Merkmal

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 Merkmal eines Datenpunkt ist eines seiner Attribute, das einfach gemessen oder berechnet werden kann, ohne menschliche Aufsicht. Die Merkmale eines Datenpunkt werden in einem Merkmalsvektor gesammelt, den eine Hypothese liest, um das Label des Datenpunkt vorherzusagen. Die Rot-Grün-Blau (RGB)-Pixelintensitäten eines digitalen Bildes oder die Signalwerte einer Audioaufnahme sind typische Merkmale. Welche Attribute eines Datenpunkt als Merkmale und nicht als Labels dienen, ist eine Designentscheidung innerhalb einer machine learning (ML)-Anwendung.

Definition

Ein digitales Foto einer Kuh ist ein Raster von Pixeln, und die Rot-, Grün- und Blauintensität jedes Pixels ist eine Zahl, die ein Computer erhält, ohne dass ein Mensch das Foto betrachtet. Eine solche Zahl ist ein Merkmal des Datenpunkt: ein Attribut des Datenpunkt $\mathbf{z}$, das ohne menschliche Aufsicht gemessen oder berechnet wird (Dodge, 2003; Everitt and Skrondal, 2010; Gujarati and Porter, 2009). Die Merkmale eines Datenpunkt werden durch eine Merkmalstransformation, eine Funktion $\mathbf{\Phi} : \mathcal{Z} \rightarrow \mathcal{X}$, zu einem Merkmalsvektor $\mathbf{x} = \mathbf{\Phi}(\mathbf{z})$ zusammengefügt, wobei der Datenpunkt selbst als seine Rohmerkmale betrachtet wird, aus denen die berechneten Merkmale gewonnen werden. Eine Hypothese nimmt den Merkmalsvektor als Eingabe und gibt eine Vorhersage des Label des Datenpunkt aus.

Neben den Rot-Grün-Blau (RGB)-Intensitäten der Pixel eines Fotos zeigt Abb. 1 ein weiteres Beispiel für Merkmale: die d Signalwerte eines Audio-Signals endlicher Dauer, die als seine Merkmale $x_1, \ldots, x_d$ verwendet werden.

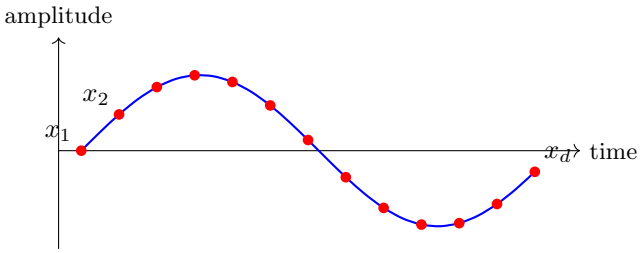


Abbildung 1: Audio-Signal (Kurve) über der Zeit und seine Signalwerte an den Abtastzeitpunkten (Punkte), die als Merkmale $x_1, \ldots, x_d$ verwendet werden können

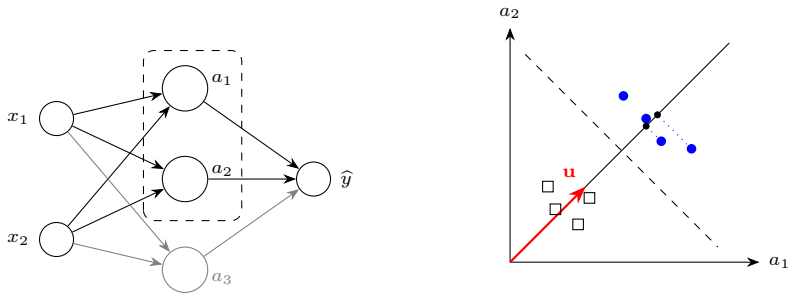
Ob ein gegebenes Attribut als Merkmal oder als Label behandelt wird, ist dem Datenpunkt nicht inhärent. Es ist eine Designentscheidung, die von der machine learning (ML)-Anwendung und den verfügbaren Ressourcen zur Bestimmung jedes Attributs abhängt. Ein Attribut sollte nur dann als Merkmal verwendet werden, wenn sein Wert für jeden Datenpunkt zu dem Zeitpunkt bestimmt werden kann, zu dem die Vorhersage benötigt wird; ein Attribut, das erst danach bekannt wird, ist stattdessen ein Kandidat für das Label.

Sobald die Merkmale ausgewählt sind, können neue Merkmale durch Transformation bestehender Merkmale konstruiert werden. Eine solche Transformation kann eine arithmetische Berechnung sein, die auf die bestehenden Merkmale angewendet wird. Wenn beispielsweise ein Datenpunkt ein numerisches Merkmal $x \in \mathbb{R}$ besitzt, können die Größen x^2, $\exp(-x)$ und $3x$ ebenfalls als Merkmale des Datenpunkt verwendet werden.

Anstelle seiner rohen Signalwerte kann das Audio-Signal in Abb. 1 durch die Beträge seiner diskreten Fourier transform-Transformation (DTFT) beschrieben werden (Oppenheim et al., 1999). Diese Merkmale beschreiben den Frequenzinhalt des Signals und bleiben bei Zeitverschiebungen unverändert, da eine Verschiebung des Signals um n_0 Abtastwerte seine DTFT mit $\exp(-\mathrm{j}\omega n_0)$ multipliziert, einem Faktor vom Betrag eins. Diese Invarianz ist nützlich für ein stationäres Signal, dessen statistische Eigenschaften sich im Laufe der Zeit nicht ändern. Ein weiteres Beispiel ist die Aktivierung eines Neurons innerhalb eines künstliches neuronales Netz (KNN): Es ist ein neues Merkmal, das aus den Eingabemerkmalen durch eine Abfolge grundlegender Berechnungen, die vom KNN durchgeführt werden, abgeleitet wird.

Die Verwendung von Aktivierungen als Merkmale liegt einer zweiten, engeren Verwendung des Wortes *Feature* in der mechanistic interpretability zugrunde: ein Feature ist ein menschlich interpretierbares Konzept, das in einem trainierten KNN kodiert ist, wie z. B. ein Kurvendetektor oder das Vorhandensein eines Rades. Die Aktivierungen einer Untermenge von d Neuronen (zum Beispiel einer ganzen Schicht oder der Neuronen, die ein Analyst zur Untersuchung auswählt) bilden einen Vektor in einem Euclidean space $\mathbb{R}^d$. Während der breitere Sinn oben eine einzelne Neuron-Aktivierung als ein Merkmal behandelt, wird hier ein Feature als Funktion dieses gesamten Aktivierung-Vektors $\mathbf{a}$ erhalten. Eine

solche Funktion ist die Projektion des Aktivierung-Vektors **a** auf einen eindimensionalen Unterraum, der von einem Einheits-Vektor (oder einer Richtung) **u** aufgespannt wird (Olah et al., 2020; Park et al., 2024). Dieser Einheits-Vektor kann als die Einheitsnormale **u** einer Hyperebene erhalten werden, die Datenpunkte trennt, die das Konzept ausdrücken, von denen, die es nicht tun (Kim et al., 2018) (siehe Abb. 2).



(a) (b)

Abbildung 2: Ein Feature als Funktion eines Aktivierung-Vektors **a**, der aus einem trainierten KNN gewonnen wird. (a) Ein KNN mit einer versteckten Schicht von drei Neuronen; die Aktivierungen a_1, a_2 einer ausgewählten Untermenge (gestrichelter Kasten) bilden den Aktivierung-Vektor $\mathbf{a} = \big(a_1, a_2\big)^\top$, während a_3 nicht verwendet wird. (b) Jeder Datenpunkt wird auf einen Punkt $(a_1, a_2)^\top$ im Aktivierung-Raum abgebildet. Datenpunkte, die das Konzept ausdrücken (gefüllte Kreise), und solche, die es nicht tun (offene Quadrate), werden durch eine Hyperebene (gestrichelte Linie) mit der Einheitsnormalen **u** (dicker Pfeil) getrennt. Der Feature-Wert ist die Projektion des Aktivierung-Vektors auf die von **u** aufgespannte Linie (dünne durchgezogene Linie), dargestellt für zwei Punkte durch gepunktete Linien, die in kleinen Punkten enden.

Eine weitere Konstruktion eines Aktivierung-basierten Features basiert auf Regionen anstelle von Richtungen im Aktivierung-Raum (Crabbé and van der Schaar, 2022). In beiden Fällen erfordert die Konstruktion der Funktion, die den Aktivierung-Vektor auf das Feature abbildet, einen Satz von Datenpunkte, die als das Konzept repräsentierend oder nicht repräsentierend gekennzeichnet sind.

Jenseits ihrer Konstruktion schränkt die Regulierung ein, welche Merkmale verwendet werden dürfen und wie genau sie sein müssen. Der EU AI Act verlangt, dass die Training-Daten von Hochrisiko-ML-Systeme (ML-System) relevant, hinreichend repräsentativ und, soweit möglich, fehlerfrei und vollständig sein müssen (European Parliament and Council of the European Union, 2024, Art. 10(3)). Wenn Merkmale personal data sind, fügt die general data protection regulation (GDPR) das data minimization principle (European Parliament and Council of the European Union, 2016, Art. 5(1)(c)) und das Genauigkeitsprinzip

(European Parliament and Council of the European Union, 2016, Art. 5(1)(d)) hinzu.

Umgekehrt kann die bewusste Störung der Merkmale nützlich sein: Das Hinzufügen kleiner Störungen zu den Merkmalen von Datenpunkte während des Training ist eine Form der Datenerweiterung, die als Regularisierung wirkt. Beispielsweise entspricht die Ridge-Regression-Strafe dem durchschnittlichen Verlust unter Gaußschen Merkmalsstörungen: Für ein lineares Modell mit Modellparameter $\mathbf{w}$ erhöht das Hinzufügen von Gaußschem Rauschen mit Nulldurchschnitt und Varianz α zu jedem Merkmal den erwarteten squared error loss um $\alpha\|\mathbf{w}\|_2^2$, die Ridge-Strafe (Bishop, 1995).

Synonyme: Kovariate, erklärende Variable, unabhängige Variable, Eingabe, Prädiktor, Regressor.

Siehe auch: Datenpunkt, Label, Merkmalsvektor, Merkmalsraum, mechanistic interpretability, Konzeptaktivierungsvektor (CAV), Datenerweiterung.

Querverweise

Literatur

1. Bishop (1995). Training with Noise is Equivalent to Tikhonov Regularization. Neural Computation, 7(1), pp. 108–116.

2. Dodge (2003). The Oxford Dictionary of Statistical Terms. Oxford Univ. Press, New York, NY, USA.

3. European Parliament and Council of the European Union (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). Official Journal of the European Union, L series, Jul. 12. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.

4. Everitt and Skrondal (2010). The Cambridge Dictionary of Statistics. 4th ed., Cambridge Univ. Press, Cambridge, U.K.

5. European Parliament and Council of the European Union (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). Official Journal of the European Union, L 119/1, May 4. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

6. Gujarati and Porter (2009). Basic Econometrics. 5th ed., McGraw-Hill/Irwin, New York, NY, USA.

7. Olah et al. (2020). Zoom In: An Introduction to Circuits. Distill, 5(3).

8. Oppenheim et al. (1999). Discrete-Time Signal Processing. 2nd ed., Prentice Hall, Englewood Cliffs, NJ.

9. Crabbé and van der Schaar (2022). Concept Activation Regions: A Generalized Framework for Concept-Based Explanations. In: Adv. Neural Inf. Process. Syst., pp. 2590–2607.

10. Kim et al. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In: Proc. 35th Int. Conf. Mach. Learn., pp. 2668–2677.

11. Park et al. (2024). The Linear Representation Hypothesis and the Geometry of Large Language Models. In: Proc. 41st Int. Conf. Mach. Learn., pp. 39643–39666.