Datenpunkt

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 Datenpunkt ist die elementare informationstragende Einheit, auf der Methoden des maschinellen Lernens arbeiten. Die Eigenschaften eines Datenpunkts fallen in zwei Kategorien: Merkmale, die leicht messbar oder berechenbar sind, und Labels, bei denen es sich um übergeordnete Tatsachen handelt, die typischerweise nur mithilfe menschlichen Fachwissens bestimmt werden können. Ob eine gegebene Eigenschaft als Merkmal oder als Label behandelt wird, ist eine Entwurfsentscheidung, die von der Anwendung des maschinellen Lernens abhängt.

Definition

Ein Bild einer Kuhherde (Abb. 1), ein Patient in einer Klinik und ein Protein sind Objekte, die eine Methode des maschinellen Lernens verarbeitet. Ein Datenpunkt ist ein Objekt, das Information trägt (Cover and Thomas, 2006); solche Objekte sind die elementaren Einheiten, auf denen Methoden des maschinellen Lernens arbeiten. Beispiele sind Studierende, Funk-Signale, Bäume, Bilder, Zufallsvariablen (ZV), reelle Zahlen oder Proteine. Datenpunkte desselben Typs werden durch zwei Kategorien von Eigenschaften beschrieben.

Die erste Kategorie umfasst Merkmale, die messbare oder berechenbare Eigenschaften eines Datenpunkts sind. Sie können automatisch mithilfe von Sensoren, Computern oder anderen Systemen zur Erfassung von Daten extrahiert oder berechnet werden. Die zweite Kategorie umfasst Labels, bei denen es sich um übergeordnete Tatsachen oder Zielgrößen handelt, deren Bestimmung typischerweise menschliches Fachwissen oder Domänenwissen erfordert, anstatt direkt messbar zu sein. Das maschinelle Lernen zielt darauf ab, das Label eines Datenpunkts aus seinen Merkmalen vorherzusagen.

Abb. 1 zeigt ein Bild als Beispiel für einen Datenpunkt. Verschiedene Eigenschaften des Bildes können als Merkmale dienen: die Farbintensitäten $x_1, \ldots, x_d$ aller Bildpixel, der Zeitstempel x_{d+1} der Bildaufnahme und der räumliche Ort x_{d+2} der Bildaufnahme. Übergeordnete Tatsachen können als Labels dienen: die Anzahl der Kühe y_1, die Anzahl der Wölfe y_2 und der Zustand der Weide y_3 (z. B. gesund oder überweidet).



Abbildung 1: Darstellung eines Datenpunkts, der aus einem Bild besteht. Verschiedene Eigenschaften des Bildes können als Merkmale dienen, wie etwa die durchschnittlichen Farbintensitäten der Bildpixel. Übergeordnete Tatsachen über das Bild, wie etwa das Vorhandensein verschiedener Objektkategorien, können als Labels dienen

Für einen Datenpunkt, der einen Patienten repräsentiert, könnte ein Merkmal das Körpergewicht sein, während das Label das Vorliegen einer Krebserkrankung sein könnte, also die Tatsache von Interesse. Die Bestimmung dieses Labels erfordert, dass ein medizinischer Experte (oder sogar ein Gremium von Experten) den Patienten untersucht. Eine Eigenschaft, die in einem Szenario als Label behandelt wird (z. B. eine Krebsdiagnose), kann in einem anderen als Merkmal dienen, in dem eine zuverlässige Automatisierung (z. B. Bildanalyse) es erlaubt, die Eigenschaft ohne menschliches Eingreifen zu berechnen.

Beide Arten von Eigenschaften sind fehleranfällig. Bei vielen Anwendungen des maschinellen Lernens ist es selten möglich, auf die wahren Labels zuzugreifen. Beispielsweise kann die Diagnose eines medizinischen Experten falsch sein, was sie zu einem verrauschten Stellvertreter für den wahren Zustand des Patienten macht. Fehler dieser Art, die als Label-Rauschen bezeichnet werden, entstehen immer dann, wenn ein Label lediglich ein Stellvertreter für die wahre Zielgröße ist, unabhängig davon, ob dieser Stellvertreter durch menschliches Urteil oder durch ein automatisiertes System erzeugt wird.

Merkmale sind ebenfalls eine Fehlerquelle innerhalb einer Methode des maschinellen Lernens. Ein Merkmal wird durch Messen oder Berechnen einer Eigenschaft eines Datenpunkts gewonnen. Physikalische und chemische Messgeräte weisen eine endliche Mess-Unsicherheit auf (Kimothi, 2002), und die Berechnung eines Merkmals auf einem endlichen Computer führt zu Rundungsfehlern (Golub and Van Loan, 2013, Abschn. 2.7). Diese Unzulänglichkeiten führen dazu, dass der aufgezeichnete Wert vom wahren Merkmalswert abweicht. Diese Abweichung ist Merkmalsrauschen.

Methoden des maschinellen Lernens sollten robust gegenüber sowohl Merkmals- als auch Label-Rauschen sein. Ihre Vorhersagen für neue Datenpunkte sollten sich nicht stark verschlechtern, wenn einige Merkmale oder Labels im Trainingsdatensatz fehlerhaft sind. Die KI-Verordnung spiegelt dieses Anliegen wider: Daten für das Training müssen, soweit möglich, „fehlerfrei“ sein

(European Parliament and Council of the European Union, 2024, Art. 10(3)), und Hochrisiko-KI-Systeme müssen ein angemessenes Maß an Robustheit erreichen (European Parliament and Council of the European Union, 2024, Art. 15(1)).

Die Robustheit gegenüber Merkmals- und Label-Rauschen kann durch Datenerweiterung verbessert werden, bei der die Merkmale oder das Label eines Datenpunkts für das Training gezielt gestört werden (siehe Abb. 2). Das Training auf solchen gestörten Datenpunkten wirkt als Regularisierung: Für den quadratischen Fehlerverlust und kleines Merkmalsrauschen ist es äquivalent zum Hinzufügen eines Strafterms auf die Ableitungen der Hypothese nach den Merkmalen (Bishop, 1995, Abschn. 3, Gl. (18)).

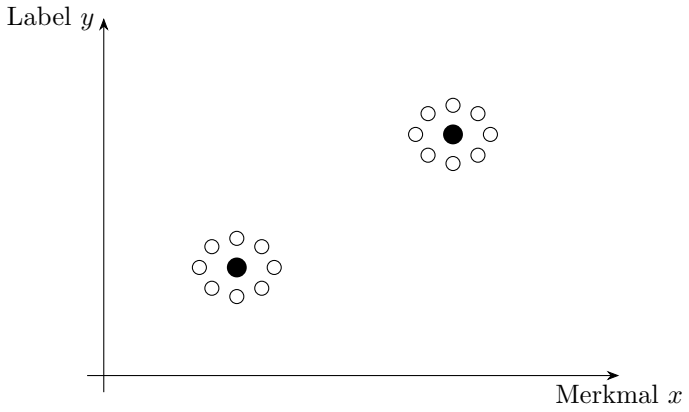

Abbildung 2: Die Datenerweiterung umgibt einen Datenpunkt (ausgefüllter Kreis) mit einer Wolke gestörter Kopien (offene Kreise). Sie ist eine Form der Regularisierung, die eine Methode des maschinellen Lernens weniger empfindlich gegenüber Merkmals- und Label-Rauschen macht

Siehe auch: Daten, Merkmal, Label, Datensatz, Datenerweiterung, Robustheit, KI-Verordnung.

Querverweise

Literatur

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

2. 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.

3. Golub and Van Loan (2013). Matrix Computations. 4th ed., The Johns Hopkins Univ. Press, Baltimore, MD, USA.

4. Cover and Thomas (2006). Elements of Information Theory. 2nd ed., Wiley, Hoboken, NJ, USA.

5. Kimothi (2002). The Uncertainty of Measurements: Physical and Chemical Metrology Impact and Analysis. American Society for Quality (ASQ), La Vergne.