Basic Correlation Analysis

Description

Correlation analysis is a basic statistical approach that can capture relationships among pairs of variables. Such variables usually represent properties of objects whose values may be stored in columns of a table. For example, in a table containing customers, we may want to measure the correlation between the variables "Age" and "Income" or between the variables "Income" and "TotalPurchase", and so on.

Correlations among variables can be negative or positive. There are different ways of computing correlations, but in most cases, the correlation is measured as a coefficient ranging from -1 to 1. A value close to 0 in this range indicates a lack of correlation. Values closer to the boundries -1 or 1 indicate strong negative or positive correlations, respectively. Usually (though not exclusively) strong positive or negative correlations may indicate a causal relationship between the variables. For example, there may be a positive correlation between the number of hours of studying for an exam and the score obtained in that exam.

Computing the Correlation Coefficient

Given two variables x and y (e.g., "Age" and "Income"), the correlation coefficient for x and y is given by the ratio of the covariance of x and y to the product of standard deviations of x and y. In other words:

$$corr(x,y) = \frac{\mathrm{cov}(x,y)}{stdev(x) \cdot stdev(y)}$$

The covariance cov(*x*, *y*) is the average of the products of deviations from the mean in each of *x* and *y*:

$$\mathrm{cov}(x,y) = \frac{1}{n}\sum_i (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

In the above, each x_i represents the value of *x* in the *i*th row, $\bar{x}$ is the mean value of *x*, and *n* is the total number of values (rows).

Example

As an example consider the following table with 3 attributes *X*, *Y*, and *Z*.

	X	Y	Z
1	3	9	200
2	2	7	180
3	4	12	175
4	5	14	123
5	6	16	120

Performing the above correlation computation among all pairs of variables, would result in the following *correlation matrix*:

	X	*Y*	*Z*
X	1.00		
Y	1.00	1.00	
Z	-0.86	-0.87	1.00

Note that the diagonal is always 1 because each variable is perfectly correlated with itself. In this case, there is also a perfect correlation between *X* and *Y* (after rounding). However, *Z* shows a strong negative correlation to both *X* and *Y*.