Machine Learning to build Intelligent Systems

Manas Dasgupta

Correlation-Covariance Matrix Comparison

Covariance vs Correlation

- "**Covariance**" indicates the **direction** of the linear relationship between variables.
- "**Correlation**" on the other hand measures both the **strength and direction** of the linear relationship between two variables.
- Correlation is a function of the covariance.

What sets them apart is the fact that correlation values are standardized whereas, covariance values are not.

You can obtain the correlation coefficient of two variables by dividing the covariance of these variables by the product of the standard deviations of the same values.

If we revisit the definition of Standard Deviation, it essentially measures the absolute variability of a Dataset's distribution.

When you divide the covariance values by the standard deviation, it essentially scales the value down to a limited range of **-1 to +1**. This is precisely the range of the correlation values.

Formulas

COVARIANCE

$$cov_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

$cov_{x,y}$ = covariance between variable a and y

x_i = data value of x

y_i = data value of y

$\bar{x}$ = mean of x

$\bar{y}$ = mean of y

N = number of data values

CORRELATION

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient

x_i = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

y_i = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

APPLICATIONS

COVARIANCE

- PCA or Principal Component Analysis is a Feature Extraction method that uses Eigenvectors and Eigenvalues. In obtaining the Eigenvectors, covariance matrix is used after the variable are brought on similar scales.

CORRELATION

- Correlation is a vital tool for feature selection and multivariate analysis in data pre-processing and exploration. Correlation helps us investigate and establish relationships between variables. This is employed in feature selection before any kind of statistical modelling or data analysis.

Hope you have liked this Video.
Please help us by providing your Ratings and Comments for this Course!

Thank You!!
Manas Dasgupta

Happy Learning!!