Zufallsexperiment

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 Zufallsexperiment ist ein physikalischer oder rechnerischer Datengenerierender Prozess, der ein Ergebnis ω aus einer bekannten Menge von Möglichkeiten, dem sample space des Experiments, erzeugt. Die Wahrscheinlichkeittheorie stellt den Prozess durch einen probability space $(\Omega, \Sigma, \mathbb{P})$ dar, wobei die σ-Algebra Σ die measurable Ereignisse sammelt und das probability measure $\mathbb{P}(\cdot)$ jedem Ereignis eine Wahrscheinlichkeit zuweist. Diese Darstellung ist ein Wahrscheinlichkeitsmodell des Prozesses, keine Eigenschaft desselben. Eine Messung des Ergebnis wird durch eine Zufallsvariable (ZV), eine measurable Funktion $x : \Omega \to \mathcal{X}$, dargestellt, und der gemessene Wert $x(\omega)$ ist eine Realisierung. Das m-malige Wiederholen des Experiments ist selbst ein Zufallsexperiment auf dem Produkt-sample space, und die Messungen der Durchläufe werden als unabhängig und identisch verteilt (u.i.v.) Zufallsvariablen modelliert. Im machine learning (ML) sind die Konstruktion eines Trainingsdatensatz, die Ziehung einer Teilmenge davon durch stochastischer Gradientenabstieg (SGD) und das für differential privacy (DP) hinzugefügte Rauschen Zufallsexperiment.

Definition

Die Wetterstation Krems an der Donau (Österreich) zeichnet die Lufttemperatur um 06:00 und um 15:00 (UTC) jedes Tages auf (Abb. 1). An einem bestimmten Tag ist keine der beiden Temperaturen im Voraus bekannt, aber der Bereich, den sie annehmen können, ist bekannt. Ein solcher Prozess ist ein Zufallsexperiment: Er erzeugt ein Ergebnis ω aus einer Menge Ω von Möglichkeiten, dem sample space des Experiments (Bertsekas and Tsitsiklis, 2008, Sect. 1.2; Billingsley, 1986). Das Ergebnis ist unsicher, und das Experiment kann unter identischen Bedingungen wiederholt werden. Hier ist ein Tag ein Durchlauf: Das Ergebnis ω ist das Wetter dieses Tages, und die beiden aufgezeichneten Temperaturen sind Messungen davon.

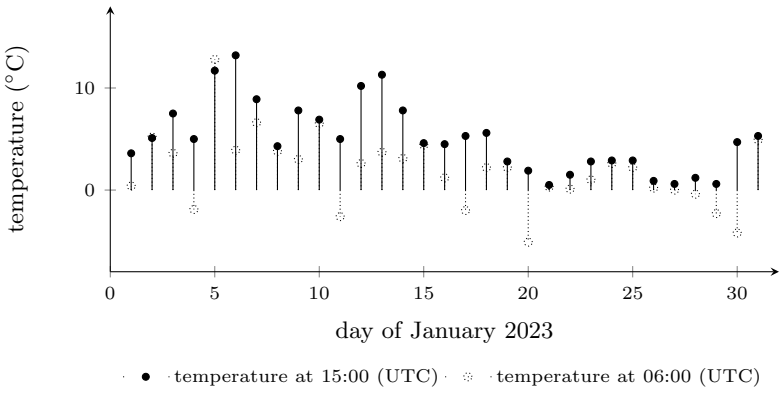

Abbildung 1: Die Lufttemperatur um 15:00 (durchgezogene Stiele, gefüllte Kreise) und um 06:00 (UTC) (gepunktete Stiele, offene Kreise) jedes Tages im Januar 2023 an der Wetterstation Krems an der Donau (Österreich). Jeder Tag ist ein Durchlauf desselben Zufallsexperiments: Das Ergebnis ist das Wetter dieses Tages, und die beiden Temperaturen sind Messungen davon. Der Monat als Ganzes kann stattdessen als ein einziges Zufallsexperiment gelesen werden, dessen Ergebnis die geordnete Liste der 31 täglichen Ergebnisse ist. Daten generiert von `pythondemos/randomexperiment.py`

Die Wahrscheinlichkeittheorie stellt ein Zufallsexperiment durch einen probability space $(\Omega, \Sigma, \mathbb{P})$ dar (Bertsekas and Tsitsiklis, 2008, Sect. 1.2; Billingsley, 1986, Sect. 2). Der sample space Ω sammelt die möglichen Ergebnisse $\omega \in \Omega$. Die σ-Algebra Σ sammelt die Ereignisse, jene Teilmengen $\mathcal{A} \subseteq \Omega$, denen eine Wahrscheinlichkeit zugewiesen wird. Das probability measure $\mathbb{P}(\cdot)$ weist jedem Ereignis $\mathcal{A} \in \Sigma$ eine Zahl $\mathbb{P}(\mathcal{A}) \in [0, 1]$ zu, sodass $\mathbb{P}(\Omega) = 1$ und die Wahrscheinlichkeit einer countable Vereinigung paarweise disjunkter Ereignisse die Summe ihrer Wahrscheinlichkeiten ist. Das Tripel $(\Omega, \Sigma, \mathbb{P})$ ist ein Wahrscheinlichkeitsmodell des Prozesses, keine Eigenschaft desselben. Es gibt prinzipielle Methoden zur Auswahl eines nützlichen Wahrscheinlichkeitsmodell $(\Omega, \Sigma, \mathbb{P})$ basierend auf den beobachteten Daten.

Das Würfeln ist ein Zufallsexperiment, das klein genug ist, um das Tripel auszuschreiben. Sein sample space ist $\Omega = \{1, \ldots, 6\}$. Seine σ-Algebra Σ ist die Sammlung aller Teilmengen von Ω: dass der Wurf gerade ist, ist das Ereignis $\mathcal{A} = \{2, 4, 6\}$. Das uniforme probability measure, $\mathbb{P}(\mathcal{A}) = |\mathcal{A}|/6$, ist das Modell eines fairen Würfels, unter dem $\mathbb{P}(\{2, 4, 6\}) = 1/2$. Ob ein gegebener Würfel fair ist, ist eine Frage über den Prozess und nicht über das Modell, beantwortet durch die relativen Häufigkeiten seiner Ergebnisse bei vielen Würfen.

Die beobachteten Daten sind Messungen des Ergebnis des Experiments. Insbesondere ist jede Messung eine measurable Funktion $x : \Omega \to \mathcal{X}$ des Ergebnis. Eine solche measurable Funktion ist eine Zufallsvariable (ZV), und der Wert $x(\omega)$, den sie für das tatsächliche Ergebnis ω annimmt, ist eine Realisierung davon (Billingsley, 1986; Gray, 2009). Die Temperatur um 13:00 Uhr ist eine

ZV für das Wetter eines Tages, und die an einem Tag abgelesene Zahl ist eine Realisierung.

Das probability measure auf Ω induziert die Wahrscheinlichkeitsverteilung der ZV, $\mathbb{P}^{(x)}(\mathcal{B}) = \mathbb{P}(x \in \mathcal{B})$ für jede measurable Menge $\mathcal{B} \subseteq \mathcal{X}$ (Billingsley, 1986, Sect. 20). Abb. 2 zeigt den Prozess neben seiner mathematischen Darstellung: das Experiment, sein Ergebnis und der gemessene Wert oben, der probability space und die ZV, die sie unten darstellen.

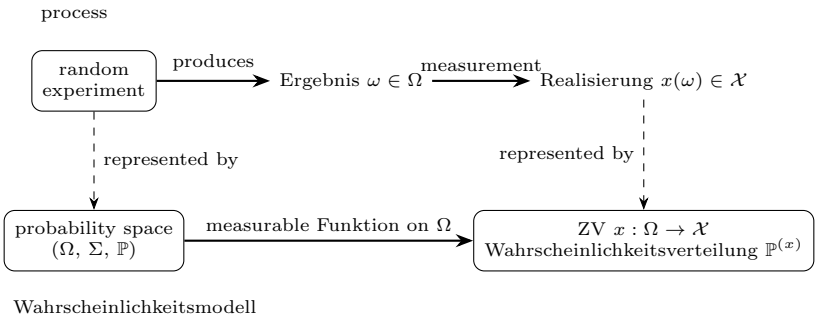

Abbildung 2: Ein Zufallsexperiment und seine mathematische Darstellung. Die obere Zeile ist der Prozess, die untere Zeile das Modell, wie die Beschriftungen links markieren. Umrahmte Elemente sind Objekte, das Experiment, der probability space und die ZV; nicht umrahmte Elemente sind die Werte, die sie erzeugen. Durchgezogene Pfeile bleiben innerhalb einer Zeile; die gestrichelten Pfeile kreuzen vom Prozess zu seinem Modell. Der Prozess erzeugt ein Ergebnis ω; eine Messung dieses Ergebnis ist eine measurable Funktion x auf Ω, eine ZV, und der Wert $x(\omega)$, den sie annimmt, ist eine Realisierung. Die Wahrscheinlichkeitsverteilung $\mathbb{P}^{(x)}$ von x wird durch $\mathbb{P}(\cdot)$ induziert.

Das Wiederholen eines Zufallsexperiments $(\Omega, \Sigma, \mathbb{P})$ eine feste Anzahl m Mal unter identischen Bedingungen definiert ein neues Zufallsexperiment $(\Omega, \Sigma, \mathbb{P})^{(m)}$, dessen Ergebnisse Folgen $(\omega^{(1)}, \ldots, \omega^{(m)})$ von Ergebnisse des ursprünglichen Experiments sind (Abb. 3). Sein sample space ist das Produkt $\Omega^{(m)} = \Omega \times \cdots \times \Omega$, und sein probability measure ist das Produkt $\mathbb{P}^{(m)} = \mathbb{P} \times \cdots \times \mathbb{P}$. Die Messungen $x^{(1)}, \ldots, x^{(m)}$ der Durchläufe sind dann unabhängig und identisch verteilt (u.i.v.) Zufallsvariablen (Billingsley, 1986, Sect. 20). Im machine learning (ML) sind diese Messungen die Datenpunkte $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(m)}$ eines Trainingsdatensatz. Während das Ergebnis eines einzelnen Experiments unsicher ist, wird das Langzeitverhalten wiederholter Experimente vorhersagbar: Das law of large numbers und der central limit theorem (CLT) präzisieren dies (Bertsekas and Tsitsiklis, 2008, Sects. 5.2 and 5.4).

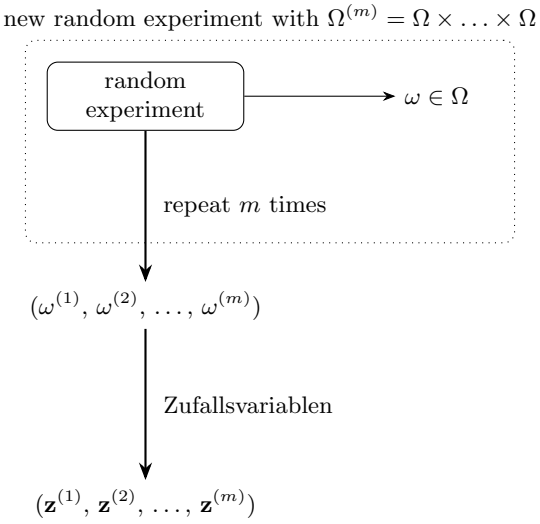

Abbildung 3: Das Wiederholen eines Zufallsexperiments $(\Omega, \Sigma, \mathbb{P})$ eine feste Anzahl m Mal ergibt ein weiteres Zufallsexperiment $(\Omega, \Sigma, \mathbb{P})^{(m)}$, dessen Ergebnisse Folgen $(\omega^{(1)}, \omega^{(2)}, \ldots, \omega^{(m)}) \in \Omega \times \ldots \times \Omega$ sind. Ein Zufallsexperiment, das in vielen ML-Anwendungen auftritt, ist die Sammlung eines Trainingsdatensatz $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(m)}$, dessen Datenpunkte Zufallsvariablen bei den wiederholten Durchläufen sind. Kästen und Pfeile folgen Abb. 2

Die Vorstellung eines wiederholten Experiments ist eine Idealisierung, die zu einem schlechten Wahrscheinlichkeitsmodell für einen physikalischen oder rechnerischen Daten-generierenden Prozess führen kann. Aufeinanderfolgende Würfe eines Würfels können typischerweise als wiederholtes Experiment betrachtet werden. Im Gegensatz dazu sind die täglichen Wetterbeobachtungen, die in Abb. 1 gezeigt werden, nicht gut als Ergebnis eines wiederholten Experiments darstellbar: Ein warmer Tag folgt tendenziell einem warmen Tag, sodass die täglichen Messwerte eher miteinander verbunden als unabhängig sind. Die autocorrelation einer solchen Aufzeichnung misst diese Beziehung (Brockwell and Davis, 1991, Sect. 1.5; Hyndman and Athanasopoulos, 2021).

Zufallsexperiment treten im ML an drei Stellen auf:

- Datensammlung: Die Datenpunkte $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(m)}$ eines Trainingsdatensatz werden als Zufallsvariablen modelliert, d.h. als Funktionen des Ergebnis $\omega \in \Omega$ eines Zufallsexperiments, und die Annahme unabhängig und identisch verteilter Daten (u.i.v.-Annahme) modelliert ihre Sammlung als Wiederholung eines Experiments. Empirische Risikominimierung (ERM) minimiert den durchschnittlichen Verlust über diese m Datenpunkte; da sie u.i.v. Zufallsvariablen mit Wahrscheinlichkeitsverteilung $\mathbb{P}^{(\mathbf{z})}$ sind, schätzt dieser Durchschnitt das Risiko einer Hypothese, einen Erwartungswert unter $\mathbb{P}^{(\mathbf{z})}$, was das Risiko zu einem sinnvollen Ziel für ERM macht.

- Stochastic Algorithmen: stochastischer Gradientenabstieg (SGD) führt in jeder Iteration ein Zufallsexperiment durch, die Ziehung einer Teilmenge des Trainingsdatensatz; das Bootstrap zieht Resamples des Trainingsdatensatz.
- Privacy protection-Methoden führen ein Zufallsexperiment durch, um die Ausgaben einer ML-Methode zu stören, was differential privacy (DP) erfordert.

Siehe auch: Ergebnis, sample space, Ereignis, ZV, Realisierung, Wahrscheinlichkeit, probability space, probability measure, Wahrscheinlichkeitsverteilung, Wahrscheinlichkeitsmodell, u.i.v., law of large numbers, CLT, autocorrelation.

Querverweise

Literatur

1. Bertsekas and Tsitsiklis (2008). Introduction to Probability. 2nd ed., Athena Scientific, Belmont, MA, USA.
2. Billingsley (1986). Probability and Measure. 2nd ed., Wiley, New York, NY, USA.
3. Brockwell and Davis (1991). Time Series: Theory and Methods. 2nd ed., Springer-Verlag, New York, NY, USA.
4. Gray (2009). Probability, Random Processes, and Ergodic Properties. 2nd ed., Springer Science+Business Media, New York, NY, USA.
5. Hyndman and Athanasopoulos (2021). Forecasting: Principles and Practice. 3rd ed., OTexts, Melbourne, Australia.