Geometric interpretation of a correlation

Zenon Gniazdowski[*]

Abstract

The study shows that the Pearson's coefficient of correlation is equivalent to the cosine of the angle between random variables. It was found that the information about the intensity of the relationship between variables is included in the value of the angle between random vectors. The paper proposes intuitive criteria for measuring the intensity of the relationship between random variables.

Keywords: *data mining, correlation coefficient, cosine, angle between vectors, stochastic dependence*

1 Preliminaries

Data mining operates on different types of data, that require the use of appropriate methods of analysis. In this article some of the statistics used in the analysis of continuous data will be interpreted.

Since the article is an attempt to present a geometrical interpretation of some statistics, the basic geometrical definitions, such as the Euclidean norm and scalar product, in order to be able to find the angle between the vectors, will be presented at the beginning. Additionally, basic statistics such as mean, variance, standard deviation, as well as measures of dependence of two random variables, such as covariance and correlation will also be presented. At the end, the basic operations on random variables such as the reduction of the constant component and standardization will be offered.

[*] Warsaw School of Computer Science.

1.1 The angle between vectors

The space with Euclidean norm and scalar product is considered. In n-dimensional space vector $a = (a_1, a_2, \dots a_n)$ is considered. The Euclidean norm $\|a\|$ of this vector is given by the formula [1]:

$$\|a\| = \sqrt{\sum_{i=1}^{n} a_i^2}. \tag{1}$$

In three-dimensional space or on a plane, Euclidean norm of vector is its length. The scalar product (dot product) of vector $a = [a_1, a_2, \dots, a_n]$ and vector $b = [b_1, b_2, \dots, b_n]$ is equal to [1]:

$$a \cdot b = \sum_{i=1}^{n} a_i b_i. \tag{2}$$

Simultaneously, the dot product of two vectors can be represented as follows:

$$a \cdot b = \|a\| \cdot \|b\| \cdot cos(a, b). \tag{3}$$

In the expression (3) $cos(a, b)$ is the cosine of the angle between two vectors:

$$cos(a, b) = \frac{a \cdot b}{\|a\| \cdot \|b\|}. \tag{4}$$

Hence, the angle between vectors can be calculated using the arccosine function.

1.2 Auxiliary Statistics

Random variable *X* is considered. In particular, there is a random sample of size *n*. Element X_i represents the *i-th* realization of the random variable. On the basis of the sample, the expected value of a random variable X can be estimated [2] [3]:

$$\mu = E(X) = \sum_{i} X_i p_i. \tag{5}$$

In the expression (5) p_i is the probability of the *i-th* event. The average value of *X* is the estimator of the expected value of the random variable *X*:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{6}$$

The measure of the scatter of a random variable is its variance [2] [3]:

$$\sigma^2(X) = E[(X - \mu)^2] = \sum_{i} (X_i - \mu)^2 p_i. \tag{7}$$

Square root of variance is called the standard deviation [3]:

$$\sigma = \sqrt{\sigma^2(X)}. \tag{8}$$

Estimator of variance calculated using the n-element sample has a form [3]:

$$s^2 = \frac{1}{l}\sum_{i=1}^{n}(X_i - \bar{x})^2. \tag{9}$$

Depending on the type of the estimator value of *l* can take one of two values [3]: for maximum likelihood estimator *l* = *n*, for unbiased estimator *l* = *n-1*.

1.3 Reduction and standardization of the random variable

If the random sample was obtained from a symmetric distribution, its average value approximates its “typical” value. For example, if the nominal size of a particular element is equal to X_0, then this value can be identified as the average of multiple measurements. To evaluate the dispersion of the random variable, its average value should be subtracted from it:

$$x_i = X_i - \bar{x}. \tag{10}$$

It is a random variable reduced by a constant component. If the variable *X* is derived from a normal distribution with mean value μ and standard deviation σ, it can be standardized by making the following transformation [3]:

$$x = \frac{X - \mu}{\sigma}. \tag{11}$$

Variable x has an average value $\mu = 0$ and standard deviation $\sigma = 1$.

2 Geometric interpretation of the Pearson’s correlation coefficient

A measure of the relationship between two random variables is the covariance [3]:

$$\sigma_{X,Y} = E[(X - \mu_X)(Y - \mu_Y)]. \tag{12}$$

Covariance normalized to unity is called the correlation coefficient [3]:

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sqrt{E[(X - \mu_X)^2]E[(Y - \mu_Y)^2]}}. \tag{13}$$

Expression (13) can be further converted to the form:

$$\rho_{X,Y} = \frac{\sum_{i=1}^{n}[(X_i - \bar{x})(Y_i - \bar{y})]}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{y})^2}}. \tag{14}$$

The correlation coefficient between two variables is equal to the covariance of variables subject to standardization. Using equation (10), formula (14) can be converted to the form:

$$\rho_{X,Y} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}. \tag{15}$$

The resulting expression is the ratio of two elements. The numerator is the scalar product of two vectors, while the denominator is the product of its lengths:

$$\rho_{X,Y} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} = \frac{x \cdot y}{\|x\| \cdot \|x\|} = cos(x,y). \tag{16}$$

Expression (16) shows the formal identity between the correlation coefficient, and the cosine of the angle between two random vectors.

3 Coefficient of determination

Model $\hat{Y}$ of dependent variable Y is created in the regression analysis. The Pearson's correlation coefficient is used to evaluate this model [4]:

$$R = \frac{\sum_{i=1}^{n}\left[(Y_i - \bar{Y})\left(\hat{Y}_i - \bar{Y}\right)\right]}{\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}\sqrt{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2}}. \tag{17}$$

Here, the coefficient belongs to the interval [0,1]. Equation (17) is analogous to formula (14), and thus represents the cosine of the angle between two vectors. One of them represents dispersion of the vector Y, and the second one represents the dispersion of model $\hat{Y}$. Expression (17) can also be represented in the equivalent form [4]:

$$R = \frac{\sqrt{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2}}{\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}. \tag{18}$$

This expression represents a ratio of the length of two vectors. Formally, the model $\hat{Y}$ is presented as a result of orthogonal projection of the vector Y on a hyperplane [2] [4]. The cosine of the angle between the projected vector and its projection is equal to the ratio of the length of the second vector to the length of the first vector. This fact expresses the formula (18).

In general, the square of the correlation coefficient (16) or (17) is called the coefficient of determination. If the first variable (x or $\hat{Y}$) is a model which explains the behavior of the second variable (respectively y or Y), the ratio of the variance of this first variable to the variance of the second variable is called the coefficient of determination [2] [5]:

$$R^2 = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}. \tag{19}$$

On the other hand, the root of determination coefficient is a ratio of the standard deviations of both variables. Coefficient of determination indicates, how the variance of the model explains the variance of modeled variable [5].

Table 1 presents the cosines of different angles (different correlation coefficients) and the corresponding coefficients of determination expressed as a percentage. Two random vectors are (almost) orthogonal, if the cosine of the angle between them (also determination coefficient) is (almost) equal to zero. This means that the random variables represented by these vectors are independent or random vectors are (near) orthogonal.

Table 1. The cosine of the angle against determination coefficient

Angle [degrees]	Angle [rad]	The cosine of the angle (correlation ρ)	Determination coefficient (ρ^2)	Explained percentage of the variance
0	0	1	1	100.00
15	$\frac{\pi}{12}$	$\frac{\sqrt{6}+\sqrt{2}}{4}$	$\frac{2+\sqrt{3}}{4}$	93.30
30	$\frac{\pi}{6}$	$\frac{\sqrt{3}}{2}$	0.75	75.00
45	$\frac{\pi}{4}$	$\frac{\sqrt{2}}{2}$	0.5	50.00
60	$\frac{\pi}{3}$	0.5	0.25	25.00
75	$\frac{5}{12}\pi$	$\frac{\sqrt{6}-\sqrt{2}}{4}$	$\frac{2-\sqrt{3}}{4}$	6.70
90	$\frac{\pi}{2}$	0	0	0.00

Similarly, if the cosine of the angle between the vectors is (almost) equal to one (determination coefficient close to unity), the vectors are (almost) parallel. Random variables represented by these vectors are highly correlated. One variable can explain most of the variance of the second variable.

4 The significance of the correlation

The value of the correlation coefficient is a random variable and its significance is a function of the number of observations. If the resulting value of the correlation coefficient is $|\varrho| = 0.7$ for a large number of observations, it is more reliable than the same coefficient but obtained for a small number of observations [4]. To assess the reliability of the correlation coefficient, the corresponding hypothesis is tested. The idea is based on the rejection of null hypothesis *H₀*, if the result is highly unlike, under the assumption that the hypothesis is true [6] [7]. The Student's test is used to test the significance of the correlation coefficient. For this purpose the following function is examined:

$$t = \frac{\rho\sqrt{N-2}}{\sqrt{1-\rho^2}}. \tag{20}$$

Function *t* is a random variable with the Student's t-distribution with $N-2$ degrees of freedom. For zero *ϱ*, the *t-statistic* is zero. However, for $\rho = \pm 1$, the value of t tends to $\pm\infty$.

The null hypothesis is established. It states that the correlation coefficient is equal to zero. If the hypothesis is false, then obtaining the correlation coefficient ρ_x (in absolute value) greater than the correlation coefficient ρ is highly unlike. With t and N can be calculated the probability of obtaining a higher value $|\rho_x|$ than the observed value $|\rho|$:

$$P(|\rho_x| > |\rho|) = P(|t_x| > |t|) = p. \tag{21}$$

Value of p can be calculated by integrating the function of Student-t distribution [8]. Counted value of the parameter *t* is used as a limit of integration:

$$p = \int_{-\infty}^{t} f(x)dx + \int_{t}^{+\infty} f(x)dx = 1 - 2\int_{0}^{t} f(x)dx. \tag{22}$$

If the resulting probability is lower than a certain level of significance, it is not possible to accept the hypothesis of no-correlation between the variables [4]. This hypothesis is rejected and it is assumed that there is a correlation. If the resulting value of probability is greater than α, the null hypothesis cannot be rejected. This means that nothing can be said about correlation.

5 The intensity of correlation

Statistics (20) that is used to assess the significance of the correlation depends on two factors. One of them is the value of the correlation coefficient. The second one is the number of degrees of freedom associated with the sample size. When the sample is large, it is easy to demonstrate statistical significance of a weak relationship. With a large or very large sample size, rejection of the false null hypothesis is almost always possible [9]. Rejection of the null hypothesis indicates that there is a significant correlation between two variables. If the null hypothesis is not rejected, it is unknown whether such relationship exists. On the other hand, information that the relationship is significant says very little. The relationship may be statistically significant, and may not be significant in other ways. Statistical test of significance only states that the correlation is nonzero [9]. The test of significance does not contribute to the assessment of intensity of the relationship. Perhaps rather give deceptive information. Thus, occurs a problem, how to measure the intensity of the relationship.

The value of the correlation coefficient, treated as the cosine of the angle between random vectors, contains information about the level of dependence of the variables. The cosine close to zero means that the vectors are (almost) orthogonal, so the random variables are independent. If the cosine is close to one or minus one, the vectors are (almost) parallel and random variables are strongly correlated.

Figure 1 shows the directions of several vectors. As a reference vector, horizontal axis in the positive direction is considered. Analysis of the figure shows that the most comprehensive and intuitive information about intensity of correlation there is in the size of the angle. If the vectors are orthogonal then variables represented by them are independent. In the range of

45 to 135 degrees, the vectors are closer to the orthogonality than collinearity. The angles of 45 degrees and 135 degrees are the limit angles. For these angles, the vector is equally far from the orthogonality and parallelism. The coefficient of determination is equal to 50%. Exactly half of the variation in one variable can be explained by the second variable.

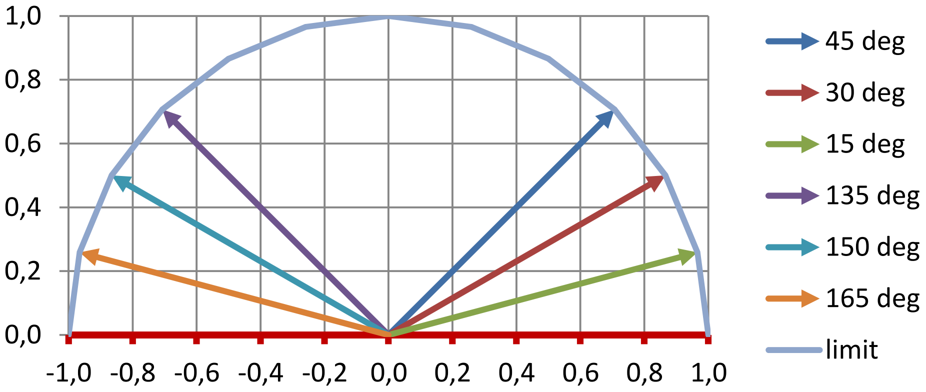

Figure 1. The directions of vectors with respect to the reference vector lying along the horizontal axis

For angles less than 45 degrees or greater than 135 degrees, vectors are closer to parallel than perpendicular – it can be assumed that the random variables are dependent. Vectors are close to parallel when they lie at an angle less than 30 degrees and greater than 150 degrees, with respect to the reference vector. If the angle is lesser than 15 degrees or greater than 165 degrees, variables are strongly correlated (Table 2).

Table 2. The level of intensity of the correlation

	Extra large	**Large**	**Weak or no**	**Large**	**Extra large**
The correlation coefficient (cosine)	≥ 0.97	≥ 0.87	$[-0.77,0.77]$	≤ -0.87	≤ -0.97
Determination coefficient [%]	≥ 93.3	≥ 75.0	$[-50,50]$	≤ -75.0	≤ -93.3
Angle [degrees]	$\leq 15^0$	$\leq 30^0$	$[45^0, 135^0]$	$\geq 150^0$	$\geq 165^0$

6 Summary

The paper presents the possibility of geometrical interpretation of the correlation. It is noted that the correlation coefficient is formally equivalent to the cosine of angle between random vectors. The variables are dependent when the vectors are almost parallel. The variables are independent, when the vectors are nearly orthogonal. Thus, independent random variables are orthogonal.

The paper also discussed the significance tests of correlation. It was found that the significance test does not provide information about the intensity of correlation. Information about the intensity of the correlation is indirectly given by the value of the correlation coefficient. Immediately it can be found in the size of the angle between random vectors. The paper proposes practical angles and the corresponding correlation coefficients that determine the intensity of correlation.

References

[1] Krejn S.G., (red.), *Analiza funkcjonalna*, PWN, Warszawa 1967

[2] Koronacki J., Mielniczuk J., *Statystyka dla studentów kierunków technicznych i przyrodniczych*, WNT, Warszawa 2006

[3] Durka P.J., *Wstęp do współczesnej statystyki*, Wydawnictwo Adamantan, Warszawa 2003

[4] Mańczak K., *Metody identyfikacji wielowymiarowych obiektów sterowania*, WNT, Warszawa 1971

[5] Sobczyk M., *Statystyka*, Wyd. Nauk. PWN, Warszawa 2005

[6] Kamys B., [online], available: http://users.uj.edu.pl/~ufkamys/BK/bb_zakladki.pdf [Date of access: May 9, 2013]

[7] Francuz P., Mackiewicz R., *Liczby nie wiedzą skąd pochodzą. Przewodnik po metodologii i statystyce*, Wydawnictwo KUL, Lublin 2005

[8] Martin F.F., *Computer Modeling and Simulation*, Wiley, New York 1968

[9] Blalock H.M., *Social Statistics*, McGraw-Hill Book Co., New York 1960

Geometryczna interpretacja korelacji

Streszczenie

W pracy pokazano, że współczynnik korelacji Pearsona jest równoważny cosinusowi kąta między wektorami losowymi. Stwierdzono, że informacja o sile związku między zmiennymi zawarta jest w wielkości kąta między wektorami losowymi. W pracy zaproponowano intuicyjne kryteria pomiaru siły związku pomiędzy zmiennymi losowymi.

Słowa kluczowe: *eksploracja danych, współczynnik korelacji, cosinus, kąt między wektorami, zależność stochastyczna*