Measures of Distance & Dissimilarity

This section presents measures of distance and dissimilarity that can be used in analysis, which directly make use of the dissimilarity between objects such as K-means clustering and hierarchical clustering analysis.

Similarity measures increase as the similarity between objects increase, while dissimilarity measures decrease as the similarity increases. Since many pattern recognition algorithms traditionally use distance metrics (which measure dissimilarity between objects) Partek Flow converts similarity measures into dissimilarity measures so that they can be interchanged with distance metrics without having to modify the algorithms that use them.

Distance Metrics

Distance Metrics tell how far apart two vectors are in n-dimensional space. Formal definitions of distance functions and distance metrics can be found in a variety of texts on cluster analysis and topology.

Let x and y denote two real vectors $(x_1,...,x_n)^T$ and $(y_1,...,y_n)^T$ (Spath 1980). A real-valued function d(x, y) is said to be a distance function if, and only if, the following three conditions are satisfied:

$$d(x,y) \geq d_0$$
$$d(x,x) = d_0$$
$$d(x,y) = d(y,x)$$

The distance function d(x, y) can further be considered a *metric* **if and only if** in addition to the above three conditions, the following two conditions are also satisfied:

$d(x,y) = d_0$ if and only if $x = y$

$d(x,y) \leq d(x,z) + d(z,y)$ for all $x, y, z \in R^n$ where R^n is n-dimensional Euclidean space and d_0 is an arbitrary real number (usually 0).

Euclidean

The Euclidean distance between vectors x and y is given by

$$d_{euc}(x,y) = \sqrt{\sum_i (x_i - y_i)^2}\ .$$

Euclidean distance is the default measure used in Partek. The Euclidean distance satisfies all conditions of a metric.

Average Euclidean

The average Euclidean distance is the same as the Euclidean distance except that it is normalized by dividing by $\sqrt{n}$:

$$d_{avgEuc}(x,y) = \sqrt{\sum_i \frac{(x_i - y_i)^2}{n}}$$

Because d_{avgEuc} is a scaled version of d_{euc} it will give the same results as d_{euc} in many algorithms. The average Euclidean distance is preferred to the Euclidean distance when the data contains missing values because it does not tend to grow larger as the vector length grows and is better suited to measuring the distance between vectors, which may contain missing values (this assumes that the data has been standardized). The average Euclidean distance satisfies all conditions of a metric.

Squared Euclidean

The squared Euclidean distance between vectors x and y is given by

$$d_{sqrEuc}(x,y) = \sum_i (x_i - y_i)^2$$

It is nearly identical to Euclidean distance. However since it does not compute square root, squared Euclidean is faster than Euclidean distance.

Minkowski Distance

The Minkowski distance is defined as the p^{th} root of the sum of the absolute value of the differences of the vector elements raised to the power p and is therefore a generalization of the Euclidean distance:

$$d_{\min k}(x,y) = \sqrt[p]{\sum_i |\, x_i - y_i \,|^p}$$

Average Minkowski Distance

Since the Minkowski distance is a generalization of the Euclidean distance, it is natural that you also provide an average Minkowski distance for the same reasons that you include the average Euclidean distance. The average Minkowski distance is the same as the Minkowski distance except that it is normalized by dividing by $\sqrt[p]{n}$:

$$d_{avgMink}(x,y) = \sqrt[p]{\sum_i \frac{|\, x_i - y_i \,|^p}{n}}$$

Mahalanobis Distance

The Mahalanobis distance is used when you want to compensate for the fact that different variables may be measured on different scales:

$$d_{mahal}(x,y) = \sqrt{(x-y)^T C^{-1} (x-y)}$$

where C is the covariance matrix of the entire data set. When C^{-1} is the identity matrix, this metric is equivalent to the Euclidean distance. It should also be noted that models that make use of this distance must save C^{-1} as part of the saved model.

Maximum Value

The maximum value distance metric can be used when you only care how close two vectors are at their farthest point. For example, it can be used to measure the maximum deviation between two observations of the same phenomena.

$$d_{\max}(x, y) = \max_i | x_i - y_i |$$

Minimum Value

The minimum value distance function is used when you only care how close two vectors are at their closest point. For example, suppose two vectors contain measurements of altitude of the ground and a high power line. In this case you may only care how close the high power line is to the ground at its closest point.

$$d_{\min}(x, y) = \min_i | x_i - y_i |$$

Absolute Value

Also known as the taxi cab distance, the absolute value distance metric is a special case of the Minkowski distance with *p*=1:

$$d_{abs}(x, y) = \sum_i | x_i - y_i |$$

You can compute an average absolute value distance by using the average Minkowski distance metric and specifying *p*=1.

Tanimoto

The Tanimoto distance is used to see how similar two chemicals are. It does this by counting the number of chemical substructures or chemical groups they have in common:

$$d(x, y) = \frac{x^t y}{x^t x + y^t y - x^t y}$$

Where $x^t y$ is number of attributes possessed by both x and y

The distance is given by the ratio between the number of groups that are occur in both, divided by this plus the number in only one, plus the number only in the other. The number that occurs in neither is ignored.

Measures of Dissimilarity

In addition to the distance metrics described above, Partek provides measures of dissimilarity. These measures tell how similar the shapes of the data profiles are. The first three are simple transformations of three measures of correlation between the vectors. The cosine dissimilarity is the cosine of the angle between the two vectors. Finally, other measures that were specifically designed to measure dissimilarity are presented.

Pearson's Dissimilarity

Pearson's dissimilarity is a transformation of the linear (Pearson's *r*) correlation between two vectors.

Linear Correlation (Pearson's r)

$$r_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$$

When used as a dissimilarity measure, it is rescaled to the interval [0,1] with 0 indicating perfect similarity (perfect positive correlation) and one indicating perfect dissimilarity (perfect negative correlation).

$$d_r(x, y) = \frac{(1 - r)}{2}$$

where r is the linear correlation.

Pearson's Absolute Dissimilarity

Pearson's Absolute Value dissimilarity is a slight modification of Pearson's dissimilarity. It is rescaled to the interval [0,1] with 0 indicating either maximum similarity or dissimilarity and 1 indicating uncorrelated.

$$d_{rabs}(x, y) = 1 - |r|$$

where r is the linear correlation.

Rank Dissimilarity

Rank dissimilarity is a transformation of Spearman's non-parametric r_s correlation between two vectors and is called for when the data is ordinal.

Correlation (Spearman's Rank coefficient)

$$r_s = \frac{\sum_i (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_i (R_i - \bar{R})^2}\sqrt{\sum_i (S_i - \bar{S})^2}}$$

where R_i is the rank of x_i in the vector x, S_i is the rank of y_i in the vector y.

When used as a dissimilarity measure, it is rescaled to the interval [0,1] with 0 indicating perfect similarity (perfect positive correlation) and 1 indicating perfect dissimilarity (perfect negative correlation).

$$d_{r_s}(x, y) = \frac{1 - r_s(x, y)}{2}$$

where r_s is Spearman's rank order coefficient.

Rank Absolute Dissimilarity

Rank absolute value dissimilarity is a slight modification of Rank dissimilarity. When used as a dissimilarity measure, it is rescaled to the interval [0,1] with 0 indicating either maximum similarity or dissimilarity and 1 indicating uncorrelated.

$$d_{rabs_s}(x, y) = 1 - |r_s(x, y)|$$

where r_s is Spearman's rank order coefficient.

Kendall's Dissimilarity

Kendall's dissimilarity is the third dissimilarity metric based on the correlation between the vectors and is computed by:

$$d_\tau(x,y) = \frac{(1-\tau)}{2}$$

where τ is Kendall's Tau correlation.

Kendall's Tau

$$\tau = \frac{concordant - discordant}{\sqrt{concordant + discordant + extraY}\sqrt{concordant + discordant + extraX}}$$

It is rescaled to the interval [0,1] with 0 indicating perfect similarity (perfect positive correlation) and one indicating perfect dissimilarity (perfect negative correlation).

Kendall's Absolute Dissimilarity

Kendall's absolute value dissimilarity is a slight modification of Kendall's dissimilarity. When used as a dissimilarity measure, it is rescaled to the interval [0,1] with 0 indicating either maximum similarity or dissimilarity and 1 indicating uncorrelated.

$$d_{\tau abs}(x,y) = 1 - |\tau|$$

where τ is Kendall's Tau correlation.

Coefficient of Shape Difference

Created by Penrose, the coefficient of shape difference is defined in the range [0, ∞] and is a function of the average Euclidean distance. The shape difference ignores additive displacement and therefore gives similar results to the cosine dissimilarity and measures based on correlation.

$$d_{shape}(x,y) = \sqrt{\frac{n}{n-1}(d_{avgEuc}(x,y)^2 - q(x,y)^2)}$$

where $d_{avgEuc}(x,y)$ is *Average Euclidean Distance* and q(x,y) is given by:

$$q(x,y) = \frac{\sum_i x_i - \sum_i y_i}{n}$$

Cosine Dissimilarity

The cosine dissimilarity is based on the cosine coefficient cos(x,y) (defined in the interval [-1,1]):

$$\cos(x,y) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^1}}$$

The cosine coefficient measures the cosine of the angle formed by the vectors x and y. Convert cos(x,y) to a measure of dissimilarity in the interval [0,1] as follows:

$$d_{cosine}(x, y) = \frac{(1 - \cos(x, y))}{2}$$

Canberra Metric

The Canberra metric is a dissimilarity measure defined on the interval [0,1] and satisfies all four conditions of a metric.

$$d_{canberra}(x, y) = \frac{1}{n} \sum_i \frac{| x_i - y_i |}{(x_i + y_i)}$$

Bray-Curtis Coefficient

The Bray-Curtis coefficient is a dissimilarity measure defined on the interval [0,1] and satisfies all four conditions of a metric.

$$d_{bc}(x, y) = \frac{\sum_i | x_i - y_i |}{\sum_i (x_i + y_i)}$$