Varianz

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

Die Varianz einer reellwertigen Zufallsvariable (ZV) ist der Erwartungswert der quadrierten Differenz zwischen der ZV und ihrem Mittelwert. Sie quantifiziert die Streuung der Wahrscheinlichkeitsverteilung der ZV um den Mittelwert: Eine kleine Varianz bedeutet, dass sich die Realisierungen nahe dem Mittelwert konzentrieren. Der Begriff bezeichnet auch die Stichprobenvarianz eines Datensatzs; die beiden Verwendungen sind konsistent, da die Stichprobenvarianz die Varianz derjenigen ZV ist, die man durch gleichverteiltes Ziehen eines Datenpunkts aus dem Datensatz erhält. Unter dem quadratischer Verlust ist die Varianz das kleinste Risiko, das eine konstante Vorhersage eines numerischen Labels erreichen kann. Für einen Zufallsvektor ist die Varianz der Erwartungswert der quadrierten euklidische Norm der Abweichung vom Mittelwert. Diese Größe entspricht der Spur der Kovarianzmatrix des Zufallsvektors, d. h. der Summe der Varianzen seiner Einträge.

Definition

Die Wetterstation Krems an der Donau (Österreich) zeichnete vom 20. bis 26. August 2024 jeweils um 15:00 Uhr (UTC) sieben Lufttemperaturen zwischen 24.9 und 33.9 °C auf (GeoSphere Austria, 2026), und jeder Messwert liegt innerhalb von sechs Grad um ihren Durchschnitt von 28.0 °C. Wie weit Messwerte um ihren Durchschnitt streuen, wird durch die Varianz gemessen.

Die Varianz einer reellwertigen Zufallsvariable (ZV) y ist der Erwartungswert $\mathbb{E}\big\{\big(y - \mathbb{E}\{y\}\big)^2\big\}$ der quadrierten Differenz zwischen y und ihrem Erwartungswert $\mathbb{E}\{y\}$, der auch als ihr Mittelwert bezeichnet wird. Der Begriff wird auch verwendet, um die Stichprobenvarianz eines endlichen Datensatzs $\mathcal{D} = \big\{y^{(1)}, \ldots, y^{(m)} \in \mathbb{R}\big\}$ zu bezeichnen. Diese beiden Verwendungen sind konsistent. Ein Datensatz definiert eine diskrete Zufallsvariable $\widetilde{y}^{(\mathcal{D})} = y^{(I)}$ auf dem Ergebnisraum $\{1, \ldots, m\}$. Hierbei wird der Index I gleichverteilt zufällig gewählt, d. h. $\mathbb{P}(I = r) = 1/m$ für alle $r = 1, \ldots, m$, sodass die Wahrscheinlichkeitsverteilung von $\widetilde{y}^{(\mathcal{D})}$ die empirische Verteilung des Datensatzs ist. Die Varianz von $\widetilde{y}^{(\mathcal{D})}$ ist genau die Stichprobenvarianz

$$\frac{1}{m}\sum_{r=1}^{m}\Big(y^{(r)} - \frac{1}{m}\sum_{r'=1}^{m} y^{(r')}\Big)^2,$$

d. h. die durchschnittliche quadrierte Abweichung der Datenpunkte von ihrem Stichprobenmittel. Abb. 1 zeigt die sieben Messwerte aus Krems vom 20. bis 26. August 2024 als Realisierungen einer reellwertigen ZV, zusammen mit ihrem Stichprobenmittel und der Abweichung jeder Messung davon.

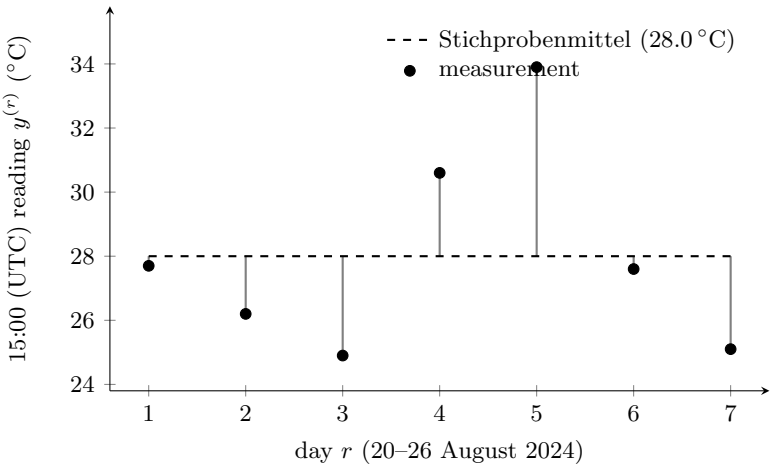


Abbildung 1: Lufttemperaturen $y^{(r)}$, aufgezeichnet um 15:00 Uhr (UTC) an der Wetterstation Krems an der Donau vom 20. bis 26. August 2024 (GeoSphere Austria, 2026) (ausgefüllte Kreise), ihr Stichprobenmittel (gestrichelte Linie) und die Abweichung jeder Messung davon (vertikale Segmente). Die Stichprobenvarianz der Messungen ist der Durchschnitt der quadrierten Abweichungen. Daten erzeugt durch `pythondemos/variance.py`

Für die sieben Temperaturen aus Abb. 1 ist die Stichprobenvarianz die Zahl $63.08/7 \approx 9.01$, die allein aus den Messwerten berechnet wird. Sie schätzt die Varianz der ZV, die sie erzeugt hat; diese Varianz ist eine Eigenschaft ihrer Wahrscheinlichkeitsverteilung und wird nie exakt beobachtet. Die Stichprobenvarianz dient auch als Schwellenwert zur Erkennung von Ausreißern: Eine gängige Regel erklärt einen Datenpunkt zu einem Ausreißer, wenn seine quadrierte Abweichung vom Stichprobenmittel das Neunfache der Stichprobenvarianz, d. h. drei Standardabweichungen, überschreitet (Pukelsheim, 1994, Abschn. 1). Der Messwert von 33.9 °C weicht um 5.9 °C ab, wovon das Quadrat 34.81 etwa das Vierfache der Stichprobenvarianz beträgt, sodass die Regel jeden Messwert beibehält.

Für eine ZV mit endlichem zweiten Moment, d. h. $\mathbb{E}\{y^2\} < \infty$, ist der Mittelwert diejenige Konstante b, die das Risiko $\mathbb{E}\{(y-b)^2\}$ minimiert. Die Varianz ist der kleinste Wert, den dieses Risiko annimmt, d. h. sein Minimum bei $b = \mathbb{E}\{y\}$ (Bertsekas and Tsitsiklis, 2008, Abschn. 8.3):

$$\mathbb{E}\big\{\big(y - \mathbb{E}\{y\}\big)^2\big\} = \min_{b\in\mathbb{R}} \mathbb{E}\big\{(y-b)^2\big\}.$$

Diese Charakterisierung über das Minimum des Risikos lässt sich direkt auf das einfachste Problem des maschinelles Lernen (ML) anwenden: die Vorhersage eines numerischen Labels y ohne Verwendung von Merkmalen. Der Hypothesenraum für dieses Problem besteht aus den konstanten Abbildungen $h(\cdot) = b$ mit $b \in \mathbb{R}$. Jede solche Hypothese läuft auf einen einzigen Bias-Term hinaus: Sie liefert für jeden Datenpunkt dieselbe Vorhersage b. Gegeben ein Trainingsdatensatz $\mathcal{D}^{(\text{train})} = \left\{y^{(1)}, \ldots, y^{(m)}\right\}$, das nur aus Labelsn $y^{(r)} \in \mathbb{R}$ besteht, läuft die empirische Risikominimierung (ERM) mit dem quadratischer Verlust darauf hinaus, diejenige Konstante zu finden, die das Label eines Datenpunkts am besten vorhersagt. Die Lösung $\widehat{b}$ ist der Stichprobenmittel der Labels (siehe Mittelwert), und der Trainingsfehler der gelernten Hypothese ist die Stichprobenvarianz der Labels (Jung, 2022). Für die Temperaturen aus Abb. 1 ist die gelernte Hypothese die gestrichelte Linie: $\widehat{b} = 28.0\,^{\circ}\text{C}$, mit einem Trainingsfehler gleich der oben berechneten Stichprobenvarianz 9.01. Die Varianz des Labels dient daher als Baseline: Eine Methode des ML, die Merkmale verwendet, ist nur dann nützlich, wenn ihr Risiko unter die Varianz des Labels fällt.

Die Varianz lässt sich von einer reellwertigen ZV auf einen Zufallsvektor $\mathbf{x}$ erweitern als der Erwartungswert der quadrierten euklidische Norm der Abweichung vom Mittelwert, $\mathbb{E}\left\{\|\mathbf{x} - \mathbb{E}\{\mathbf{x}\}\|_2^2\right\} = \text{tr}\left(\mathbf{C}^{(\mathbf{x})}\right)$. Die quadrierte euklidische Norm ist die Summe der quadrierten Abweichungen der Einträge von $\mathbf{x}$, sodass ihr Erwartungswert die Summe von deren Varianzen ist. Die Varianzen der Einträge sind die Diagonaleinträge der Kovarianzmatrix $\mathbf{C}^{(\mathbf{x})}$ (Murphy, 2012, Abschn. 2.5.1), und die Summe der Diagonaleinträge ist die Spur.

Synonyme: zweites zentrales Moment.

Siehe auch: ZV, Zufallsvektor, Erwartungswert, Mittelwert, Stichprobenvarianz, Standardabweichung, Wahrscheinlichkeitsverteilung, ERM, quadratischer Verlust, Baseline, Kovarianzmatrix, Spur, Ausreißer.

Querverweise

Literatur

1. Bertsekas and Tsitsiklis (2008). Introduction to Probability. 2nd ed., Athena Scientific, Belmont, MA, USA.

2. 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.

3. Jung (2022). Machine Learning: The Basics. Springer Nature, Singapore, Singapore.

4. Murphy (2012). Machine Learning: A Probabilistic Perspective. MIT Press, Cambridge, MA, USA.

5. Pukelsheim (1994). The Three Sigma Rule. The American Statistician, 48(2), pp. 88–91.