Daten

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

Daten sind Repräsentationen von Informationen, die in einer Form aufgezeichnet wurden, die für Speicherung, Kommunikation und Verarbeitung geeignet ist. Der Begriff hat drei technische Bedeutungen im maschinellen Lernen. Im allgemeinen Sinn bezeichnet er aufgezeichnete Informationen in einer Form, die für die Verarbeitung durch ML-Methoden geeignet ist. Als Datenpunkt ist er eine elementare, informationstragende Einheit; als Datensatz eine indizierte Sammlung von Datenpunkte. Ein verwandter regulatorischer Begriff, personenbezogene Daten, ist in der Datenschutz-Grundverordnung (DSGVO) definiert und wird von der KI-Verordnung übernommen.

Definition

Eine Wetterstation misst die Temperatur an ihrem Standort und speichert den Wert 23.4 °C zusammen mit einem Zeitstempel (siehe Abb. 1(a)). Solche aufgezeichneten Werte sind Daten: Repräsentationen von Informationen, die in einer Form aufgezeichnet wurden, die für Speicherung, Kommunikation und Verarbeitung durch ML-Systeme (ML-System) geeignet ist (International Organization for Standardization and International Electrotechnical Commission, 2015). Jede maschinelle Lernen-Methode trainiert ein Modell auf Daten, und der Nutzen der durch die Methode gelernten Hypothese ist durch die Qualität und Quantität der verfügbaren Daten begrenzt.

Das Wort *Daten* wird in drei technischen Bedeutungen verwendet, die in Abb. 1 zusammengefasst sind, und zusätzlich in einer juristischen Bedeutung. Im weitesten Sinne bezeichnen Daten aufgezeichnete Informationen. Diese Bedeutung gilt, wenn es unerheblich ist, ob ein einzelner Datenpunkt oder eine Sammlung davon gemeint ist. Beispiele sind rohe Sensormessungen wie die Temperaturmessung in Abb. 1(a) und zusammengesetzte Begriffe wie Datenparallelismus, vernetzte Daten und Datenerweiterung.

Als formale Einheit ist ein Datenpunkt die grundlegende informationstragende Einheit (Abb. 1(b)). Seine Information ist in zwei Arten von Eigenschaften enthalten: Merkmale, die leicht messbar oder berechenbar sind, und Labels, die schwer messbar oder berechenbar sind oder eine menschliche Annotation erfordern. Das formale Modell einer Domänenmenge, einer Label-Menge und einer

Zuordnungsfunktion, die jedem Punkt der Domäne sein Label zuweist, wird von Shalev-Shwartz and Ben-David (2014), Abschnitt 2.1 gegeben.

Als formale Sammlung ist ein Datensatz $\mathcal{D} = \{\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(m)}\}$ die indizierte Sammlung von Datenpunkte, auf der Modell-Training und Validierung durchgeführt werden (Abb. 1(c)) (Goodfellow et al., 2016, Abschnitt 5.1). Die Qualität und Quantität der Daten, die eine gelernte Hypothese begrenzen, sind Eigenschaften dieser Sammlung: die Labels ihrer Datenpunkte und deren Anzahl.

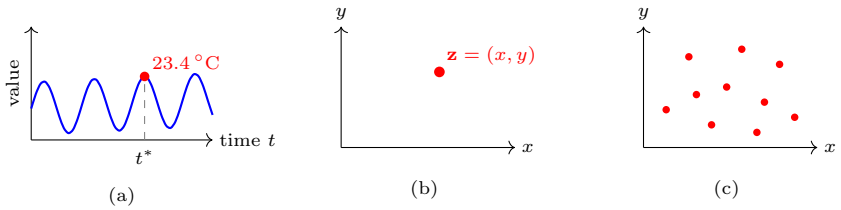

Abbildung 1: Drei technische Bedeutungen des Wortes *Daten* in ML-Anwendungen. (a) Daten im allgemeinen Sinn: eine zu einem Zeitpunkt t^* gemessene Temperatur. (b) Ein Datenpunkt $\mathbf{z} = (x, y)$ mit einem einzelnen Merkmal x und Label y. (c) Ein Datensatz, das aus einer endlichen Anzahl von Datenpunkte besteht.

Das Wort *Daten* hat auch eine regulatorische Bedeutung. Die Datenschutz-Grundverordnung (DSGVO) definiert personenbezogene Daten als alle Informationen, die sich auf eine identifizierte oder identifizierbare natürliche Person beziehen (European Parliament and Council of the European Union, 2016, Art. 4(1)). Die KI-Verordnung übernimmt diese Definition und definiert biometrische Daten als personenbezogene Daten, die aus einer bestimmten technischen Verarbeitung resultieren (European Parliament and Council of the European Union, 2024, Art. 3(34) und 3(50)). Mehrere regulatorische Verpflichtungen für ML-Systeme, darunter das Datensparsamkeitsprinzip, das Recht auf Löschung und das Recht auf Erklärung, beziehen sich auf Daten in diesem juristischen Sinn.

Siehe auch: Datenpunkt, Datensatz, personenbezogene Daten, DSGVO, KI-Verordnung.

Querverweise

Literatur

1. 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.

2. 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.

3. Goodfellow et al. (2016). Deep Learning. MIT Press, Cambridge, MA, USA.

4. International Organization for Standardization and International Electrotechnical Commission (2015). Information technology — Vocabulary. ISO/IEC 2382:2015. URL: https://www.iso.org/obp/ui/#iso:std:iso-iec:2382:ed-1:v2:en.

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