Thailand Statistician
2015; 13(1): 125-126
http://statassoc.or.th
Short communication

One More Geometric Interpretation of Pearson's Correlation

Ildar Batyrshin $^*[a]$ **and Vladik Kreinovich** $[b]$

$[a]$ Centro de Investigación en Computación (CIC)
Instituto Politécnico Nacional (IPN), México, D.F.

$[b]$ Department of Computer Science
University of Texas at El Paso, El Paso, TX 79968, USA

* corresponding author; e-mail: batyr1@gmail.com

Received: 12 August 2014
Accepted: 7 October 2014

Abstract

It is known that Pearson's correlation coefficient r is equal to the cosine of the angle between the vectors $\vec{x}$ and $\vec{y}$ describing the centered data. We use this known fact to show that r describes how closer the corresponding unit vector $\vec{e}_x = \dfrac{\vec{x}}{\|\vec{x}\|}$ is to the y-unit vector $\vec{e}_y = \dfrac{\vec{y}}{\|\vec{y}\|}$ than to its opposite $-\vec{e}_y$: namely, r is equal to the (scaled) difference between the squared distances $d^2(\vec{e}_x, -\vec{e}_y)$ and $d^2(\vec{e}_x, \vec{e}_y)$. In particular, the correlation is positive if $\vec{e}_x$ is closer to $\vec{e}_y$ and negative if $\vec{e}_x$ is closer to $-\vec{e}_y$.

Known geometric interpretation: correlation as the cosine of the angle between two centered data vectors. It is known that the sample Pearson's correlation r is equal to $\cos(\alpha)$, where $\alpha \stackrel{\text{def}}{=} \angle(\vec{x}, \vec{y})$ is the angle between the centered data vectors $\vec{x} \stackrel{\text{def}}{=} (X_1 - \overline{X}, \ldots, X_n - \overline{X})$ and $\vec{y} \stackrel{\text{def}}{=} (Y_1 - \overline{Y}, \ldots, Y_n - \overline{Y})$; see, e.g., Rodgers and Nicewander [1].

Specifically, in terms of these vectors $\vec{x}$ and $\vec{y}$, the correlation

$$r = \frac{\sum\limits_{i=1}^{n} (X_i - \overline{X}) \cdot (Y_i - \overline{Y})}{\sqrt{\sum\limits_{i=1}^{n} (X_i - \overline{X})^2} \cdot \sqrt{\sum\limits_{i=1}^{n} (Y_i - \overline{Y})^2}}$$

can be described as $r = \dfrac{\vec{x} \cdot \vec{y}}{\|\vec{x}\| \cdot \|\vec{y}\|}$. Since for every two vectors $\vec{a}$ and $\vec{b}$, the scalar (dot) product is equal to $\vec{a} \cdot \vec{b} = \|\vec{a}\| \cdot \|\vec{b}\| \cdot \cos(\angle(\vec{a}, \vec{b}))$, the vector-based expression for r leads to the desired formula $r = \cos(\alpha)$.

Towards a new interpretation. The angle between the two vectors $\vec{x}$ and $\vec{y}$ is the same as the angle between the corresponding unit vectors $\vec{e}_x \stackrel{\text{def}}{=} \dfrac{\vec{x}}{\|\vec{x}\|}$ and $\vec{e}_y \stackrel{\text{def}}{=} \dfrac{\vec{y}}{\|\vec{y}\|}$. Thus, for these unit vectors, $\vec{e}_x \cdot \vec{e}_y = \|\vec{e}_x\| \cdot \|\vec{e}_y\| \cdot \cos(\alpha) = \cos(\alpha)$.

Therefore,

$$d^2(\vec{e}_x, \vec{e}_y) = \|\vec{e}_x - \vec{e}_y\|^2 = (\vec{e}_x)^2 + (\vec{e}_y)^2 - 2\vec{e}_x \cdot \vec{e}_y = 2 - 2\cos(\alpha) = 2 - 2r$$

and

$$d^2(\vec{e}_x, -\vec{e}_y) = \|\vec{e}_x - (-\vec{e}_y)\|^2 = (\vec{e}_x)^2 + (\vec{e}_y)^2 + 2\vec{e}_x \cdot \vec{e}_y = 2 + 2\cos(\alpha) = 2 + 2r.$$

So, we conclude that

$$r = \frac{1}{4} \cdot (d^2(\vec{e}_x, -\vec{e}_y) - d^2(\vec{e}_x, \vec{e}_y)).$$

In other words, the correlation coefficient describes whether the unit centered data vector $\vec{e}_x$ corresponding to X_i is closer to the similar vector $\vec{e}_y$ corresponding to Y_i or to the opposite vector $-\vec{e}_y$. In particular, the correlation is positive if $\vec{e}_x$ is closer to $\vec{e}_y$ and negative if $\vec{e}_x$ is closer to $-\vec{e}_y$.

Comment. The above formula for r is a particular case of a more general formula presented in Batyrshin [2].

References

[1] Rodgers, J. L., and Nicewander, W. A., Thirteen ways to look at the correlation coefficient, *The American Statistician*, 1988; 42(1): 59-66.

[2] Batyrshin, I., Constructing time series shape association measures: Minkowski distance and data standardization, Proceedings of the 1st BRICS Countries Congress on Computational Intelligence BRIC-SCCI 2013, Recife, Brazil. September 811, 2013. http://arxiv.org/pdf/1311.1958v3.