Datensatz

Autor und Herausgeber

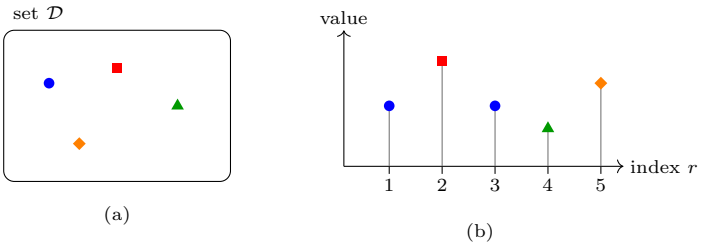
Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland (ORCID 0000-0001-7538-0990)
Editors: Konstantina Olioumtsevits (ORCID 0000-0003-2697-8659), Ekkehard Schnoor (ORCID 0000-0001-8252-1499), Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

Zusammenfassung

Ein Datensatz ist eine endliche Sammlung von Datenpunkte, an denen Modell Training und validation durchgeführt werden. Streng genommen ist ein Datensatz eine Menge von unterschiedlichen Datenpunkte. Mit leichter terminologischer Ungenauigkeit wird der Begriff auch synonym mit einer Stichprobe verwendet, einer endlichen Folge, die Wiederholungen enthalten kann. Als solche Folge gelesen, repräsentiert ein Datensatz eine Wahrscheinlichkeitsverteilung durch seine empirische Verteilung, wobei ein wiederholter Datenpunkt so oft zählt, wie er aufgeführt ist. Die Größe eines Datensatzes wird am einfachsten durch die Anzahl seiner Datenpunkte gemessen, was jedoch den Informationsgehalt korrelierter Datenpunkte überschätzt und nichts darüber aussagt, wie gut sie den Merkmalsraum abdecken. In der Praxis greifen machine learning (ML)-Methoden nicht auf die Datenpunkte selbst zu, sondern auf eine Repräsentation davon, wie z.B. eine Tabelle des relational model, die deren Attribute erfasst, aber nichts über die Eignung des Datensatzes für einen bestimmten Zweck aussagt.

Definition

Die Kühe in Abb. 2 oder die Zeilen von Tabelle 1, die sie erfassen, bilden einen Datensatz. Ein Datensatz ist eine Sammlung $\mathcal{D} = \{\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(m)}\}$ von Datenpunkte. Machine learning (ML)-Methoden verwenden einen Datensatz für das Modell Training und die validation. Streng genommen ist ein Datensatz eine ungeordnete Sammlung von unterschiedlichen Datenpunkte, d.h. eine Menge ohne Wiederholungen. In der ML-Literatur wird der Begriff jedoch oft für eine Stichprobe verwendet: eine Folge von m Datenpunkte, indiziert durch ihre Position $r = 1, \ldots, m$ in der Folge, die Wiederholungen enthalten kann (siehe Abb. 1).



(a) (b)

Abbildung 1: Zwei Lesarten des Wortes *Datensatz*. (a) Streng genommen ist ein Datensatz eine Menge von unterschiedlichen Datenpunkte, ohne Reihenfolge und ohne Wiederholungen. (b) In der ML-Literatur bezeichnet ein Datensatz oft eine Stichprobe: eine endliche Folge von Datenpunkte, indiziert durch r, die sich wiederholen können (Positionen $r = 1$ und $r = 3$ enthalten denselben Datenpunkt)

Als Liste gelesen, repräsentiert ein Datensatz auch eine Wahrscheinlichkeitsverteilung. Seine empirische Verteilung $\mathbb{P}^{(\mathcal{D})}$ weist jedem Eintrag der Liste die Wahrscheinlichkeit $1/m$ zu, sodass ein Datenpunkt, der zweimal aufgeführt ist, die Wahrscheinlichkeit $2/m$ erhält. Eine Menge, die einen Datenpunkt nicht zweimal auflisten kann, hat keine solchen Wiederholungen. Die Wiederholungen sind daher überall dort wesentlich, wo ein Datensatz eine Wahrscheinlichkeitsverteilung repräsentiert. Der Bootstrap zieht z.B. neue Datensätze aus $\mathbb{P}^{(\mathcal{D})}$ und listet so einige Datenpunkte mehrmals auf.

Die Größe eines Datensatzes hat mehr als ein Maß. Das grundlegendste ist die Anzahl m der Datenpunkte in der Liste. Wenn die Datenpunkte jedoch Realisierungen korrelierter Zufallsvariablen (ZV) sind, überschätzt m ihren Informationsgehalt: aufeinanderfolgende Tage einer Wetteraufzeichnung sind korreliert, sodass ihr Durchschnitt stärker streut als der Durchschnitt über ebenso viele zufällig aus der gesamten Aufzeichnung gezogene Tage (siehe Stichprobengröße). Die Größe eines Datensatzes wird dann nur relativ zu einem Wahrscheinlichkeitsmodell der Datengenerierung definiert, als die Anzahl der unabhängig und identisch verteilt (u.i.v.) Datenpunkte, die dieselbe Information tragen würden. Ein drittes Maß fragt, wie gut die Datenpunkte den Merkmalsraum abdecken: die größte Abstand von einem Punkt des interessierenden Bereichs zum nächsten Datenpunkt (siehe covering), die klein ist für einen Datensatz, dessen Merkmale verteilt sind, und groß für einen, der Teile des Merkmalsraum leer lässt, unabhängig davon, wie viele Datenpunkte er enthält.

Der Begriff eines Datensatzes ist weit gefasst: Datenpunkte können konkrete physische Entitäten (wie Menschen oder Tiere) oder abstrakte Objekte (wie Zahlen) repräsentieren. Zur Veranschaulichung zeigt Abb. 2 einen Datensatz, dessen Datenpunkte Kühe sind.



Abbildung 2: Kuhherde irgendwo in den Alpen

In ML-Anwendungen ist der Datensatz selbst oft nicht zugänglich. Zum Beispiel würde der Zugriff auf den Datensatz in Abb. 2 einen Besuch der Kuhherde erfordern. Was eine ML-Methode stattdessen verarbeitet, ist eine Repräsentation des Datensatzes innerhalb eines Daten Modells, eines Formalismus zur Darstellung und Verarbeitung von Daten (Abiteboul et al., 1995; Hoberman, 2009; Ramakrishnan and Gehrke, 2002; Silberschatz et al., 2019). Zum Beispiel organisiert das relational model Daten als eine Sammlung von Tabellen oder Relationen (Codd, 1970; Silberschatz et al., 2019).

Eine einzelne Tabelle besteht aus Zeilen und Spalten, wobei jede Zeile einem einzelnen Datenpunkt entspricht und jede Spalte ein spezifisches Attribut eines Datenpunkt repräsentiert. Die Reihenfolge der Zeilen spielt keine Rolle, und jedes Attribut ist mit einer Definitionsbereich verbunden, die seine Menge der erlaubten Werte festlegt. ML-Methoden verwenden diese Attribute als die Merkmale oder das Label eines Datenpunkt, sodass die Attribut-Definitionsbereiche dem Merkmalsraum und dem Labelraum entsprechen. Tabelle 1 zeigt ein Beispiel, das aus dem Datensatz in Abb. 2 gewonnen wurde.

TABLE 1
Relation (or Table) Representing a Dataset of Cows

Name	**Weight**	**Age**	**Height**	**Stomach temperature**
Zenzi	100	4	100	25
Berta	140	3	130	23
Resi	120	4	120	31

Eine Tabelle erfasst den Inhalt eines Datensatzes, aber nicht, ob dieser Inhalt für einen bestimmten Zweck geeignet ist. Ein in ethischen und rechtmäßigen ML-Systeme (ML-System) verwendeter Datensatz muss weitere Anforderungen erfüllen. Zum Beispiel verlangt der EU AI Act, dass Datensätze, auf denen hochriskante ML-Systeme trainiert werden, „ausreichend repräsentativ“ für die beabsichtigten Anwendungsfälle sind (European Parliament and Council of the European Union, 2024, Art. 10(3)).

Vertrauenswürdige künstliche Intelligenz (vertrauenswürdige KI) erfordert auch die Dokumentation, wie ein Datensatz zusammengestellt wurde, zum Beispiel durch ein Datenblatt, das die Motivation, Zusammensetzung und den Erhebungsprozess des Datensatzes festhält (Gebru et al., 2021). Wenn die Datenpunkte personal data sind, wird dies zu einer rechtlichen Pflicht: Nach dem Grundsatz der Zweckbindung verlangt die general data protection regulation (GDPR), dass solche Daten für „festgelegte, ausdrückliche und legitime Zwecke“ erhoben werden (European Parliament and Council of the European Union, 2016, Art. 5(1)(b)). Diese Zwecke werden zum Zeitpunkt der Erhebung festgelegt.

Siehe auch: Datenpunkt, Stichprobe, Daten, relational model, Merkmal, empirische Verteilung, Bootstrap, Stichprobengröße, covering, EU AI Act.

Querverweise

Literatur

1. Gebru et al. (2021). Datasheets for datasets. Commun. ACM, 64(12), pp. 86–92.

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. 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.

4. Abiteboul et al. (1995). Foundations of Databases. Addison-Wesley, Reading, MA, USA.

5. Codd (1970). A Relational Model of Data for Large Shared Data Banks. Commun. ACM, 13(6), pp. 377–387.

6. Hoberman (2009). Data Modeling Made Simple: A Practical Guide for Business and IT Professionals. 2nd ed., Technics Publications, Basking Ridge, NJ, USA.

7. Ramakrishnan and Gehrke (2002). Database Management Systems. 3rd ed., McGraw-Hill, New York, NY, USA.

8. Silberschatz et al. (2019). Database System Concepts. 7th ed., McGraw-Hill Education, New York, NY, USA.