Zufallsvariable (ZV)

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

Eine Zufallsvariable (ZV) ist eine messbare Funktion, die jedes Ergebnis eines Zufallsexperiments auf ein Element eines messbaren Raums abbildet. Zu den wichtigen Arten von ZV zählen binäre ZV, diskrete Zufallsvariablen, reellwertige ZV, Zufallsvektoren und Zufallsmatrizen. Die Analyse von Methoden des maschinelles Lernen (ML)s stützt sich häufig auf Wahrscheinlichkeitsmodelle, die Datenquellen als ZV darstellen. Das am weitesten verbreitete Wahrscheinlichkeitsmodell nimmt an, dass Datenpunkte unabhängig und identisch verteilt (u.i.v.) ZV mit einer gemeinsamen Wahrscheinlichkeitsverteilung sind, die im Allgemeinen unbekannt ist.

Definition

Das Wetter an einem gegebenen Tag in der österreichischen Stadt Krems an der Donau ist das Ergebnis eines Zufallsexperiments (Hyndman and Athanasopoulos, 2021). Jedes mögliche Ergebnis ω ist durch seine Lufttemperatur um 15:00 (UTC) charakterisiert, und das Erfassen dieser Temperatur ordnet dem Ergebnis eine reelle Zahl $x(\omega)$ zu. Diese Zuordnung ist eine ZV x. Die an aufeinanderfolgenden Tagen erfassten Temperaturen sind Realisierungen dieser ZV (siehe Abb. 1).

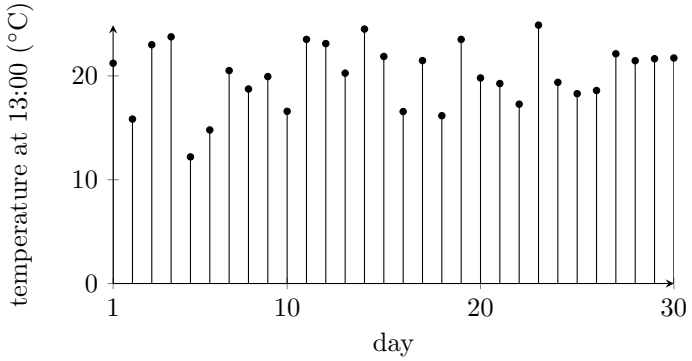

Abbildung 1: Die tägliche Lufttemperatur um 15:00 (UTC) ist eine Realisierung der reellwertigen ZV x; das Stabdiagramm zeigt eine Realisierung pro Tag über 30 aufeinanderfolgende Tage. Daten erzeugt durch `pythondemos/rv.py`

Allgemeiner bildet eine ZV die Ergebnisse eines Zufallsexperiments auf Elemente eines messbaren Raums ab (Billingsley, 1986; Gray, 2009). Mathematisch gesehen ist eine ZV eine messbare Funktion $x : \Omega \to \mathcal{X}$, deren Definitionsbereich der Ergebnisraum Ω eines Wahrscheinlichkeitsraums und deren Zielbereich ein messbarer Raum $\mathcal{X}$ ist. Messbarkeit bezieht sich auf die σ-Algebren der beiden Räume: Das Urbild $x^{-1}(B)$ jeder messbaren Menge $B \subseteq \mathcal{X}$ muss zur σ-Algebra des Wahrscheinlichkeitsraums gehören, sodass die Wahrscheinlichkeit dafür, dass x einen Wert in B annimmt, definiert ist. Der Wert $x(\omega) \in \mathcal{X}$, den x einem Ergebnis $\omega \in \Omega$ zuordnet, ist das Bild von ω unter x (siehe Abb. 2).

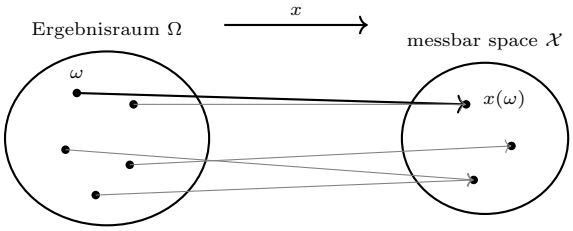

Abbildung 2: Eine ZV x bildet jedes Ergebnis ω eines Zufallsexperiments auf sein Bild $x(\omega)$ in einem messbaren Raum $\mathcal{X}$ ab. Der schwarze Pfeil folgt einem Ergebnis; die grauen Pfeile zeigen die übrigen Ergebnisse, von denen zwei dasselbe Bild teilen. Eine ZV kann daher einen großen Ergebnisraum in einen kleinen messbaren Raum abbilden

Zu den verschiedenen Arten von ZV gehören:

- binäre ZV, die jedes Ergebnis auf ein Element einer binären Menge abbilden (z. B. $\{-1, 1\}$ oder $\{\text{Katze}, \text{keine Katze}\}$);

- diskrete Zufallsvariablen, die Werte in einer abzählbaren Menge annehmen (die endlich oder abzählbar unendlich sein kann; z. B. die Anzahl von „Kopf“ bei zehn Münzwürfen mit Werten in $\{0, 1, \ldots, 10\}$);
- reellwertige ZV, die Werte in den reellen Zahlen $\mathbb{R}$ annehmen (z. B. die Temperatur um 15:00 (UTC) in Krems);
- Zufallsvektoren, die Ergebnisse in den euklidischer Raum $\mathbb{R}^d$ abbilden (z. B. die Rot-Grün-Blau-(RGB-)Werte eines Pixels, die einen Vektor in $\mathbb{R}^3$ bilden);
- Zufallsmatrizen, die Ergebnisse in einen Raum von Matrizen $\mathbb{R}^{m \times d}$ abbilden (z. B. die Daten-Matrix eines Datensatzes).

Der Ergebnisraum und die darauf definierten ZV sind eine Modellierungsentscheidung und nicht durch die Daten festgelegt. Die täglichen Temperaturen in Krems lassen zwei Standardkonstruktionen zu (Abb. 3). In der ersten hat jeder Tag r seinen eigenen Wahrscheinlichkeitsraum mit Ergebnisraum $\Omega^{(r)}$, und die Temperatur dieses Tages ist eine ZV $x^{(r)} : \Omega^{(r)} \to \mathbb{R}$; der erfasste Wert ist eine Realisierung $x^{(r)}(\omega^{(r)})$. In der zweiten repräsentiert ein Wahrscheinlichkeitsraum mit Ergebnisraum Ω alle Tage auf einmal: Ein einziges Ergebnis $\omega \in \Omega$ legt die Temperaturen aller Tage fest, und jeder Tag ist eine ZV $x^{(r)} : \Omega \to \mathbb{R}$ auf diesem gemeinsamen Raum. Die unabhängig und identisch verteilt (u.i.v.)-Annahme ist eine Aussage über die zweite Konstruktion: Die $x^{(r)}$ besitzen dieselbe Wahrscheinlichkeitsverteilung und sind unabhängig, eine Eigenschaft, die für Zufallsvariablen (ZV) auf einem gemeinsamen Wahrscheinlichkeitsraum definiert ist (Billingsley, 1986, Abschn. 20). Die Konstruktion mit gemeinsamem Raum wird für das Lernen aus Daten verwendet, da dann eine einzige Wahrscheinlichkeitsverteilung einen gesamten Datensatz bestimmt.

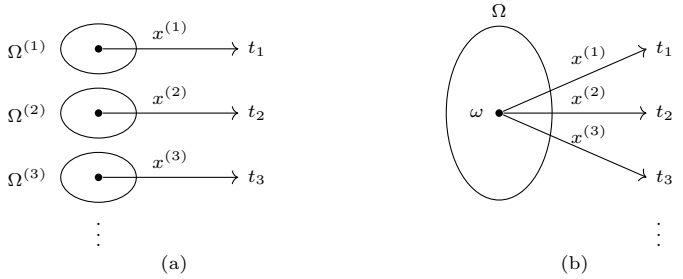

Abbildung 3: Zwei Konstruktionen für die täglichen Temperaturen t_r in Krems an der Donau. (a) Jeder Tag r hat seinen eigenen Wahrscheinlichkeitsraum mit Ergebnisraum $\Omega^{(r)}$ und ZV $x^{(r)}$, mit einem separaten Ergebnis pro Tag. (b) Ein Wahrscheinlichkeitsraum mit Ergebnisraum Ω repräsentiert alle Tage auf einmal, sodass ein einziges Ergebnis ω die Temperatur jedes Tages $t_r = x^{(r)}(\omega)$ festlegt; nur diese Konstruktion erlaubt die Modellierung der Tage als u.i.v. ZV

Im maschinelles Lernen (ML) repräsentieren ZV Datenquellen, sodass das Verhalten von Lernmethoden analysiert werden kann. Beispielsweise wird die Erzeugung eines Trainingsdatensatzes oder Validierungsdatensatzes in der Konstruktion mit gemeinsamem Raum durch u.i.v. ZV dargestellt, die eine zugrunde liegende Wahrscheinlichkeitsverteilung teilen. Diese Wahrscheinlichkeitsverteilung wird entweder postuliert oder aus tatsächlichen Datensätze geschätzt, beispielsweise mittels der maximale Likelihood. Idealerweise sollte eine ML-Methode auf Datensätze gut funktionieren, die als Realisierungen mit hoher Wahrscheinlichkeit auftreten. Eine gute Leistung auf solchen Datensätze ist das Ziel der Generalisierung, das die empirische Risikominimierung (ERM) verfolgt, indem sie den durchschnittlichen Verlust auf dem Trainingsdatensatz als Näherung für den erwarteten Verlust der gelernten Hypothese auf einem neuen, aus der gemeinsamen Wahrscheinlichkeitsverteilung gezogenen Datenpunkt minimiert.

Siehe auch: Funktion, Zufallsexperiment, Ergebnisraum, Wahrscheinlichkeitsraum, Vektor, euklidischer Raum, Wahrscheinlichkeit, messbar, Realisierung, Wahrscheinlichkeitsmodell, Zufallsvektor, Zufallsmatrix, sub-Gaußsche Zufallsvariable (sub-Gaußsche ZV).

Querverweise

Literatur

1. Billingsley (1986). Probability and Measure. 2nd ed., Wiley, New York, NY, USA.

2. Gray (2009). Probability, Random Processes, and Ergodic Properties. 2nd ed., Springer Science+Business Media, New York, NY, USA.

3. Hyndman and Athanasopoulos (2021). Forecasting: Principles and Practice. 3rd ed., OTexts, Melbourne, Australia.