Covariance

Covariance is used to determine how much two random variables vary together.

"Covariance" indicates the direction of the linear relationship between variables.

cov(X, Y) = (sum (x - mean(X)) * (y - mean(Y))) * 1/(n-1)

Correlation

In simple terms, correlation is a measure of how strongly one variable depends on another.

Correlation is used to determine when a change in one variable can result in a change in another.

- Positive Correlation: both variables change in the same direction.
- Neutral Correlation: No relationship in the change of the variables.
- Negative Correlation: variables change in opposite directions.

"Correlation" on the other hand measures both the strength and direction of the linear relationship between two variables.

Correlation is a normalized form of covariance and not affected by scale. Both covariance and correlation measure the linear relationship between variables but cannot be used interchangeably.

Pearson's correlation coefficient = covariance(X, Y) / (stdv(X) * stdv(Y))

Multicollinearity

The performance of some algorithms can deteriorate if two or more variables are tightly related, called multicollinearity.

Cosine Similarity

Cosine Similarity measures the cosine of the angle between two non-zero vectors

$$A.B = |A||B| \cos\theta$$