Stichproben-Kovarianzmatrix

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 Stichproben-Kovarianzmatrix eines Datensatzes ist die Kovarianzmatrix des Zufallsvektors, der durch gleichverteilt zufälliges Auswählen eines seiner Datenpunkte entsteht. Ihre Diagonaleinträge sind die Stichprobenvarianzen der einzelnen Merkmale, und ihre Nebendiagonaleinträge sind die Kovarianzen zwischen verschiedenen Merkmalen. Zwei Konventionen für den Divisor sind gebräuchlich: Das Teilen der Summe der äußeren Produkte durch die Anzahl m der Datenpunkte liefert die Kovarianzmatrix unter der empirische Verteilung, während das Teilen durch $m-1$ einen erwartungstreuen Schätzer der Kovarianzmatrix einer zugrundeliegenden Wahrscheinlichkeitsverteilung ergibt. Die beiden unterscheiden sich um den Faktor $m/(m-1)$ und besitzen daher dieselben Eigenvektoren, sodass eine darauf aufbauende Dimensionsreduktion nicht von dieser Wahl abhängt.

Definition

Betrachtet wird ein Datensatz $\mathcal{D}$ aus m Datenpunkten, die durch Merkmalsvektoren $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)} \in \mathbb{R}^d$ charakterisiert sind. Ein fortlaufendes Beispiel ist der Datensatz der 1824 Tage der Jahre 2020 bis 2024 in Krems an der Donau (Österreich), für die beide Messwerte vorliegen, wobei jeder Tag durch seine Lufttemperatur um 06:00 Uhr und um 15:00 Uhr (UTC) charakterisiert ist. Die Stichproben-Kovarianzmatrix erfasst, wie stark jedes Merkmal über die Datenpunkte hinweg variiert und wie stark zwei Merkmale gemeinsam variieren, ohne ein Wahrscheinlichkeitsmodell hinter den Messungen vorauszusetzen.

Die Kovarianzmatrix der Stichprobe $\mathcal{D}$ ist definiert als die Kovarianzmatrix bezüglich der durch $\mathcal{D}$ induzierten empirische Verteilung $\mathbb{P}^{(\mathcal{D})}$. Sie ist explizit gegeben durch

$$\widehat{\mathbf{C}} = \frac{1}{m} \sum_{r=1}^{m} \big(\mathbf{x}^{(r)} - \widehat{\boldsymbol{\mu}}\big)\big(\mathbf{x}^{(r)} - \widehat{\boldsymbol{\mu}}\big)^{\top},$$

wobei $\widehat{\boldsymbol{\mu}}$ den Stichprobenmittel von $\mathcal{D}$ bezeichnet. Äquivalent dazu ist $\widehat{\mathbf{C}}$ die Kovarianzmatrix des Zufallsvektors, der durch gleichverteilt zufälliges Auswählen eines Datenpunkts aus $\mathcal{D}$ entsteht. Für die Wettermessungen betragen die beiden Diagonaleinträge von $\widehat{\mathbf{C}}$ 63.01 und 85.72 Quadratgrad Celsius und beide Nebendiagonaleinträge 65.40 (`pythondemos/samplecovmtx.py`).

Für den Divisor sind zwei Konventionen gebräuchlich. Die obige Definition teilt durch m, wodurch $\widehat{\mathbf{C}}$ zur Kovarianzmatrix eines gleichverteilt zufällig ausgewählten Datenpunkts wird und was die maximale Likelihood-Methode für ein gaußsches Wahrscheinlichkeitsmodell liefert (Bishop, 2006, Sect. 2.3.4). Viele Texte und Softwarebibliotheken teilen stattdessen durch $m - 1$. Unter der Annahme unabhängig und identisch verteilter Daten (u.i.v.-Annahme), bei der die Datenpunkte als Realisierungen von unabhängig und identisch verteilt (u.i.v.) Zufallsvektorenn modelliert werden, ist diese zweite Version ein erwartungstreuer Schätzer der Kovarianzmatrix der zugrundeliegenden Wahrscheinlichkeitsverteilung, während die obige Version im Mittel um den Faktor $(m - 1)/m$ kleiner ist (Bishop, 2006, Sect. 2.3.4). Der Grund dafür ist, dass $\widehat{\boldsymbol{\mu}}$ aus denselben Datenpunkten berechnet wird: Die zentrierten Vektoren summieren sich zum Null-Vektor, sodass nur $m - 1$ davon frei gewählt werden können.

Die beiden Versionen unterscheiden sich um den Faktor $m/(m - 1)$, sodass die Wahl nur bei einem kleinen Datensatz ins Gewicht fällt. Für die 1824 Tage verschiebt dies den ersten Diagonaleintrag von 63.01 auf 63.05 Quadratgrad Celsius. Abb. 1 zeigt, was bei wenigen Datenpunkten geschieht: Gemittelt über viele aus den 1824 Tagen gezogene Datensätze, liefert der Divisor m bei zwei Datenpunkten die Hälfte der Varianz der Temperatur um 06:00 Uhr und liegt bei einhundert immer noch zwei Prozent darunter, während der Divisor $m - 1$ durchgehend genau darauf liegt.

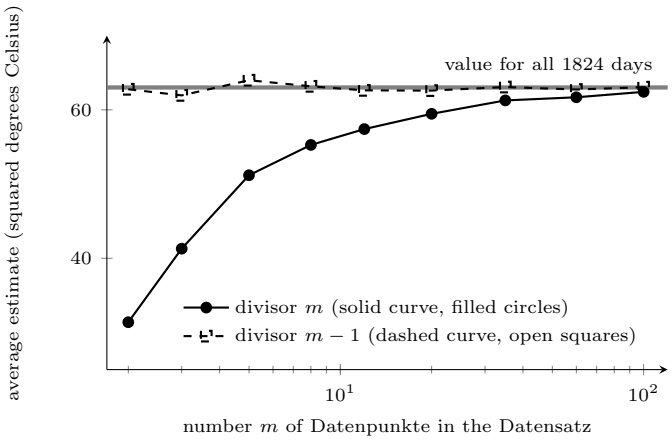

Abbildung 1: Die beiden Divisorkonventionen für die Varianz der Stichprobe der Temperatur um 06:00 Uhr in Krems an der Donau, gemittelt über jeweils 4000 Datensätze jeder Größe, gezogen aus den 1824 Tagen der Jahre 2020 bis 2024. Der Divisor m (durchgezogene Kurve, gefüllte Kreise) nähert sich dem Wert für alle 1824 Tage (dicke graue Linie) von unten; der Divisor $m - 1$ (gestrichelte Kurve, offene Quadrate) trifft ihn bei jeder Größe. Daten erzeugt mit `pythondemos/samplecovmtx.py`

Welche Konvention zu verwenden ist, ergibt sich aus dem Zweck der Matrix. Als Zusammenfassung des vorliegenden Datensatzes ist der Divisor m zu wählen: Er macht $\widehat{\mathbf{C}}$ zur Kovarianzmatrix eines gleichverteilt zufällig ausgewählten Datenpunkts, sodass jede Identität, die für eine Kovarianzmatrix gilt, auch hier gilt. Als Schätzer der Kovarianzmatrix eines Wahrscheinlichkeitsmodells hinter den Datenpunkten beseitigt der Divisor $m-1$ die Verzerrung. Keine der beiden Konventionen liefert den kleinsten erwarteten quadratischen Fehler: Ein Schrumpfen der Einträge in Richtung eines Vielfachen der Einheits-Matrix tauscht eine Verzerrung gegen eine geringere Streuung ein und übertrifft beide, wenn d in der Größenordnung von m liegt (Chen et al., 2010).

Im maschinelles Lernen (ML) ist die Stichproben-Kovarianzmatrix der Merkmalsvektoren eines Trainingsdatensatzes das, was ein Verfahren ausliest, wenn es die Merkmale untereinander statt mit dem Label vergleicht. Die Hauptkomponentenanalyse (PCA) extrahiert die zu den größten Eigenwerten gehörenden Eigenvektoren von $\widehat{\mathbf{C}}$ und verwendet sie als neue Merkmale (siehe Dimensionsreduktion). Die Wahl des Divisors verändert diese Richtungen nicht: Die beiden Versionen unterscheiden sich um einen positiven skalaren Faktor, besitzen also dieselben Eigenvektoren, und ihre Eigenwerte unterscheiden sich um eben diesen Faktor (`pythondemos/samplecovmtx.py`). Bei zentrierten Merkmalen ist $\widehat{\mathbf{C}}$ zudem die aus der Merkmalsmatrix $\mathbf{X}$ gebildete Matrix $(1/m)\mathbf{X}^\top\mathbf{X}$, deren Eigenwerte die Konditionszahl festlegen, welche steuert, wie viele Schritte der Gradientenabstieg für die empirische Risikominimierung (ERM) der lineare Regression benötigt (siehe Konditionszahl).

Siehe auch: Kovarianzmatrix, Kovarianz, Matrix, Stichprobenmittel, Stichprobenvarianz, empirische Verteilung, Varianz, Schätzer, PCA.

Querverweise

Literatur

1. Bishop (2006). Pattern Recognition and Machine Learning. Springer Science+Business Media, New York, NY, USA.

2. Chen et al. (2010). Shrinkage Algorithms for MMSE Covariance Estimation. IEEE Trans. Signal Processing, 58(10), pp. 5016–5029.