Similarity Measures

Continuous Case

Euclidean Distance

The length of the shortest possible path through space between two point that could be taken if there were no obstacles:

$$d\left(x,\, y\right) = \sqrt{\sum_{k=1}^{n} \left(x_k - y_k\right)^2}$$

where n is the number of dimensions (attributes) and x_k and y_k are the k^{th} attributes (components) of data objects x and y.

Minkowski Distance

This is a generalization of Euclidean Distance:

$$d\left(x,\, y\right) = \left(\sum_{k=1}^{n} \left|x_k - y_k\right|^r\right)^{\frac{1}{r}}$$

where r is the degree of the distance.

Mahalanobis Distance

The Mahalanobis distance is defined as:

$$d\left(x,\, y\right) = \sqrt{\left(x - y\right) \boldsymbol{\sigma}^{-1} \left(x - y\right)^T}$$

where $\boldsymbol{\sigma}^{-1}$ is the covariance matrix of the data.

- For $r = 1$, the *city block*, (*Manhattan*, *taxicab* or *L1 norm*) distance.
- For $r = 2$, the *Euclidean* distance.
- For $r = \infty$, the *supremum* (L_{max} *norm* or L_∞ *norm*) distance

Cosine Similarity

The similarity is computed as the cosine of the angle that they form:

$$\cos\left(x,\, y\right) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{\|\boldsymbol{x}\| \, \|\boldsymbol{y}\|}$$

where $\cdot$ indicates vector dot product and $\|\boldsymbol{x}\|$ is the norm of vector $\boldsymbol{x}$. This is also known as the *L2 Norm. The term is used as the cosine of the angle only as a convenient mechanism for calculating the angle itself and is no part of the meaning.*

Pearson Correlation

Pearson correlation uses:

$$Pearson\,(x,\,y) = \frac{\Sigma\,(x,\,y)}{\sigma_x \sigma_y}$$

where Σ is the covariance of data points x and y, and σ is the standard deviation. The Pearson correlation is $+1$ in the case of a perfect direct (increasing) linear relationship (correlation), -1 in the case of a perfect decreasing (inverse) linear relationship (anticorrelation), and some value in the open interval $(-1,\,1)$ in all other cases, indicating the degree of linear dependence between the variables. As it approaches zero there is less of a relationship (closer to uncorrelated). The closer the coefficient is to either -1 or 1, the stronger the correlation between the variables.

Binary Case

M_{xy} is the number of attributes presented by x and y which are binary numbers 0 or 1.

Simple Matching coefficient

$$SMC = \frac{M_{00} + M_{11}}{M_{00} + M_{11} + M_{01} + M_{10}}$$

The simple matching distance (SMD), which measures dissimilarity between sample sets, is given by $1 - SMC$.

Jaccard coefficient

The Jaccard index, also known as *Intersection* over Union and the *Jaccard similarity coefficient* is a statistic used for comparing the *similarity* and *diversity* of sample sets:

$$JC = \frac{M_{11}}{M_{11} + M_{01} + M_{10}}$$

The *Jaccard distance*, which measures dissimilarity between sample sets, is complementary to the Jaccard coefficient and is obtained by subtracting the Jaccard coefficient from 1, or, equivalently, by dividing the difference of the sizes of the union and the intersection of two sets by the size of the union $1 - JC$.

Extended Jaccard (Tanimoto) coefficient

$$d = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{\|\boldsymbol{x}\|^2 + \|\boldsymbol{y}\|^2 - \boldsymbol{x} \cdot \boldsymbol{y}}$$