Korrelation

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 Korrelation misst die Stärke und Richtung der linearen Beziehung zwischen zwei reellwertigen Zufallsvariablen (ZV). Sie wird erhalten, indem ihre covariance durch ihre standard deviations normiert wird, wodurch sie dimensionslos und über Größen hinweg vergleichbar wird, die auf verschiedenen Skalen gemessen werden. Ihre extremen positiven und negativen Werte charakterisieren perfekte zunehmende und abnehmende affine Beziehungen, während ihr zentraler Wert unkorrelierte Variablen charakterisiert. Unabhängige Zufallsvariablen sind unkorreliert, aber unkorrelierte Zufallsvariablen müssen nicht unabhängig sein, da die Korrelation nur die lineare Abhängigkeit quantifiziert. Die Korrelation zwischen den predictions der Basislerner eines Ensemble begrenzt die Varianz-Reduktion, die durch deren Mittelung erreicht werden kann.

Definition

An einem Tag mit einem ungewöhnlich warmen Morgen in Krems an der Donau (Österreich) ist der Nachmittag tendenziell ebenfalls ungewöhnlich warm. Abb. 1 zeigt einen Datensatz von 1824 täglichen Wetterdaten an dieser Station. Jeder Datenpunkt repräsentiert einen Tag, charakterisiert durch die Lufttemperatur um 06:00 und um 15:00 (UTC) als seinen Merkmalsvektor $\mathbf{x} = (x_1, x_2)^\top$. Die zufällige Auswahl eines Datenpunkt aus dem Datensatz definiert einen random vector $\mathbf{x}$. Die Korrelation quantifiziert, ob ein Tag, dessen Temperatur um 06:00 ungewöhnlich hoch ist, auch tendenziell eine ungewöhnlich hohe Temperatur um 15:00 aufweist.

Allgemeiner ausgedrückt, für zwei reellwertige Zufallsvariablen (ZV) x und y auf einem gemeinsamen probability space mit nicht-null standard deviations σ_x und σ_y ist ihre Korrelation ihre covariance, normiert durch das Produkt dieser standard deviations (Bertsekas and Tsitsiklis, 2008, Sect. 4.2),

$$\rho := \frac{\mathrm{cov}\,(x,y)}{\sigma_x\,\sigma_y} \in [-1,1].$$

Der Wert ρ wird auch als Pearson correlation coefficient bezeichnet.

Im Gegensatz zur covariance ist die Korrelation dimensionslos. Das Hinzufügen einer Konstanten zu einer der Zufallsvariablen oder deren Multiplikation mit einer positiven Konstanten ändert ρ nicht; die Multiplikation einer ZV mit

einer negativen Konstanten kehrt ihr Vorzeichen um (Anderson, 2003, Sect. 2.3, Thm. 2.3.2). Die Extremwerte $\rho = 1$ und $\rho = -1$ treten genau dann auf, wenn y eine zunehmende oder abnehmende affine function von x mit probability eins ist. Tatsächlich ist $|\rho| = 1$ der Gleichheitsfall in der Cauchy-Schwarz inequality für die zentrierten Zufallsvariablen $x - \mathbb{E}\{x\}$ und $y - \mathbb{E}\{y\}$, was genau dann gilt, wenn die eine ein konstantes Vielfaches der anderen mit probability eins ist und das Vorzeichen dieser Konstanten das Vorzeichen von ρ ist. Die Stichprobenkorrelation zweier Datenvektoren hat dieselbe Eigenschaft (Meyer, 2000, Sect. 5.4). Die Zufallsvariablen sind unkorreliert, wenn $\rho = 0$ (Bertsekas and Tsitsiklis, 2008, Sect. 4.2). Unabhängige Zufallsvariablen sind unkorreliert, aber unkorrelierte Zufallsvariablen müssen nicht unabhängig sein, da ρ nur die lineare Abhängigkeit quantifiziert (Bertsekas and Tsitsiklis, 2008, Sect. 4.2). Für den random vector $\mathbf{x}$, der durch die gleichmäßige Zufallsauswahl eines der in Abb. 1 gezeigten täglichen Temperaturpaare erhalten wird, beträgt die Kovarianz zwischen den beiden Temperaturen 65.40 Quadratgrad Celsius, und ihre Varianzen betragen 63.01 und 85.72 Quadratgrad Celsius. Somit

$$\rho = \frac{65.40}{\sqrt{63.01 \cdot 85.72}} \approx 0.89.$$

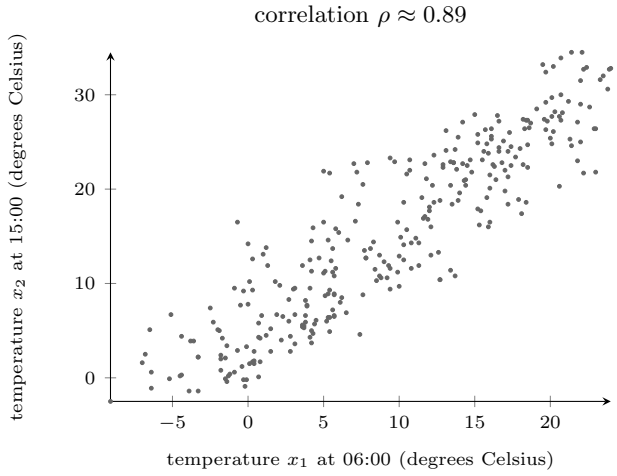


Abbildung 1: Tägliche realizations des Wetter-random vector $\mathbf{x}$: die Temperatur um 06:00 und um 15:00 (UTC) an jedem sechsten der 1824 Tage von 2020 bis 2024, für die beide Messwerte vorliegen. Die Korrelation $\rho \approx 0.89$, berechnet aus allen 1824 Tagen, spiegelt die Tendenz wider, dass ein warmer Morgen einen warmen Nachmittag begleitet. Daten generiert von `pythondemos/correlation.py`

Die Korrelation zwischen den predictions der Basislerner eines Ensemble begrenzt, wie sehr deren Mittelung helfen kann. Ein Mittelwert von M identisch

verteilten predictions, jede mit Varianz σ^2 und paarweiser Korrelation ρ, hat eine Varianz von $\rho\sigma^2 + (1-\rho)\sigma^2/M$ (Hastie et al., 2009, Sect. 15.2). Der zweite Term verschwindet, wenn M wächst, während der erste bestehen bleibt, weshalb Bagging den Trainingsdatensatz neu abtastet und ein Random Forest die Merkmale bei jeder Aufteilung randomisiert: beides senkt die Korrelation zwischen ihren predictions.

Querverweise

Literatur

1. Anderson (2003). An Introduction to Multivariate Statistical Analysis. 3rd ed., Wiley, Hoboken, NJ.

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

3. Meyer (2000). Matrix Analysis and Applied Linear Algebra. SIAM, Philadelphia, PA, USA.

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