# Distance/Similarity Measures

## Terminology

**Similarity**: measure of how close to each other two instances are. The "closer" the instances are to each other, the larger is the similarity value.

**Dissimilarity**: measure of how different two instances are. Dissimilarity is large when instances are very different and is small when they are close.

**Proximity**: refers to either similarity or dissimilarity

**Distance metric**: a measure of dissimilarity that obeys the following laws (laws of triangular norm):

- $d(x, y) \leq 0$; $d(x, y) = 0$ **iff** $x = y$;
- $d(x, y) = d(y, x)$;
- $d(x, y) + d(y, z) \geq d(x, z)$.

### Conversion of similarity and dissimilarity measures.

Typically, given a similarity measure, one can "revert" it to serve as the dissimilarity measure and vice versa.

Conversions may differ. E.g., if $d$ is a distance measure, one can use

$$s(x, y) = \frac{1}{d(x, y)}$$

or

$$s(x, y) = \frac{1}{d(x, y) + 0.5}$$

as the corresponding similarity measure. If $s$ is the similarity measure that ranges between 0 and 1 (so called degree of similarity), then the corresponding dissimilarity measure can be defined as

$$d(x, y) = 1 - s(x, y)$$

or

$$d(x,y) = \sqrt{(1 - s(x,y))}.$$

In general, **any** *monotonically decreasing transformation* can be applied to convert similarity measures into dissimilarity measures, and **any** *monotonically increasing transformtaion* can be applied to convert the measures the other way around.

## Distance Metrics for Numeric Attributes

When the data set is presented in a *standard* form, each instance can be treated as a vector $\bar{x} = (x_1, \ldots, x_N)$ of measures for attributes numbered $1, \ldots, N$.

Consider for now only non-nominal scales.

**Eucledian Distance.**

$$d_E(\bar{x}, \bar{y}) = \sqrt{\left(\sum_{k=1}^{N} (x_k - y_k)^2\right)}.$$

**Squared Eucledian Distance**

$$d_E(\bar{x}, \bar{y}) = \sum_{k=1}^{N} (x_k - y_k)^2.$$

**Manhattan Distance.**

$$d_m(\bar{x}, \bar{y}) = \sum_{k=1}^{N} |x_k - y_k|.$$

**Minkowski Distance.**

Generalization of Eucledian and Manhattan distances:

$$d_{M,\lambda}(\bar{x}, \bar{y}) = \left(\sum_{k=1}^{N} (x_k - y_k)^{\lambda}\right)^{\frac{1}{\lambda}}.$$

In particular: $d_{M,1} = d_m$, $d_{M,2} = d_E$. Also of interest:

**Chebyshev Distance.**

$$d_{M,\infty}(\bar{x}, \bar{y}) = \max_{k=1,\ldots,N} (|x_k - y_k|).$$

**Additivity.**

Eucledian distance is *additive*: contributions to the distance for each attribute are independent and are summed up.

**Commesurability.** Different attributes may have different scales of measurement. Attributes are commesurable, when their numeric values contribute equally to the actual distance/proximity between instantes.

For example, is instances represent 3D positions of points in space, all attributes are commesurable.

**Standardized (Normalized) Eucledian Distance.**

When some attributes are not commesurable with others, it may be possible "normalize" them by dividing the attribute values by the standard deviation of the attribute over the entire dataset.

**Range standadization.** Each data point is standardized by mapping from its current range to [0,1]. Each attribute value $x_i$ of the data point $\bar{x} = (x_1, \ldots, x_N)$ is standardized as follows:

$$x_i' = \frac{x_i - \min_{\bar{y} \in D}(y_i)}{\max_{\bar{y} \in D}(y_i) - \min_{\bar{y} \in D}(y_i)}.$$

**$z$-score standadization** . Assumes normal distribution of of attribute values. Normalizes the data by using mean and standard deviation of the values of the attribute.

Standard deviation for $i$th attribute:

$$\hat{\sigma}_i = \sqrt{\left(\frac{1}{n-1}\sum_{i=j}^{n}(\bar{x_j}[i] - \mu_i)^2\right)},$$

where $n$ is the number of instances in the data set, and

$$\mu_i = \frac{\sum_{j=1}^{n} \bar{x_j}[i]}{n},$$

is the mean of the $i$th attribute.

The $z$-score standardization of a vector $x = (x_1, \ldots x_N)$ is:

$$\hat{x} = (x_1', \ldots, x_N') = \left(\frac{x_1 - \mu_1}{\hat{\sigma}_1}, \ldots \frac{x_N - \mu_N}{\hat{\sigma}_N}\right)$$

**Standardized Eucledian distance** is then:

$$d_{SE}(\bar{x}, \bar{y}) = d_E(\hat{x}, \hat{y}).$$

**Weighted Distances.**

Different attributes may also be of different *importance* for the purposes of determining distance. Often, this importance is quantified as the *attribute weight*. Given a vector $w = (w_1, \ldots, w_N)$ of attribute weights, the weighted Minkowski distance is computed as:

$$d_{WM,\lambda}(\bar{x}, \bar{y}) = \left(\sum_{k=1}^{N} w_k \cdot (x_k - y_k)^\lambda\right)^{\frac{1}{\lambda}}.$$

From here, we can derive formulas for weighted Eucledian and weighted Manhattan distances.

# Distance Measures for Categorical Attributes

### Distance Measures for Binary Vectors

**Binary vectors.** Vectors $\bar{v} = (v_1, \ldots, v_n) \in \{0,1\}^n$.

**Confusion matrix for binary vectors.** Let $\bar{x} = (x_1, \ldots, x_n)$ and $\bar{y} = (y_1, \ldots, y_n)$ be two binary vectors.

For each attribute $i = 1 \ldots n$, four cases are possible:

| No. | $x_i$ | $y_i$ |
|---|---|---|
| (1) | 1 | 1 |
| (2) | 1 | 0 |
| (3) | 0 | 1 |
| (4) | 0 | 0 |

We count the incidence of each of the four cases and organize these numbers in a **confusion matrix** form:

| | $x_i = 1$ | $x_i = 0$ |
|---|---|---|
| $y_i = 1$ | $A$ | $B$ |
| $y_i = 0$ | $C$ | $D$ |

**Symmetric attributes.** Binary attributes are **symmetric** if both 0 and 1 values have equal importance (e.g. Male and Female or McCain and Obama).

If binary vectors have symmetric attributes, the following distance computations can be performed:

**Simple Matching Distance:**

$$d_s(\bar{x}, \bar{y}) = \frac{B + C}{A + B + C + D}.$$

**Simple Weighted Matching Distance:**

$$d_{s,\alpha}(\bar{x}, \bar{y}) = \frac{\alpha \cdot (B + C)}{A + D + \alpha \cdot (B + C)},$$

or

$$d_{s,\alpha}(\bar{x}, \bar{y}) = \frac{B + C}{\alpha \cdot (A + D) + B + C}.$$

**Assymetric Attributes.** Binary attributes are **assymetric** if one of the states is more imortant than the other (e.g., true and false, present and absent). We assume that 1 is more important than 0.

**Jaccard distance:**

$$d_J(\bar{x}, \bar{y}) = \frac{B + C}{A + B + C}.$$

**Weighted Jaccard Distance**

$$d_{J,\alpha}(\bar{x}, \bar{y}) = \frac{\alpha \cdot (B + C)}{A + \alpha \cdot (B + C)}.$$

$$d_{J,\alpha}(\bar{x}, \bar{y}) = \frac{B + C}{\alpha \cdot A + B + C}.$$

**Non-binary categorical attributes**

**Simple Matching distance.**

$$d_s(\bar{x}, \bar{y}) = \frac{n - q}{n},$$

where:

$n$ : number of attributes in $\bar{x}$ and $\bar{y}$.

$q$ : number of attributes in $\bar{x}$ and $\bar{y}$ that have matching values.

## Using Covariances to compute distances.

Sometimes, some attributes/dimensions *correlate* with each other (e.g., different measurements of the same feature). If not accounted for, such attributes may "hijack" the distance computation.

Geometric intuition: generally, we consider all attributes to correspond to *independent orthogonal dimensions*. Attributes that are not independent do not correspond to *orthogonal dimensions*.

We can use *correlation coefficients* and *covariance coefficients* to correct our distance computation.

**Sample Covariance**

$$cov(i, j) = \frac{1}{n} \sum_{k=1}^{n} (x_{ki} - \mu_i)(x_{kj} - \mu_j),$$

where $\mu_i$ and $\mu_j$ are sample means for $i$th and $j$th attributes respectively. We can construct matrix $C = (cov(i, j))$ of covariances. $C$ is symmetric.

We can also standardize covariance coefficients. *Correlation coefficient* is computed as:

$$\rho(i, j) = \frac{\sum_{k=1}^{n} (x_{ki} - \mu_i)(x_{kj} - \mu_j)}{(\sum_{k=1}^{n} (x - \mu_i)^2 \sum_{k=1}^{n} (x - \mu_j)^2)^{\frac{1}{2}}}.$$

We can form the matrix $\mathcal{S}$ of correlation coefficients $\rho(i, j)$.

Covariance/correlation coefficients can only capture linear dependency between the variables. Non-linear relations are left "out".

**Mahalanobis Distance.**

$$d_{MH}(\bar{x}, \bar{y}) = (\bar{x} - \bar{y})^T \mathcal{S}^{-1} (\bar{x} - \bar{y}).$$

Note: here $\bar{x}$, $\bar{y}$ are treated as **columns**.