Definition

Ein Verfahren des maschinelles Lernen (ML)s kann einen Tag in Krems an der Donau (Österreich) durch vier Zahlen darstellen. Dabei handelt es sich um die Lufttemperatur um 06:00 und um 15:00 Uhr (UTC) dieses Tages sowie dieselben beiden Messwerte des Vortages. Für den 16. Juli 2024 sind dies 24.6 und 30.8 °C sowie 22.2 und 31.5 °C am Vortag. Fasst man diese Messwerte zu einem Vektor $\big(24.6, 30.8, 22.2, 31.5\big)^{\top} \in \mathbb{R}^4$ zusammen, so erhält man einen Merkmalsvektor, der diesen Tag charakterisiert.

Allgemein sind die Einträge eines Merkmalsvektors $\mathbf{x} = \big(x_1, \ldots, x_d\big)^{\top}$ die Merkmale $x_1, \ldots, x_d$ eines Datenpunkts: Attribute, die ohne nennenswerte menschliche Aufsicht gemessen oder berechnet werden (Duda et al., 2001). Eine genaue Festlegung, wie diese Attribute gemessen werden, definiert eine Merkmalstransformation

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

Die Funktion $\mathbf{\Phi}$ bildet einen Datenpunkt $\mathbf{z} \in \mathcal{Z}$ auf seinen Merkmalsvektor $\mathbf{x}$ im Merkmalsraum $\mathcal{X}$ ab (Abb. 1). Die Merkmalsextraktion, ein Schritt der Vorverarbeitung, berechnet $\mathbf{\Phi}(\mathbf{z})$ für einen gegebenen Datenpunkt (Bishop, 2006, S. 2). Beim überwachtes Lernen wird der extrahierte Merkmalsvektor in eine Hypothese $h : \mathcal{X} \to \mathcal{Y}$ eingespeist, die eine Vorhersage $\widehat{y} = h(\mathbf{x})$ des Labels $y \in \mathcal{Y}$ des Datenpunkts liefert. Die Merkmalsvektoren $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$ der Datenpunkte in einem Datensatz $\mathcal{D}$ vom dataset size m bilden die Zeilen seiner Merkmalsmatrix $\mathbf{X} \in \mathbb{R}^{m \times d}$.

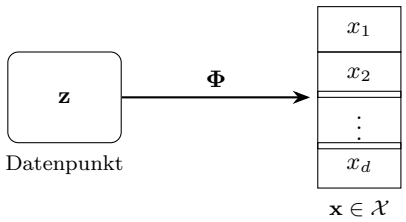


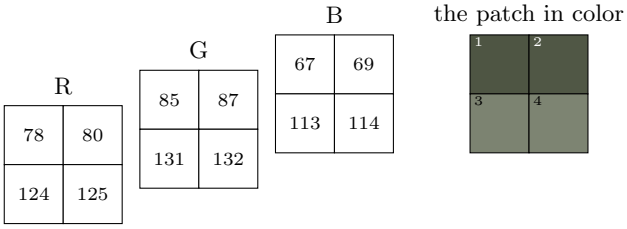
Abbildung 1: Ein Merkmalsvektor $\mathbf{x}$ fasst die einzelnen Merkmale $x_1, \ldots, x_d$ eines Datenpunkts $\mathbf{z}$ zusammen, die durch die Merkmalstransformation $\mathbf{\Phi}$ berechnet werden. Er ist ein Punkt im Merkmalsraum $\mathcal{X}$

Eine weit verbreitete Wahl für den Merkmalsraum ist ein euklidischer Raum $\mathbb{R}^d$ fester Dimension d (Axler, 2024, Abschn. 1A). Für das obige Wetterbeispiel ist der Merkmalsraum $\mathcal{X} = \mathbb{R}^4$. Der Merkmalsraum $\mathbb{R}^d$ dient Anwendungen des MLs, deren Merkmale numerische Größen sind, wie etwa die vier Temperaturen. Auf unterschiedlichen Skalen gemessene Merkmale werden vor dem Training häufig durch Merkmalsstandardisierung auf eine gemeinsame Skala gebracht. Die Dimension d reicht von einer Handvoll Messwerten bis hin zu Millionen von Pixelintensitäten; ein großes d begünstigt den Fluch der Dimensionalität und motiviert eine Dimensionsreduktion.

Auch nichtnumerische Merkmale können Einträge eines Merkmalsvektors sein, sobald ihre Werte als reelle Zahlen kodiert werden (Goodfellow et al., 2016; Hastie et al., 2009). Ein binäres Merkmal, etwa ob es im Laufe des Tages geregnet hat, wird als 0 oder 1 kodiert. Ein kategoriales Merkmal mit K möglichen Werten, wie etwa die Jahreszeit (Winter, Frühling, Sommer oder Herbst), wird One-Hot-kodiert. Es belegt K Einträge des Merkmalsvektors, die alle gleich 0 sind, bis auf einen einzigen Eintrag gleich 1, dessen Position den Wert angibt. Am 16. Juli 2024 fielen 18.7 Millimeter Regen, und die Jahreszeit ist Sommer, sodass die beiden Merkmale als die Einträge 1 bzw. $\big(0, 0, 1, 0\big)$ kodiert werden. Unter die vier Temperaturen gestapelt ergeben sie den Merkmalsvektor $\mathbf{x} = \big(24.6, 30.8, 22.2, 31.5, 1, 0, 0, 1, 0\big)^\top \in \mathbb{R}^9$ des Tages. Datensätze, die numerische und nichtnumerische Attribute auf diese Weise mischen, mit einer Zeile pro Datenpunkt, sind tabellarische Daten.

Während der Wertebereich der Merkmalstransformation $\mathbf{\Phi}$ oft als $\mathbb{R}^d$ gewählt werden kann, variiert ihre Definitionsbereich $\mathcal{Z}$ stark zwischen den Anwendungen des MLs. Beispielsweise kann jedes Element von $\mathcal{Z}$ ein Text-

dokument sein, dessen Merkmalsvektor aus Worthäufigkeiten oder aus einem Durchschnitt der Worteinbettungen seiner Wörter besteht. In der Computervision sind die Elemente von $\mathcal{Z}$ digitale Bilder, jeweils ein Array von Pixeln. Der Merkmalsvektor führt dann die Rot-Grün-Blau-(RGB-)Intensitäten dieser Pixel auf, mit einem Eintrag pro Intensität (Abb. 2). Der Merkmalsvektor einer Audioaufnahme besteht typischerweise aus Werten des Signals zu einer Folge von Zeitpunkten, mit einem Eintrag x_j für jeden Zeitpunkt $j = 1, \ldots, d$ (Abb. 3), oder aus den Einträgen ihres spectrograms. Anstatt $\boldsymbol{\Phi}$ von Hand festzulegen, lernen Verfahren des Deep Learnings sie aus Daten: Die Schichten eines tiefes Netzs bis zur vorletzten berechnen einen Merkmalsvektor, und die letzte Schicht ist die darauf angewandte Hypothese (Repräsentationslernen; Goodfellow et al., 2016, Kap. 15).



$$\mathbf{x} = \big(\underbrace{78, 85, 67}_{\text{pixel 1}},\ \underbrace{80, 87, 69}_{\text{pixel 2}},\ \underbrace{124, 131, 113}_{\text{pixel 3}},\ \underbrace{125, 132, 114}_{\text{pixel 4}}\big)^\top \in \mathbb{R}^{12}$$

Abbildung 2: Ein 2×2-Ausschnitt eines digitalen Bildes und sein Merkmalsvektor. Der Ausschnitt besitzt ein Array (oder Gitter) von Intensitäten pro Farbkanal, bezeichnet mit R, G und B, und jedes Array enthält dieselben vier Pixel an denselben Positionen. Rechts ist der Ausschnitt in Farbe dargestellt, seine Pixel sind von 1 bis 4 nummeriert: Pixel 1 hat die Farbe mit den drei Intensitäten 78, 85 und 67. Stapelt man diese drei Werte für Pixel 1, dann die für Pixel 2 und so weiter, erhält man einen Merkmalsvektor mit $d = 12$ Einträgen

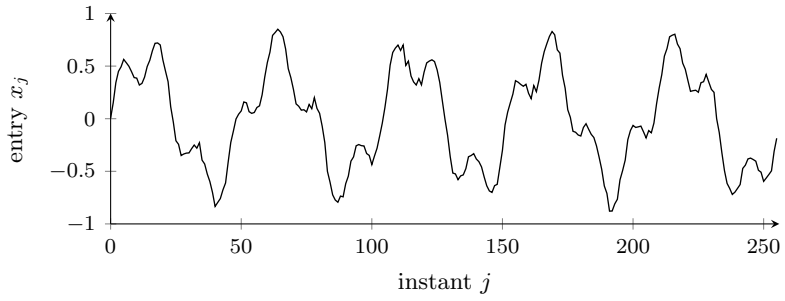


Abbildung 3: Der Merkmalsvektor einer zu 256 Zeitpunkten abgetasteten Audioaufnahme. Der Eintrag x_j ist der Wert, den ihr Signal zum Zeitpunkt j annimmt, sodass dieser Datenpunkt ein Punkt in $\mathbb{R}^{256}$ ist

Die Definitionsbereich $\mathcal{Z}$ kann auch die Knotenmenge $\mathcal{V}$ eines gewichteter Graphen $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ sein, eines Ähnlichkeitsgraphen, dessen Knoten $r \in \mathcal{V} := \{1, \ldots, m\}$ der r-te Datenpunkt ist. Eine Kante $\{r, r'\} \in \mathcal{E}$ mit Kantengewicht $A_{r,r'} > 0$ hält fest, dass die beiden Datenpunkte in einem domänenspezifischen Sinn ähnlich sind, wie etwa zwei miteinander verlinkte Webseiten oder zwei wenige Kilometer voneinander entfernte Wetterstationen. Im Gegensatz zu einem k-Nächste-Nachbarn-Graphen, der aus bereits vorliegenden Merkmale aufgebaut wird, setzt ein solcher Graph keine Merkmale voraus; er liefert sie. Seien $\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(d)}$ Eigenvektoren der Laplace-Matrix $\mathbf{L}^{(\mathcal{G})}$, jeweils ein Vektor in $\mathbb{R}^m$, zu ihren d kleinsten Eigenwerte $\lambda_1 \leq \ldots \leq \lambda_d$. Der Merkmalsvektor von Knoten r fasst die r-ten Einträge dieser Eigenvektoren zusammen,

$$\mathbf{x}^{(r)} = \left(u_r^{(1)}, \ldots, u_r^{(d)}\right)^\top \in \mathbb{R}^d$$

(von Luxburg, 2007, Abschn. 4; Hastie et al., 2009, Abschn. 14.5.3). Ein Eigenvektor zu einem kleinen Eigenwert ändert sich über Kanten mit großem Kantengewicht nur wenig (siehe Laplace-Matrix), sodass durch solche Kanten verbundene Knoten nahe beieinander liegende Merkmalsvektoren erhalten: Die Abbildung $r \mapsto \mathbf{x}^{(r)}$ ist eine Einbettung der Knoten in $\mathbb{R}^d$. Der Eigenvektor $\mathbf{u}^{(2)}$ zur algebraische Konnektivität λ_2 ist der Fiedler-Vektor von $\mathcal{G}$. Das Spektrales Clustering wendet diese Konstruktion an, bevor die resultierenden Merkmalsvektoren geclustert werden (von Luxburg, 2007, Abschn. 4), und ein Graph Neural Network (GNN) lernt Knotenmerkmalsvektoren aus demselben Graphen durch message passing entlang seiner Kanten.

Der Begriff eines euklidischen Merkmalsvektors lässt sich verallgemeinern, indem der Merkmalsraum $\mathbb{R}^d$ durch einen beliebigen Hilbertraum $\mathcal{H}$ ersetzt wird. Hier ist ein Merkmalsvektor das Bild einer Merkmalstransformation

$$\mathbf{\Phi} : \mathcal{Z} \to \mathcal{H},$$

und jeder Hilbertraum kann als Merkmalsraum dienen (Shalev-Shwartz and Ben-David, 2014, Abschn. 16.1). Eine Konstruktion eines solchen Merkmalsvek-

tors verwendet einen Kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. Der Kernel ordnet jedem $\mathbf{z} \in \mathbb{R}^d$ die Funktion $k\big(\mathbf{z}, \cdot\big) : \mathbb{R}^d \to \mathbb{R}$ zu, deren Wert an der Stelle $\mathbf{z}' \in \mathbb{R}^d$ gleich $k\big(\mathbf{z}, \mathbf{z}'\big)$ ist. Diese Funktion ist selbst ein Vektor in einem RKHS (RKHS) $\mathcal{H}$, einem Hilbertraum, dessen Elemente Funktionen auf $\mathbb{R}^d$ sind (Aronszajn, 1950, Teil I, Abschn. 2; Schölkopf and Smola, 2002, Abschn. 2.2.2). Eine Kernel-Methode, wie etwa die Stützvektormaschine (SVM), die Kernel-Ridge-Regression oder ein Gauß-Prozess, verwendet Elemente von $\mathcal{H}$ als Merkmalsvektoren, ohne sie jemals explizit zu bilden. Viele Verfahren des MLs nutzen Merkmalsvektoren nur über ihre Skalarprodukte, welche die reproduzierende Eigenschaft auf Auswertungen des Kernels reduziert, $\langle k\big(\mathbf{z}, \cdot\big), k\big(\mathbf{z}', \cdot\big)\rangle_{\mathcal{H}} = k\big(\mathbf{z}, \mathbf{z}'\big)$ (Schölkopf and Smola, 2002, Gl. (2.30)).

Synonyme: Kovariatenvektor, Eingabevektor, Prädiktorvektor.
Siehe auch: Merkmal, Merkmalsraum, Merkmalstransformation, Datenpunkt, Vektor, euklidischer Raum, Merkmalsmatrix, Eingabevektor, Vorverarbeitung, Repräsentationslernen, Dimensionsreduktion, tabellarische Daten, Graph, gewichteter Graph, Ähnlichkeitsgraph, Laplace-Matrix, Eigenvektor, Einbettung, Spektrales Clustering, Kernel, RKHS, Kernel-Methode, Hilbertraum.

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. Duda et al. (2001). Pattern Classification. 2nd ed., Wiley, New York, NY, USA.
4. Goodfellow et al. (2016). Deep Learning. MIT Press, Cambridge, MA, USA.
5. Schölkopf and Smola (2002). Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA, USA.
6. von Luxburg (2007). A tutorial on spectral clustering. Statist. Comput., 17(4), pp. 395–416.
7. Shalev-Shwartz and Ben-David (2014). Understanding Machine Learning: From Theory to Algorithms. Cambridge Univ. Press, New York, NY, USA.
8. Aronszajn (1950). Theory of Reproducing Kernels. Trans. Am. Math. Soc., 68(3), pp. 337–404.
9. Hastie et al. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed., Springer Science+Business Media, New York, NY, USA.