correlation

Author and editors

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

Abstract

Correlation measures the strength and direction of the linear relationship between two real-valued random variables (RVs). It is obtained by normalizing their covariance by their standard deviations, making it dimensionless and comparable across quantities measured on different scales. Its extreme positive and negative values characterize perfect increasing and decreasing affine relationships, while its central value characterizes uncorrelated variables. Independent RVs are uncorrelated, but uncorrelated RVs need not be independent, since correlation quantifies only linear dependence. The correlation between the predictions of the base learners of an ensemble bounds the variance reduction that averaging them can achieve.

Definition

Fig. 1 shows a dataset that consists of 1827 daily weather measurements at Krems an der Donau. Each data point represents a day, characterized by the minimum and the maximum temperature as its feature vector $\mathbf{x} = (x_1, x_2)^\top$. Selecting a data point from the dataset uniformly at random defines a random vector $\mathbf{x}$. Correlation quantifies whether a day whose minimum temperature is unusually high also tends to have an unusually high maximum temperature.

More generally, for two real-valued random variables (RVs) x and y on a common probability space with nonzero standard deviations σ_x and σ_y, their correlation is their covariance normalized by the product of these standard deviations (Bertsekas and Tsitsiklis, 2008),

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

The value ρ is also referred to as the Pearson correlation coefficient.

Unlike the covariance, correlation is dimensionless. Adding a constant to either RV, or multiplying it by a positive constant, does not change ρ; multiplying one RV by a negative constant reverses its sign (Anderson, 2003, Sec. 2.3, Thm. 2.3.2). The extreme values $\rho = 1$ and $\rho = -1$ occur exactly when y is an increasing or a decreasing affine function of x with probability one (Meyer, 2000, Sec. 5.4). The RVs are uncorrelated when $\rho = 0$ (Bertsekas and Tsitsiklis, 2008, Sec. 4.2). Independent RVs are uncorrelated, but uncorrelated RVs need

not be independent, since ρ quantifies only linear dependence (Bertsekas and Tsitsiklis, 2008, Sec. 4.2). For the random vector $\mathbf{x}$ obtained by selecting uniformly at random one of the daily minimum-and-maximum temperature pairs shown in Fig. 1, the covariance between the two temperatures is 55.08 squared degrees Celsius, and their variances are 49.89 and 83.41 squared degrees Celsius. Thus,

$$\rho = \frac{55.08}{\sqrt{49.89 \cdot 83.41}} \approx 0.85.$$

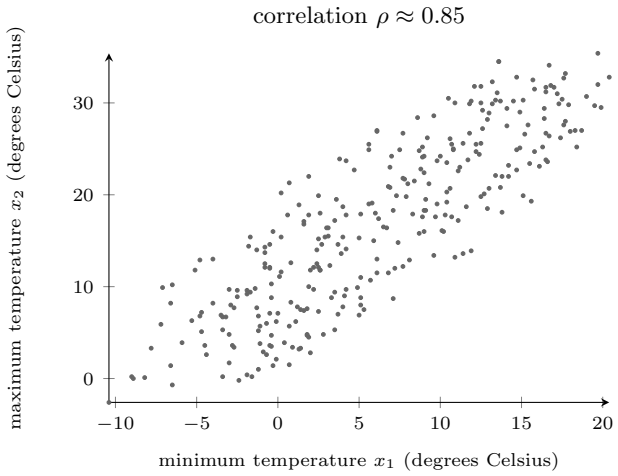


Figure 1: Daily realizations of the weather random vector $\mathbf{x}$: every sixth day from 2020 to 2024. The correlation $\rho \approx 0.85$, computed from all 1827 days, reflects the tendency of high daily minima to accompany high daily maxima. Data generated by `pythondemos/correlation.py`

The correlation between the predictions of the base learners of an ensemble limits how much averaging them can help. An average of M identically distributed predictions, each with variance σ^2 and pairwise correlation ρ, has variance $\rho\sigma^2 + (1-\rho)\sigma^2/M$ (Hastie et al., 2009, Sect. 15.2). The second term vanishes as M grows, while the first remains, which is why bagging resamples the training set and a random forest randomizes the features at each split: both lower the correlation between their predictions.

Cross References

References

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.