tabellarische 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

Tabellarische Daten bestehen aus Datenpunkte, die sich eine gemeinsame, feste Menge von Attributen teilen. Die gemessenen Attribute dienen als die Merkmale oder als das Label eines Datenpunkts. Die Datenpunkte bilden die Zeilen oder Datensätze einer Tabelle, und die Attribute sind ihre Spalten. Jede Zeile hat eine Zelle pro Spalte, und die Zellen innerhalb einer Spalte enthalten Werte desselben Attributs. Die feste Attributmenge unterscheidet tabellarische Daten von anderen Arten von Daten, wie Text oder Graphen, deren Datenpunkte keine gemeinsame Menge von Attributen aufweisen.

Definition

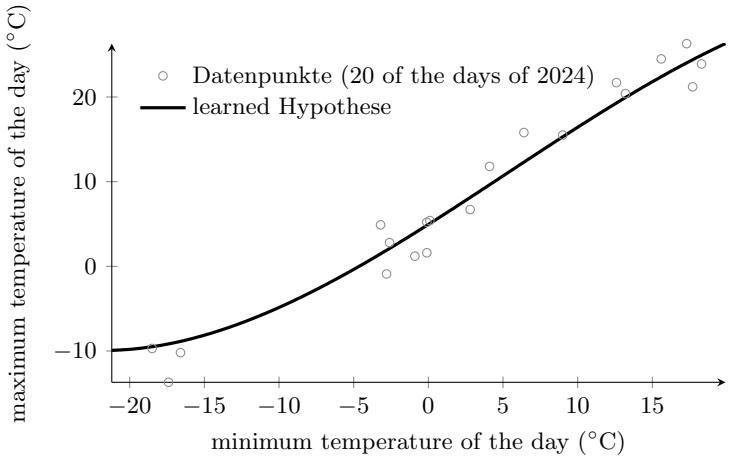
Eine Wetterstation des Finnisches Meteorologisches Institut (FMI) misst das Wetter jedes Tages, und die Messungen werden als eine Tabelle mit einer Zeile für jeden Tag dargestellt (Abb. 1(a)). Die Tabelle hat vier Spalten, eine pro Attribut eines Tages: das Datum, die morgendliche Tiefsttemperatur, den Niederschlag und die tägliche Höchsttemperatur (Finnish Meteorological Institute, 2024). Das Datum identifiziert den Datenpunkt, und die drei gemessenen Attribute sind die Kandidaten für seine Merkmale und sein Label. Tabellarische Daten bestehen aus Datenpunkte, die sich eine solche gemeinsame, feste Menge von Attributen teilen (Everitt, 2002). Wie der Name schon sagt, können tabellarische Daten als eine Tabelle gespeichert werden: Die Zeilen der Tabelle sind die Datenpunkte, und die Spalten der Tabelle sind die Attribute.

In Abb. 1(b) besteht die Lernaufgabe darin, die tägliche Höchsttemperatur eines Tages aus seiner morgendlichen Tiefsttemperatur vorherzusagen. Eine Methode des maschinelles Lernen (ML)s liest die Spalte der Tiefsttemperatur als das Merkmal und die Spalte der Höchsttemperatur als das Label ein, während die Niederschlagsspalte ungenutzt bleibt. Sie lernt eine Hypothese aus den 366 Datenpunkte, dargestellt als ein Streudiagramm mit der gelernten Kurve.

	Merkmal		Label
date	min. temp.	precip.	max. temp.
Jan 8	−19.8	−1.0	−2.9
Jan 9	−4.0	0.1	2.5
Jan 10	−0.2	−1.0	2.4
Jan 11	−9.2	2.1	2.4

one Datenpunkt per row

(a)



(b)

Abbildung 1: (a) Vier der 366 Zeilen der Tabelle täglicher Wettermessungen, die im Jahr 2024 an der Station Helsinki Kaisaniemi des FMI aufgezeichnet wurden (8.–11. Januar). Jede Zeile ist ein Datenpunkt, nämlich ein Tag. Die Kopfzeilen kürzen die morgendliche Tiefsttemperatur (min. temp.), den Niederschlag (precip.) und die tägliche Höchsttemperatur (max. temp.) ab. Jede Zeile hat eine Zelle pro Spalte. Die beiden schattierten Spalten sind diejenigen, die in Teilabbildung (b) verwendet werden, und die geschweiften Klammern darüber benennen die Rollen: Die hellgraue Spalte enthält das Merkmal und die dunkelgraue Spalte das Label eines Tages. (b) Streudiagramm von 20 der 366 Tage (jeder 19.), mit der Tiefsttemperatur eines Tages als seinem Merkmal und der Höchsttemperatur als seinem Label; die Kurve zeigt eine Hypothese, die aus allen 366 Datenpunkte gelernt wurde. Welche Spalten als Merkmal und Label einer Lernaufgabe dienen, ist eine Entwurfsentscheidung, und die Teilabbildung zeigt eine solche Wahl. Daten generiert durch `pythondemos/tabulardata.py`

Das charakteristische Merkmal tabellarischer Daten ist die starre Form einer Tabelle. Jede Zeile enthält dieselbe Anzahl von Zellen, eine pro Spalte, und die Zellen innerhalb einer Spalte enthalten Werte desselben Attributs der Da-

tenpunkte. Der Wertebereich muss für jede Spalte definiert sein, und jede Zelle einer Spalte enthält einen Wert aus diesem Bereich. Die Niederschlagsspalte in Abb. 1(a) enthält zum Beispiel eine Tagesmenge in Millimetern oder den speziellen Wert -1.0, der erfasst, dass kein Niederschlag beobachtet wurde. Eine zweite, andere Verwendung eines speziellen Wertes ist eine Markierung wie „N/A“, die für eine leere Zelle reserviert ist (Tennison et al., 2015). Fehlende Daten werden durch eine leere Zelle dargestellt: Ein Attribut, das für einen Datenpunkt nicht erfasst wurde, lässt seine Zelle leer, während die Tabelle ihre Form behält. Die Imputation füllt eine solche Zelle mit einem Schätzwert, der aus den erfassten Zellen berechnet wird.

Die feste Attributmenge ist das, was tabellarische Daten von anderen Arten von Daten unterscheidet. Bei den Wettermessungen zeichnet die Station jeden Tag dieselben Größen auf, sodass die Attributmenge durch den Messprozess festgelegt ist und von jedem Datenpunkt geteilt wird. Ein Text hingegen ist eine Folge von Tokens, deren Länge von Datenpunkt zu Datenpunkt variiert, und ein Graph variiert in seinen Knoten- und Kantenmengen. Solche Datenpunkte weisen von sich aus keine Attributmenge auf. Man kann eine vorgeben, zum Beispiel die Anzahl der Vorkommen jedes Wortes aus einer festen Wortliste, aber eine solche Menge verwirft die Reihenfolge der Tokens. Die Tabelle der Wettermessungen verliert auf diese Weise nichts, da ein Tag durch seine vier Attribute vollständig beschrieben wird, während ein Text mit seinen Tokens in einer anderen Reihenfolge ein anderer Text ist. Stattdessen werden Methoden des Repräsentationslernens entwickelt, die solche Datenpunkte auf Merkmalsvektoren (oder Einbettungen) abbilden (Kipf and Welling, 2017; Mikolov et al., 2013).

In der Datenbanktheorie formalisiert das relationales Modell eine Tabelle als eine Relation (Codd, 1970). Eine Relation ist eine Menge von Tupeln von Attributwerten. Das relationales Modell stellt auch die Operationen bereit, wie das Auswählen von Zeilen oder das Zusammenführen von Tabellen, die Datenbanksysteme zur Manipulation tabellarischer Daten verwenden (Silberschatz et al., 2019).

Siehe auch: Daten, Datenpunkt, Datensatz, Merkmal, Label, Merkmalsmatrix, relationales Modell, fehlende Daten, Imputation, Gradienten-Boosting-Entscheidungsbaum (GBDT), Repräsentationslernen.

Querverweise

Literatur

1. Everitt (2002). The Cambridge Dictionary of Statistics. 2nd ed., Cambridge Univ. Press, Cambridge, U.K.

2. Finnish Meteorological Institute (2024). The Finnish Meteorological Institute's open data. Open data web feature service, daily weather observations. URL: https://opendata.fmi.fi/wfs.

3. Mikolov et al. (2013). Efficient Estimation of Word Representations in Vector Space. In: Int. Conf. Learn. Represent. (ICLR), Workshop Track.

4. Kipf and Welling (2017). Semi-Supervised Classification with Graph Convolutional Networks. In: 5th Int. Conf. Learn. Representations.

5. Tennison et al. (2015). Model for Tabular Data and Metadata on the Web. W3C Recommendation. URL: https://www.w3.org/TR/tabular-data-model/.

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

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