Linear Correlation

Updated: July 23, 2015

Calculates the linear correlation between column values in a dataset

Category: Statistical Functions (https://msdn.microsoft.com/en-us/library/azure/dn905867.aspx)

Module Overview

You can use the **Linear Correlation** module to compute a set of Pearson correlation coefficients for each possible pair of variables in the input dataset.

The Pearson correlation coefficient, sometimes called Pearson’s R test, is a statistical value that measures the linear relationship between two variables. By examining the coefficient values, you can infer something about the strength of the relationship between the two variables, and whether they are positively correlated or negatively correlated.

How to Use Linear Correlation

There are no parameters to set for **Linear Correlation**. However, some other restrictions apply:

- The **Linear Correlation** module can process only numeric values. All other types of values, including missing values, non-numeric values, and categorical values, are treated as NaNs.

  To avoid creating a lot of unnecessary columns of NaNs, we recommend that you use Project Columns (https://msdn.microsoft.com/en-us/library/azure/dn905883.aspx) to pass in only the columns for which you want to compute coefficients.

- Pearson’s correlation is calculated for all numeric columns in the dataset that are passed as input.

- **Linear Correlation** is intended to be used with data that has no missing values. Before running this module, use Clean Missing Data (https://msdn.microsoft.com/en-us/library/azure/dn906028.aspx) or other methods to impute missing values.

- Be sure to replace placeholders with other appropriate values before using this module.

  For example, if NaNs were inserted for missing values when the dataset was loaded from the source, it could cause an error. If a placeholder (such as 999 or -1) was used, it could cause bad results.

- The correlation coefficient should not be calculated if the relationship is not linear. You can visually assess the linearity of two variables by using a scatter plot, or you can calculate a regression equation for the two variables.

Results

Given two feature columns, the **Linear Correlation** module returns the scalar Pearson product moment (sample) correlation coefficient.

Given a matrix, the **Linear Correlation** module returns a set of Pearson product moment correlations between each pair of feature columns.

The Pearson correlation coefficient (often denoted as *r*) ranges in value from *+1* to *-1*, where *+1* indicates a strong linear relationship, and *-1* indicates no linear relationship between the two variables.

The interpretation of the coefficients depends very much on the problem you are modeling and the variables you are studying.

- For example, if you are certain the variables are unrelated and yet the Pearson's correlation coefficient is positive, you should investigate further.
- Conversely, if you use linear correlation on two variables that you know to be perfectly correlated, and the coefficient values are not what you expect, it might indicate a problem in the data.

Thus it is important to understand the context of the data when reporting and interpreting Pearson's correlation coefficient.

Examples

To see how this module is used in machine learning experiments, see these sample experiments in the Model Gallery (http://gallery.azureml.net/):

- In the Data Processing and Analysis (http://go.microsoft.com/fwlink/?LinkId=525733) sample experiment, **Linear Correlation** is used to identify potential feature columns.

Technical Notes

If the column that is passed as input contains scalars, the input arrays (*x* and *y*) are treated as vectors and the Pearson product moment correlation is computed as follows:

$$r = \frac{\sum_{k=1}^{n} \left(x_k - \mu_y\right)\left(y_k - \mu_y\right)}{\sqrt{\sum_{k=1}^{n} \left(x_k - \text{ìì}\ _x\right)^2 \sum_{k=1}^{n} \left(y_k - \mu_y\right)^2}}$$

In this formula, each array contains *n* elements and the means of the *x* and *y* samples are μ_x and μ_y respectively.

For a matrix, a matrix of data (*X*) is input, in which each column represents a vector of values. The data matrix should be *n*-by-*m*. The output is the *m*-by-*m* matrix, *R* as defined by

$$\mathrm{R}_{i,j} = \frac{\sum_{k=1}^{n} \left(X_{k,i} - \mu_{x_i}\right)\left(X_{k,j} - \mu_{x_j}\right)}{\sqrt{\sum_{k=1}^{n} \left(X_{k,i} - \mu_{x_i}\right)^2 \sum_{k=1}^{n} \left(X_{k,j} - \mu_{x_j}\right)^2}}$$

In this formula, μ_x represents the mean value of the column x_i. The elements at $_{\mathrm{I,j}}$ always equal 1, as they represent the correlation of a vector with itself.

Expected Input

Name	Type	Description
Dataset	Data Table (https://msdn.microsoft.com/en-us/library/azure/dn905851.aspx)	Input dataset

Output

Name	Type	Description
Results dataset	Data Table (https://msdn.microsoft.com/en-us/library/azure/dn905851.aspx)	Correlations matrix

Exceptions

For a complete list of error messages, see Machine Learning Module Error Codes (https://msdn.microsoft.com/en-us/library/azure/dn905910.aspx).

Exception	Description
Error 0003 (https://msdn.microsoft.com/en-us/library/azure/dn906003.aspx)	Exception occurs if one or more of inputs are null or empty.
Error 0020 (https://msdn.microsoft.com/en-us/library/azure/dn906040.aspx)	Exception occurs if the number of columns in some of the datasets passed to the module is too small.
Error 0021 (https://msdn.microsoft.com/en-us/library/azure/dn905802.aspx)	Exception occurs if the number of rows in some of the datasets passed to the module is too small.

See Also

Statistical Functions (https://msdn.microsoft.com/en-us/library/azure/dn905867.aspx)
A-Z List of Machine Learning Studio Modules (https://msdn.microsoft.com/en-us/library/azure/dn906033.aspx)