Week 6

Data Mining and Knowledge Discovery

Dr John Evans
j.evans8@herts.ac.uk

University of Hertfordshire UH

Plan for today

Dissimilarity of Numeric Data

Proximity measures for Ordinal Attributes

Correlation

Recap

Last week:

- We saw discretisation and entropy.
- We saw measures of similarity and dissimilarity.
- In particular, we saw the SMC and Jaccard Coefficient.

Dissimilarity of numeric data

The most popular distance measure is the *Euclidean distance* measure, which is a simple generalisation of the well-known Pythagorean Theorem.

Dissimilarity of numeric data

The most popular distance measure is the *Euclidean distance* measure, which is a simple generalisation of the well-known Pythagorean Theorem.

It can also be thought of as 'straight line' distance or 'as the crow flies'. Suppose we are in $\mathbb{R}^n$ (that is, we are representing elements as n-dimensional vectors). Then we can write,

$$\begin{aligned} x_i &= \begin{pmatrix} x_{i1} & x_{i2} & \cdots & x_{in} \end{pmatrix}^T \\ x_j &= \begin{pmatrix} x_{j1} & x_{j2} & \cdots & x_{jn} \end{pmatrix}^T. \end{aligned}$$

The Euclidean distance between the objects x_i, x_j can then be defined as,

$$d(i, j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{in} - x_{jn})^2}.$$

Minkowski distance

Recall: We can write square roots using indices as $\sqrt{x} = x^{\frac{1}{2}}$. In this sense, we can generalise the Euclidean distance to the *Minkowski* distance measure given as follows:

$$d(i, j) = \left(\sum_{k=1}^{n} |x_{ik} - x_{jk}|^r \right)^{\frac{1}{r}},$$

for some parameter r.

Minkowski distance

Recall: We can write square roots using indices as $\sqrt{x} = x^{\frac{1}{2}}$. In this sense, we can generalise the Euclidean distance to the *Minkowski* distance measure given as follows:

$$d(i, j) = \left(\sum_{k=1}^{n} |x_{ik} - x_{jk}|^r \right)^{\frac{1}{r}},$$

for some parameter r. Of course, when $r = 2$, this is just the usual Euclidean metric.

Important Point: The r parameter should not be confused with the number of dimensions/attributes n.

Varying r

The three most common examples of Minkowski metrics are as follows:

Varying r

The three most common examples of Minkowski metrics are as follows:

- $r = 1$, City block/Manhattan/Taxi-cab distance (also called L_1 distance/norm). This is so named because it is the distance in blocks between any two points in a city, such as 2 blocks down and 3 blocks over for a total of 5 blocks. It is written,

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{in} - x_{jn}|.$$

  A common example is the *Hamming distance*, which is the number of bits that is different between two objects that have only binary attributes.

Varying r

The three most common examples of Minkowski metrics are as follows:

- $r = 1$, City block/Manhattan/Taxi-cab distance (also called L_1 distance/norm). This is so named because it is the distance in blocks between any two points in a city, such as 2 blocks down and 3 blocks over for a total of 5 blocks. It is written,

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \cdots + |x_{in} - x_{jn}|.$$

  A common example is the *Hamming distance*, which is the number of bits that is different between two objects that have only binary attributes.

- $r = 2$, Euclidean distance. Again, sometimes called the L_2 distance/norm. We can generalise the earlier formula. If each attribute is assigned a weight according to its perceived importance, the *weighted* Euclidean distance can be computed as

$$d(i, j) = \sqrt{w_1(x_{i1} - x_{j1})^2 + w_2(x_{i2} - x_{j2})^2 + \cdots + w_n(x_{in} - x_{jn})^2}.$$

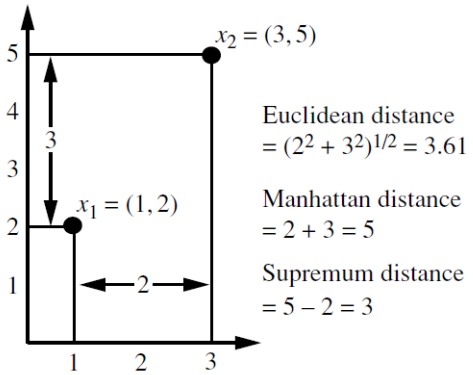
Varying r

- $r \to \infty$, Supremum distance/L_∞ norm.
  This is the maximum distance between any attribute of the objects. We can write the L_∞ norm as,

$$d(i, j) = lim_{r \to \infty} \left(\sum_{k=1}^{n} |x_{ik} - x_{jk}|^r \right)^{\frac{1}{r}}.$$

Example

Consider the following diagram in which we have two objects x_1, x_2.



Example continued

- If we want to work out the Euclidean distance, we compute $\sqrt{(3-1)^2+(5-2)^2} = \sqrt{13}$.

Example continued

- If we want to work out the Euclidean distance, we compute $\sqrt{(3-1)^2+(5-2)^2} = \sqrt{13}$.
- By contrast, the Manhattan distance is given by $|3-1|+|5-2| = 2+3 = 5$.
- Finally, the supremum distance is taken by solving,

$$\lim_{r\to\infty}(2^r + 3^r) = 3.$$

While is is trickier to understand, an intuitive way to think of it is that as r gets large, $2^r + 3^r$ starts to look more and more like 3^r. This leaves us with $(3^r)^{\frac{1}{r}} = 3$. WARNING! This is *very* hand wavy and you should not think you can always do this.

Working with limits

While you should be very careful when limits are involved, in these cases with finitely many attributes and everything reasonably well behaved, we can make the following statement,

$$\lim_{r\to\infty}\left(\sum_{k=1}^{n}|x_{ik}-x_{jk}|^{r}\right)^{\frac{1}{r}} = max\{|x_{i1}-x_{j1}|,\cdots,|x_{in}-x_{jn}|\}.$$

Exercise: Read pp. 68-69 of notes to understand the idea of a metric.

Example

Consider now four points in $\mathbb{R}^2$ represented in the following table:

Point	x coordinate	y coordinate
x_1	0	2
x_2	2	0
x_3	3	1
x_4	5	1

We can now compute the L_1, L_2, L_∞ norms as follows:

L_1	x_1	x_2	x_3	x_4
x_1	0	4	4	6
x_2	4	0	2	4
x_3	4	2	0	2
x_4	6	4	2	0

L_2	x_1	x_2	x_3	x_4
x_1	0	2.8	3.2	5.1
x_2	2.8	0	1.4	3.2
x_3	3.2	1.4	0	2
x_4	5.1	3.2	2	0

L_∞	x_1	x_2	x_3	x_4
x_1	0	2	3	5
x_2	2	0	1	3
x_3	3	1	0	2
x_4	5	3	2	0

Proximity measures for ordinal attributes

Recall: An ordinal attribute is one in which we also take order into account (compared to nominal).

Proximity measures for ordinal attributes

Recall: An ordinal attribute is one in which we also take order into account (compared to nominal).

For simplicity, suppose we have just the one attribute. For example, perhaps this attribute measures the quality of a product on the scale $\{$*poor, fair, okay, good, amazing*$\}$.

- It is reasonable to say that a product P_1 which is rated *amazing* is closer to a product P_2 which is rated *good* than it would be to a product P_3 which is rated *okay*.

Proximity measures for ordinal attributes

Recall: An ordinal attribute is one in which we also take order into account (compared to nominal).

For simplicity, suppose we have just the one attribute. For example, perhaps this attribute measures the quality of a product on the scale $\{$*poor, fair, okay, good, amazing*$\}$.

- It is reasonable to say that a product P_1 which is rated *amazing* is closer to a product P_2 which is rated *good* than it would be to a product P_3 which is rated *okay*.
- To make this quantitative, the values of the ordinal attribute are mapped to successive integers (often starting at 0 or 1), e.g. $\{poor = 0,\ fair = 1,\ okay = 2,\ good = 3,\ amazing = 4\}$. In this case, $d(P_1,\ P_2) = 4 - 3$. If we wanted dissimilarity to fall between 0 and 1, then we could set $d(P_1,\ P_2) = \frac{4-3}{4} = 0.25$. As ever, a similarity attribute can then be defined as $s = 1 - d$.

Using our work for numerical attributes

Suppose that f is an attribute from a set of ordinal attributes describing m objects. The dissimilarity computation with respect to f involves the following steps:

1. The value of f for the i^{th} object is x_{if}, and f has M_f ordered states, representing the ranking $1, \ldots, M_f$. Replace each x_{if} by its corresponding rank $r_{if} \in \{1, \ldots, M_f\}$ (again, can start at 0 if preferred).

Using our work for numerical attributes

Suppose that f is an attribute from a set of ordinal attributes describing m objects. The dissimilarity computation with respect to f involves the following steps:

1. The value of f for the i^{th} object is x_{if}, and f has M_f ordered states, representing the ranking $1, \ldots, M_f$. Replace each x_{if} by its corresponding rank $r_{if} \in \{1, \ldots, M_f\}$ (again, can start at 0 if preferred).
2. Since each ordinal attribute can have a different number of states, it is often necessary to map the range of each attribute onto $[0, 1]$ so that each attribute has equal weight. We perform such data normalisation by replacing the rank r_{if} of the i^{th} object in the f^{th} attribute by,

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}.$$

Using our work for numerical attributes

3. Dissimilarity can then be computed using any of the distance measures described above for numeric attributes, using z_{if} to represent the f value for the i^{th} object.

Example

Return to the table of data from earlier. This time, we only have the object identifier and the continuous ordinal attribute *Test 2*:

Object Identifier	Test 1 - Nominal	Test 2 - Ordinal	Test 3 - Numeric
1	Code A	Excellent	45
2	Code B	Fair	22
3	Code C	Good	64
4	Code A	Excellent	28

There are three states for *Test 2: fair, good* and *excellent*, i.e. $M_f = 3$.

- Step 1 - replace each value for *Test 2* by its rank. The four objects are assigned the ranks 3, 1, 2, 3, respectively.
- Step 2 - normalise by mapping rank 1 to 0, rank 2 to 0.5 and rank 3 to 1.
- Step 3 - use the Euclidean distance (say) to compute distances. This results in the following dissimilarity matrix:

Example continued

$$\begin{pmatrix} 0 & & & \\ 1 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1 & 0.5 & 0 \end{pmatrix}.$$

Using this, objects 1 and 2 are the most dissimilar, as are objects 2 and 4 (i.e. $d(2, 1) = d(4, 2) = 1$).

Exercise: Read pp. 71-72 to understand how to compute proximity measures of attributes of mixed type.

Cosine similarity

- Documents are often represented as vectors, where each component (attribute) represents the frequency with which a particular term (word) occurs in the document.
- Thus, each document is an object represented by what is called a *term-frequency* vector.

Example

Consider the following table which represents articles on each of the top four teams in the Premier League this season (22/23) after 7 games:

Document	Team	Manager	Yellow Cards	Won	Lost	Drawn	Scored	Conceded	Intl Break
Article 1	4	7	3	1	4	2	3	2	1
Article 2	2	5	0	2	6	4	2	3	0
Article 3	1	5	0	4	4	1	1	1	2
Article 4	1	8	0	3	3	0	0	0	0

Example

Consider the following table which represents articles on each of the top four teams in the Premier League this season (22/23) after 7 games:

Document	Team	Manager	Yellow Cards	Won	Lost	Drawn	Scored	Conceded	Intl Break
Article 1	4	7	3	1	4	2	3	2	1
Article 2	2	5	0	2	6	4	2	3	0
Article 3	1	5	0	4	4	1	1	1	2
Article 4	1	8	0	3	3	0	0	0	0

So, in Article 1, there were 7 mentions of the word *manager*, but only 1 mention of *won* and in Articles 2-4 there are no mentions of *Yellow Cards*. Other examples include information retrieval, text document clustering, biological taxonomy, and gene feature mapping. Such data can be highly asymmetric.

Term-frequency vectors and sparsity

- Term-frequency vectors are typically very long and *sparse* (i.e. they have many 0 values).
- As such, much like transaction data, similarity should not depend on the number of shared 0 values because any two documents are likely to ‘not contain’ many of the same words, and therefore if 0-0 matches are counted, most documents will be highly similar to most other documents.

Term-frequency vectors and sparsity

- Term-frequency vectors are typically very long and *sparse* (i.e. they have many 0 values).
- As such, much like transaction data, similarity should not depend on the number of shared 0 values because any two documents are likely to 'not contain' many of the same words, and therefore if 0-0 matches are counted, most documents will be highly similar to most other documents.
- Consequently, a similarity measure for documents needs to ignore 0-0 matches like the Jaccard measure, but also must be able to handle non-binary vectors.

Cosine similarity

Let $\boldsymbol{x}$, $\boldsymbol{y}$ be two vectors for comparison. Using the cosine measure as a similarity function, we have

$$sim(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{||\boldsymbol{x}|| \, ||\boldsymbol{y}||},$$

where

- $||\boldsymbol{x}||$ is the Euclidean norm of $\boldsymbol{x} = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix}^T$ given by $||\boldsymbol{x}|| = \sqrt{x_1^2 + \cdots + x_n^2}$.
- $\boldsymbol{x} \cdot \boldsymbol{y}$ is the scalar product of the two vectors (also called the dot product[1]) and can be written in two ways:

$$\begin{aligned} \boldsymbol{x} \cdot \boldsymbol{y} &= x_1 y_1 + x_2 y_2 + \cdots + x_n y_n \\ &= ||\boldsymbol{x}|| \, ||\boldsymbol{y}|| \cos(\theta), \end{aligned}$$

  where θ is the angle between $\boldsymbol{x}$ and $\boldsymbol{y}$.

[1] In the literature, it is also called an inner product.

What cosine similarity means

- The measure computes the cosine of the angle between the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$.

What cosine similarity means

- The measure computes the cosine of the angle between the vectors ***x*** and ***y***.
- Recall, a cosine value of 0 means that the two vectors are at 90 degrees to each other (orthogonal) and have no match. The closer the cosine value isto 1, the smaller the angle and the greater the match between vectors.
- As $\cos(\theta)$ ranges from -1 to 1, we might expect to get negative similarity.This will not happen if we only have non-negative values in our document vectors. However, if we allow negative values in our vectors (dimensionality reduction via PCA might do this, for example), then we can get negative similarity. In this case, the vectors are pointing in opposite directions and we can think of this as 'anti'-similarity.

Examples

1. Suppose that $\boldsymbol{x}$, $\boldsymbol{y}$ are the first two term-frequency vectors in the above table, i.e.

$$\begin{aligned} \boldsymbol{x} &= \begin{pmatrix} 4 & 7 & 3 & 1 & 4 & 2 & 3 & 2 & 1 \end{pmatrix}^T \\ \boldsymbol{y} &= \begin{pmatrix} 2 & 5 & 0 & 2 & 6 & 4 & 2 & 3 & 0 \end{pmatrix}^T \end{aligned}$$

Then, $\boldsymbol{x} \cdot \boldsymbol{y} = 89$, $||\boldsymbol{x}|| = \sqrt{109}$ and $||\boldsymbol{y}|| = 7\sqrt{2}$. Putting this all together gives,

$$sim(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{||\boldsymbol{x}|| \, ||\boldsymbol{y}||} = \frac{89}{7\sqrt{109}\sqrt{2}} = 0.861\ldots$$

So, if we were using cosine similarity to compare these documents, they would be considered reasonably similar.

Examples Continued

2. Suppose that $\boldsymbol{x}$, $\boldsymbol{y}$ are the first two term-frequency vectors in the above table, i.e.

$$\boldsymbol{x} = \begin{pmatrix} 3 & 2 & 0 & 5 & 0 & 0 & 0 & 2 & 0 & 0 \end{pmatrix}^T$$
$$\boldsymbol{y} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 2 \end{pmatrix}^T$$

Then, $\boldsymbol{x} \cdot \boldsymbol{y} = 5$, $||\boldsymbol{x}|| = 6.48$ and $||\boldsymbol{y}|| = 2.45$. Putting this all together gives,

$$sim(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{||\boldsymbol{x}|| \, ||\boldsymbol{y}||} = 0.31$$

So, if we were using cosine similarity to compare these documents, they would be considered reasonably dissimilar.

Correlation

- Correlation is frequently used to measure the linear relationship between two sets of values that are observed together.
- Thus, correlation can measure the relationship between two variables (height and weight), or between two objects (a pair of temperature time series).

Correlation

- Correlation is frequently used to measure the linear relationship between two sets of values that are observed together.
- Thus, correlation can measure the relationship between two variables (height and weight), or between two objects (a pair of temperature time series).
- Correlation is used much more frequently to measure the similarity between attributes since the values in two data objects can come from different attributes, which can have very different attribute types and scales.
- There are many types of correlation, and indeed correlation is sometimes used in a general sense to mean the relationship between two sets of values that are observed together.
- For nominal data, the χ^2 test (chi-square) is popular, while for numeric attributes, Pearson's correlation is popular. For us, we will just discuss the latter.

Pearson's correlation

Suppose $\boldsymbol{x}$, $\boldsymbol{y}$ are two vectors. Then we define Pearson's correlation as,

$$corr(\boldsymbol{x}, \boldsymbol{y}) = \frac{covariance(\boldsymbol{x}, \boldsymbol{y})}{standard\ deviation(\boldsymbol{x}) \times standard\ deviation(\boldsymbol{y})} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}.$$

Pearson's correlation

Suppose $\boldsymbol{x}$, $\boldsymbol{y}$ are two vectors. Then we define Pearson's correlation as,

$$corr(\boldsymbol{x}, \boldsymbol{y}) = \frac{covariance(\boldsymbol{x}, \boldsymbol{y})}{standard\ deviation(\boldsymbol{x}) \times standard\ deviation(\boldsymbol{y})} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}.$$

Recall, we define covariance and standard deviation as follows:

$$\begin{aligned}
\sigma_{xy} &= \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})(y_k - \bar{y}) \\
\sigma_x &= \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})^2} \\
\sigma_y &= \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (y_k - \bar{y})^2} \\
\bar{x} &= \frac{1}{n} \sum_{k=1}^{n} x_k \\
\bar{y} &= \frac{1}{n} \sum_{k=1}^{n} y_k.
\end{aligned}$$

Example (Perfect correlation)

Correlation is always in the range -1 to 1. A correlation of 1 (resp. -1) means that $\boldsymbol{x}$, $\boldsymbol{y}$ have a perfect positive (resp. negative) linear relationship, i.e. $x_k = ay_k + b$, for some constants a, b.

$$\begin{aligned} \boldsymbol{x} &= \begin{pmatrix} -3 & 6 & 0 & 3 & -6 \end{pmatrix}^T \\ \boldsymbol{y} &= \begin{pmatrix} 1 & -2 & 0 & -1 & 2 \end{pmatrix}^T \\ corr(\boldsymbol{x}, \boldsymbol{y}) &= -1, \text{ and } x_k = -3y_k \end{aligned}$$

$$\begin{aligned} \boldsymbol{x} &= \begin{pmatrix} 3 & 6 & 0 & 3 & 6 \end{pmatrix}^T \\ \boldsymbol{y} &= \begin{pmatrix} 1 & 2 & 0 & 1 & 2 \end{pmatrix}^T \\ corr(\boldsymbol{x}, \boldsymbol{y}) &= 1, \text{ and } x_k = 3y_k \end{aligned}$$

Example (Nonlinear relationships)

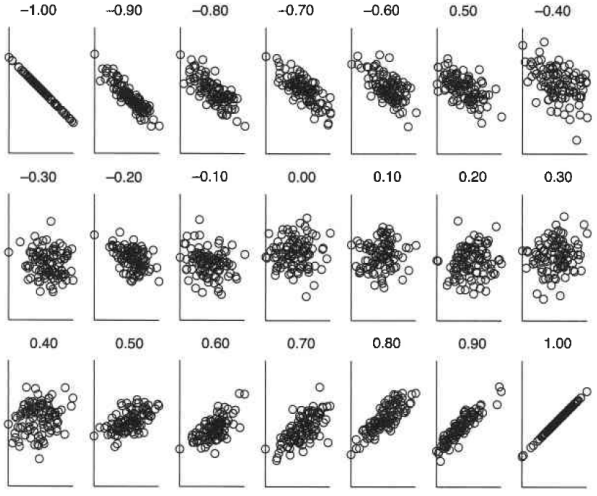
If the correlation is 0, then there is no linear relationship between the two sets of values. However, nonlinear relationships can still exist. In the following example, $y_k = x_k^2$, but their correlation is 0:

$$\begin{aligned} \boldsymbol{x} &= \begin{pmatrix} -3 & -2 & -1 & 0 & 1 & 2 & 3 \end{pmatrix}^T \\ \boldsymbol{y} &= \begin{pmatrix} 9 & 4 & 1 & 0 & 1 & 4 & 9 \end{pmatrix}^T. \end{aligned}$$

Example (Visualising correlation)

- We can also judge the correlation between two vectors by plotting pairs of corresponding values of $\boldsymbol{x}$, $\boldsymbol{y}$ in a scatter plot.
- In the image on the next slide, we have a number of scatter plots for $\boldsymbol{x}$, $\boldsymbol{y}$ consisting of a set of 30 pairs of values that are randomly generated (with a normal distribution) so that the correlation of $\boldsymbol{x}$, $\boldsymbol{y}$ ranges from -1 to 1.
- Each circle in a plot represents one of the 30 pairs of $\boldsymbol{x}$, $\boldsymbol{y}$ values; its x coordinate is the value of that pair for $\boldsymbol{x}$, while its y coordinate is the value of the same pair for $\boldsymbol{y}$.

Example (Visualising correlation)



Relationship with the scalar product

- If we transform $\boldsymbol{x}$, $\boldsymbol{y}$ by subtracting off their means and then normalising so that their lengths are 1, then we can calculate their correlation by taking the scalar product.
- We denote these transformed vectors of $\boldsymbol{x}$, $\boldsymbol{y}$ by $\boldsymbol{x}'$, $\boldsymbol{y}'$, respectively.

Relationship with the scalar product

- If we transform $\boldsymbol{x}$, $\boldsymbol{y}$ by subtracting off their means and then normalising so that their lengths are 1, then we can calculate their correlation by taking the scalar product.
- We denote these transformed vectors of $\boldsymbol{x}$, $\boldsymbol{y}$ by $\boldsymbol{x}'$, $\boldsymbol{y}'$, respectively.
- This highlights an interesting relationship between the correlation measure and the cosine measure.
  - Specifically, the correlation between $\boldsymbol{x}$, $\boldsymbol{y}$ is identical to the cosine between $\boldsymbol{x}'$, $\boldsymbol{y}'$.
  - However, the cosine between $\boldsymbol{x}$, $\boldsymbol{y}$ is not the same as the cosine between $\boldsymbol{x}'$, $\boldsymbol{y}'$, even though they both have the same correlation measure.
- In general, the correlation between two vectors is equal to the cosine measure only in the special case when the means of the two vectors are 0.

Summary

- We have seen Euclidean distance and its generalisation to Minkowski distance.
- We have seen proximity measures for ordinal attributes.
- We have seen how to compute the cosine similarity, i.e.

$$sim(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{||\boldsymbol{x}|| \, ||\boldsymbol{y}||}.$$

- We have seen how to compute Pearson's correlation. Suppose $\boldsymbol{x}$, $\boldsymbol{y}$ are two vectors. Then we define Pearson's correlation as,

$$corr(\boldsymbol{x}, \boldsymbol{y}) = \frac{covariance(\boldsymbol{x}, \boldsymbol{y})}{standard\ deviation(\boldsymbol{x}) \times standard\ deviation(\boldsymbol{y})} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}.$$

.