Bootstrap

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

Das Bootstrap ist ein stochastischer Algorithmus, der gestörte Varianten, sogenannte Bootstrap-Stichproben, eines gegebenen Datensatz erzeugt. Jede Bootstrap-Stichprobe wird durch zufälliges Auswählen von Datenpunkte mit Zurücklegen aus dem ursprünglichen Datensatz gewonnen. Die Bootstrap-Stichproben haben die gleiche Größe wie das ursprüngliche Datensatz. Sie können jedoch einige der ursprünglichen Datenpunkte wiederholen, während sie andere weglassen. Mehrere Bootstrap-Stichproben ermöglichen die Untersuchung der Eigenschaften einer machine learning (ML)-Methode. Zum Beispiel liefert eine robuste ML-Methode ähnliche Modellparameter für zwei Datensätze, die sich nur in den zufälligen Schwankungen der Datengenerierung unterscheiden. Bootstrap-Stichproben können auch verwendet werden, um ein confidence interval für einen Validierungsfehler zu schätzen. Eine weitere Anwendung des Bootstrap ist Bagging, bei dem die aus den Bootstrap-Stichproben gelernten Hypothesen zu einer einzigen aggregiert werden.

Definition

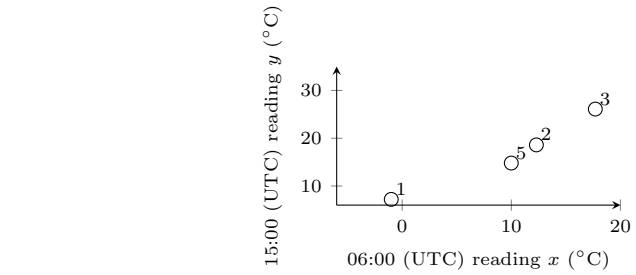
Fünf Tage Wetteraufzeichnungen in Krems an der Donau (Österreich) sind alle verfügbaren Daten. Jeder Tag ist ein Datenpunkt, dessen Merkmal die Lufttemperatur um 06:00 (UTC) und dessen Label die Temperatur um 15:00 (UTC) desselben Tages ist (Abb. 1). Dieses Datensatz ist klein: Es erlaubt kein Training eines hochdimensionalen Modell, und eine daraus gelernte Hypothese hängt davon ab, welche fünf Tage zufällig aufgezeichnet wurden. Das Bootstrap ist ein stochastischer Algorithmus, wie in Algorithmus 1 beschrieben, der zusätzliche Datensätze aus dem aufgezeichneten generiert, die sogenannten Bootstrap-Stichproben (Efron and Tibshirani, 1993), und so diese Abhängigkeit messbar macht. Abb. 1 zeigt zwei Bootstrap-Stichproben, die mit Algorithmus 1 gewonnen wurden. Eine Bootstrap-Stichprobe kann einen Datenpunkt von $\mathcal{D}$ mehrmals enthalten, daher ist sie eher eine Liste von Datenpunkte als eine Menge. Das Bootstrap basiert auf der Annahme unabhängig und identisch verteilter Daten (u.i.v.-Annahme).

Algorithm 1 Das Bootstrap (Efron, 1979, Sect. 2)

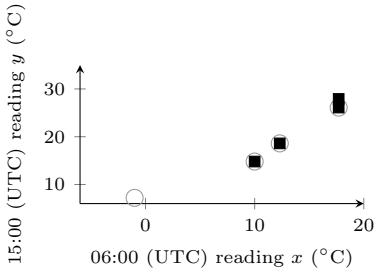
Input: Datensatz $\mathcal{D} = \left\{\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(m)}\right\}$, number B of bootstrap samples

1: **for** $b = 1, \ldots, B$ **do**
2: **for** $r = 1, \ldots, m$ **do**
3: draw an index i uniformly at random from $\{1, \ldots, m\}$
4: $\widetilde{\mathbf{z}}^{(r)} := \mathbf{z}^{(i)}$
5: **end for**
6: $\widetilde{\mathcal{D}}^{(b)} := \left(\widetilde{\mathbf{z}}^{(1)}, \ldots, \widetilde{\mathbf{z}}^{(m)}\right)$
7: **end for**

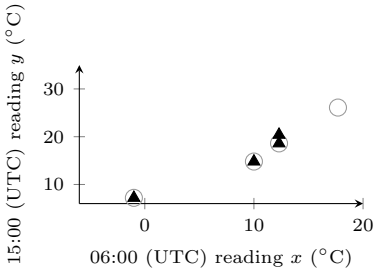
Output: bootstrap samples $\widetilde{\mathcal{D}}^{(1)}, \ldots, \widetilde{\mathcal{D}}^{(B)}$



(a) the Datensatz $\mathcal{D}$ of five days



(b) bootstrap sample $\widetilde{\mathcal{D}}^{(1)}$



(c) bootstrap sample $\widetilde{\mathcal{D}}^{(2)}$

Abbildung 1: Ein Datensatz von fünf Datenpunkte, die die täglichen Wetterbedingungen in Krems an der Donau am 15. Februar, April, Juni, August und Oktober 2022 darstellen, und darunter zwei Bootstrap-Stichproben, die daraus mit Algorithmus 1 gewonnen wurden. In (a) ist jeder Tag ein Datenpunkt (offener Kreis, nummeriert). In (b) und (c) wiederholen die grauen offenen Kreise die fünf Tage, und die gefüllten Markierungen sind die fünf Ziehungen der Bootstrap-Stichprobe: $\widetilde{\mathcal{D}}^{(1)}$ (gefüllte Quadrate) enthält die Tage $2, 3, 3, 4, 5$ und lässt Tag 1 aus, während $\widetilde{\mathcal{D}}^{(2)}$ (gefüllte Dreiecke) die Tage $1, 2, 2, 4, 5$ enthält und Tag 3 auslässt. Die zweite Markierung eines zweimal gezogenen Tages sitzt leicht über der ersten, sodass sich zwei Ziehungen desselben Tages nicht gegenseitig verdecken. Wetterdaten: GeoSphere Austria (GeoSphere Austria, 2026). Daten generiert von `pythondemos/bootstrap.py`

Hier wird ein Datensatz $\mathcal{D} = \left\{\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(m)}\right\}$ als die Realisierungen von m unabhängig und identisch verteilt (u.i.v.) Zufallsvariablen (ZV) mit einer gemeinsamen Wahrscheinlichkeitsverteilung $\mathbb{P}$ interpretiert. Diese $\mathbb{P}$ ist unbekannt und muss aus $\mathcal{D}$ selbst geschätzt werden. Das Bootstrap schätzt sie durch die empirische Verteilung $\mathbb{P}^{(\mathcal{D})}$ von $\mathcal{D}$ (Efron, 1979; Efron and Tibshirani, 1993; Hastie et al., 2009, Sect. 8.2.1). Für jedes Ereignis $\mathcal{A}$ approximiert der Anteil der Datenpunkte $\mathbf{z}^{(r)}$, die hineinfallen, seine Wahrscheinlichkeit durch das law of large numbers (Billingsley, 1986, Thm. 6.1; Efron and Tibshirani, 1993, Sect. 4.2),

$$\frac{1}{m}\left|r : \mathbf{z}^{(r)} \in \mathcal{A}\right| \approx \mathbb{P}\left(\mathcal{A}\right).$$

Die empirische Verteilung $\mathbb{P}^{(\mathcal{D})}$ ist eine vollständig spezifizierte Wahrscheinlichkeitsverteilung und kann beliebig oft abgetastet werden. So können beliebig viele Datenpunkte und Datensätze aus dem gesammelten Datensatz generiert werden. Der Gewinn sind Kopien, nicht Informationen. Jede Kopie behält die Größe m von $\mathcal{D}$, da die imitierte Größe die Streuung eines Datensatz dieser Größe ist, das aus $\mathbb{P}$ gezogen wurde. Auch enthält eine Kopie keine neuen Datenpunkte: $\mathbb{P}^{(\mathcal{D})}$ konzentriert ihre gesamte Masse auf die m beobachteten Datenpunkte, sodass keine Ziehung einen erzeugen kann, den $\mathcal{D}$ nicht bereits enthält. Eine Variante, das geglättete Bootstrap, zieht aus einer geglätteten Version von $\mathbb{P}^{(\mathcal{D})}$, bei der jeder Datenpunkt durch eine kleine, beulenförmige Wahrscheinlichkeitsdichte ersetzt wird; es erzeugt Datenpunkte, die $\mathcal{D}$ nicht enthält, was das Bootstrap selbst niemals tut.

Das Ziehen von m Datenpunkte aus $\mathbb{P}^{(\mathcal{D})}$ ist dasselbe wie das m-fache Ziehen aus $\mathcal{D}$ mit Zurücklegen, was Algorithmus 1 tut: Jede Ziehung erfolgt aus dem gesamten $\mathcal{D}$, sodass ein Datenpunkt zweimal und ein anderer überhaupt nicht gezogen werden kann (Abb. 2). Die Anzahl B der Bootstrap-Stichproben ist nur durch den Rechenaufwand begrenzt, den sie verursachen.

$\mathcal{D}$	1	2	3	4	5	6	7	8	
$\widetilde{\mathcal{D}}^{(1)}$	2	5	6	6	7	7	8	8	omits $1, 3, 4$
$\widetilde{\mathcal{D}}^{(2)}$	1	1	3	3	4	7	7	8	omits $2, 5, 6$
$\widetilde{\mathcal{D}}^{(3)}$	1	2	3	3	3	4	7	7	omits $5, 6, 8$

Abbildung 2: Drei Bootstrap-Stichproben $\widetilde{\mathcal{D}}^{(1)}$, $\widetilde{\mathcal{D}}^{(2)}$, $\widetilde{\mathcal{D}}^{(3)}$, gezogen aus einem Datensatz $\mathcal{D}$ von $m = 8$ Datenpunkte (schattiert), jeweils durch 8-maliges Ziehen mit Zurücklegen. Jede Zeile hat die gleiche Länge wie $\mathcal{D}$, wiederholt einige Datenpunkte und lässt andere aus; jede der drei enthält zufällig 5 der 8 unterschiedlichen Datenpunkte, im Gegensatz zu den $0.63 \cdot 8 \approx 5$, die die Berechnung unten vorhersagt

Ein Datenpunkt wird mit Wahrscheinlichkeit $(1 - 1/m)^m$ aus einer Bootstrap-Stichprobe weggelassen, sodass er mit Wahrscheinlichkeit $1 - (1 - 1/m)^m$ in einer erscheint, was sich $1 - 1/e \approx 0.63$ nähert, wenn m wächst, da $(1 - 1/m)^m$ gegen $1/e$ konvergiert. Die Anzahl der unterschiedlichen Datenpunkte in einer Bootstrap-Stichprobe ergibt sich allein aus dieser Wahrscheinlichkeit. Sei I_r die ZV, die den Wert eins annimmt, wenn $\mathbf{z}^{(r)}$ in $\widetilde{\mathcal{D}}^{(b)}$ erscheint, und null sonst. Die Anzahl der unterschiedlichen Datenpunkte ist $\sum_{r=1}^{m} I_r$, sodass ihr Erwartungswert ist

$$\mathbb{E}\Big\{ \sum_{r=1}^{m} I_r \Big\} = \sum_{r=1}^{m} \mathbb{P}\left(I_r = 1\right) = m\big(1 - (1 - 1/m)^m\big) \approx 0.63\, m.$$

Die I_r sind nicht unabhängig: Ein zweimal gezogener Tag zwingt dazu, einen

anderen wegzulassen. Der Erwartungswert einer Summe ist dennoch die Summe der Erwartungswerte.

Eine Bootstrap-Stichprobe enthält daher etwa $0.63\,m$ unterschiedliche Datenpunkte und überlappt $\mathcal{D}$ stark. Diese Überlappung unterscheidet das Bootstrap von der k-fache Kreuzvalidierung, die $\mathcal{D}$ in Blöcke zerlegt, die sich überhaupt nicht überlappen. Naiv verwendet, um den Vorhersage-Fehler einer gelernten Hypothese zu schätzen, bewertet das Bootstrap diese daher auf Datenpunkte, die in ihr eigenes Training eingeflossen sind, und die Schätzung fällt optimistisch aus (Hastie et al., 2009, Sect. 7.11).

Machine learning (ML)-Training-Methoden können separat auf jede Bootstrap-Stichprobe angewendet werden, was die gelernten Hypothesen $\widehat{h}^{(1)}, \ldots, \widehat{h}^{(B)}$ ergibt. Eine aus $\widetilde{\mathcal{D}}^{(b)}$ berechnete Größe ist eine Bootstrap-Replikation (Efron and Tibshirani, 1993), wie die Hypothese $\widehat{h}^{(b)}$, die durch empirische Risikominimierung (ERM) aus $\widetilde{\mathcal{D}}^{(b)}$ gelernt wurde, oder ihre Vorhersage $\widehat{h}^{(b)}(\mathbf{x})$ für einen festen Datenpunkt. Die Streuung der B Bootstrap-Replikationen schätzt, wie stark sich die Ausgabe der ML-Methode aufgrund zufälliger Schwankungen in der Datengenerierung ändert (Hastie et al., 2009, Sect. 8.2.1).

Abb. 3 zeigt das Bootstrap zum Lernen einer linearen Hypothese aus Wetterdaten, die 2024 in Krems an der Donau aufgezeichnet wurden, den $m = 365$ Tagen, die beide Messwerte tragen. Jeder Punkt ist ein Datenpunkt, dessen Merkmal der 06:00-Messwert und dessen Label der 15:00-Messwert des Tages ist; die Abbildung zeigt 20 davon. Die dicke Linie ist die Hypothese, die aus allen 365 durch ERM mit dem squared error loss gelernt wurde. Die dünnen Linien zeigen Bootstrap-Replikationen $\widehat{h}^{(b)}$, $b = 1, \ldots, B$, die durch dasselbe ERM-Problem aus Bootstrap-Stichproben gelernt wurden. Ihre Streuung ist hier gering, und das 95%-confidence interval für die Steigung ist $[1.00, 1.09]$, um die Steigung 1.04, die aus allen 365 Tagen gelernt wurde. An einem Viertel der Tage ist dieses Intervall 1.7-mal so breit.

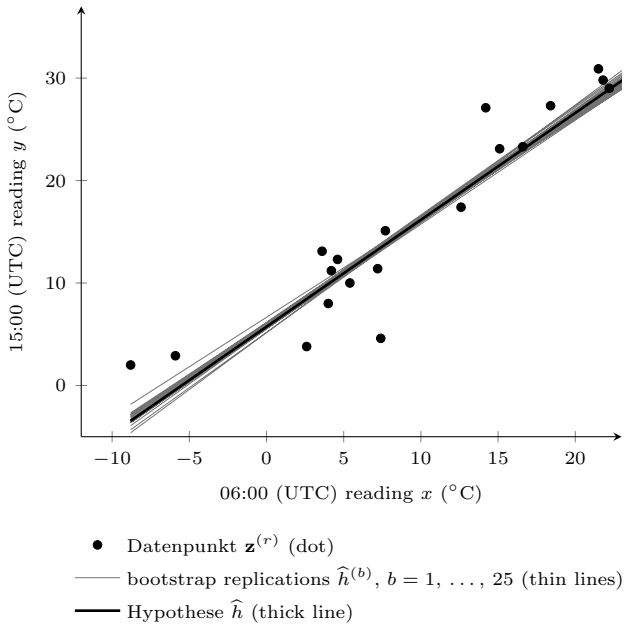


Abbildung 3: Die Lufttemperatur um 06:00 (UTC) und 15:00 (UTC) in Krems an der Donau im Jahr 2024. Die Punkte sind 20 der $m = 365$ Datenpunkte, jeder 19. Tag des Jahres. Die dicke Linie ist die Hypothese $\widehat{h}$, die aus allen m Datenpunkte gelernt wurde, die dünnen Linien sind die Bootstrap-Replikationen $\widehat{h}^{(b)}$, die aus den ersten 25 von $B = 500$ Bootstrap-Stichproben gelernt wurden, und die Streuung aller 500 ergibt das confidence interval für die Steigung. Wetterdaten: GeoSphere Austria (GeoSphere Austria, 2026). Daten generiert von `pythondemos/bootstrap.py`

Während die grundlegende validation eine Punktschätzung für das Risiko einer Hypothese liefert, d.h. den Validierungsfehler, fügt das Bootstrap dieser Punktschätzung ein confidence interval hinzu. Eine Hypothese $\widehat{h}$ wird einmal aus dem Trainingsdatensatz $\mathcal{D}^{(\text{train})}$ gelernt, und das Validierungsdatensatz $\mathcal{D}^{(\text{val})}$ wird B-mal mit Zurücklegen neu abgetastet, wobei jede Bootstrap-Stichprobe die Größe von $\mathcal{D}^{(\text{val})}$ hat. Die Bewertung derselben $\widehat{h}$ auf jeder von ihnen liefert B Schätzungen des Risiko. Der Stichprobenmittel dieser Schätzungen ist die Risiko-Schätzung, und die Quadratwurzel ihrer Stichprobenvarianz ist ihr Bootstrap-Standardfehler (Efron and Tibshirani, 1993, Sect. 6.2). Die Schätzung plus oder minus ein Standardnormalquantil mal diesem Standardfehler ist ein confidence interval (Efron and Tibshirani, 1993, Sect. 13.2). Für eine Schwellenwertregel in Krems, die um 06:00 vor Frost warnt, wenn der 15:00-Messwert des Tages unter 3.3 Grad fällt, gelernt an 250 Tagen und beurteilt an den verbleibenden 115, beträgt der Validierungsfehler 0.087 und die 500 Bootstrap-Schätzungen ergeben das confidence interval $[0.043, 0.139]$. Dieses Intervall ver-

birgt, was die einzelne Zahl nicht zeigt: Immer keinen Frost vorherzusagen, erzielt 0.070 an denselben Tagen, gut innerhalb des Intervalls, sodass der Validierungsfehler allein nicht beweist, dass die Regel die konstante übertrifft.

Bagging ist das Bootstrap, das anders eingesetzt wird: Anstatt die Streuung von $\widehat{h}^{(1)}, \ldots, \widehat{h}^{(B)}$ zu messen, mittelt es diese zu einer Hypothese, wodurch ein Random Forest aus Entscheidungsbäume aufgebaut wird.

Siehe auch: empirische Verteilung, u.i.v., ZV, Wahrscheinlichkeitsverteilung, Varianz, k-fache Kreuzvalidierung, Bagging, confidence interval, validation, Validierungsdatensatz, Validierungsfehler.

Querverweise

Literatur

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

2. Efron and Tibshirani (1993). An Introduction to the Bootstrap. Chapman & Hall, New York, NY, USA.

3. Efron (1979). Bootstrap Methods: Another Look at the Jackknife. Ann. Statist., 7(1), pp. 1–26.

4. GeoSphere Austria (2026). GeoSphere Austria climate station archive. Data hub, station historical resource klima-v2-1d. URL: https://dataset.api.hub.geosphere.at/v1/station/hist v2-1d.

5. Hastie et al. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed., Springer Science+Business Media, New York, NY, USA.