Similarity and Dissimilarity

M. H. Bashogh

March 2025

1 Types of Data

In data mining, data attributes can be categorized into two main types:

1.1 Categorical (Qualitative) Data

- **Nominal:** Categorical data with no intrinsic ordering, e.g., colors (red, blue, green).
- **Ordinal:** Categorical data with a meaningful order but no fixed interval, e.g., ratings (low, medium, high).

1.2 Numeric (Quantitative) Data

Numeric data can be further divided into:

- **Interval:** Numerical data with equal intervals but no true zero, e.g., temperature in Celsius.
- **Ratio:** Numerical data with a true zero point, e.g., weight, height, income.

interval and ratio data can be classified as:

- **Discrete:** Has a finite or countably infinite set of values, often represented as integer variables. Examples include zip codes, counts, and word occurrences in documents. A special case is **binary attributes** (e.g., yes/no, 0/1).
- **Continuous:** Has real numbers as attribute values, typically represented as floating-point variables. Examples include temperature, height, and weight. Real values can only be measured and represented using a finite number of digits.

2 Types of Data Set

data sets can be categorized into different types based on their structure and characteristics:

- **Record (Tabular) Data:** Structured data stored in tabular form.
  - **Data Matrix:** A matrix where rows represent instances and columns represent attributes.

  - **Document Data:** Text-based data represented as term-frequency matrices.
  - **Transaction Data:** Data representing transactions, such as market basket analysis.

- **Graph Data:** Data represented as nodes and edges, such as:
  - **World Wide Web:** Hyperlinked web pages.
  - **Molecular Structures:** Chemical compounds represented as graphs.

- **Ordered Data:** Data with an inherent order, such as:
  - **Spatial Data:** Data with geographic or spatial relationships.
  - **Temporal Data:** Data that changes over time, such as stock prices.
  - **Sequential Data:** Ordered sequences, such as time-series data.
  - **Genetic Sequence Data:** Biological sequences such as DNA or protein structures.

- **Spatial/Image Data:** Data that includes spatial and image-based representations.
  - **Maps:** Geographic information represented in layers.
  - **Images:** Photographic or graphical data used in analysis.

3 Central tendency, variation and spread

3.1 Measuring the Central Tendency

Central tendency measures describe the center or typical value of a dataset. The key measures include:

- **Mean:** The arithmetic average of all data points, given by:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- **Weighted Arithmetic Mean:** A mean where each data point is assigned a weight reflecting its importance, given by:

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

- **Trimmed Mean:** The mean calculated after removing a fixed percentage of the smallest and largest values.
- **Median:** The middle value in a sorted dataset, estimated by interpolation for grouped data:

$$\text{median} = L_1 + \left(\frac{n/2 - (\sum freq)_l}{freq_{median}} \right) \text{width}$$

L_1: low inteval limit.

n: Total number of observations.

$freq_{median}$: Frequency of the median class..

$(\sum freq)_l$: Cumulative frequency before the median class.

width: difference between upper and lower boundaries of the class interval.

- **Mode:** The most frequently occurring value in a dataset, estimated using the empirical formula:

$$\text{mean} - \text{mode} = 3 \times (\text{mean} - \text{median})$$

$$\text{Mode} = 3\text{median} - 2\text{mean}$$

3.2 Measuring the Dispersion of Data

Dispersion measures the spread of data points in a dataset. The following metrics are commonly used to quantify dispersion:

3.2.1 Variance

Variance (σ^2) measures the average squared deviation from the mean:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2 \tag{1}$$

where x_i are the data points, μ is the mean, and N is the total number of observations.

3.2.2 Sample Variance

When the mean is unknown, the sample variance (s^2) is computed as:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right] \tag{2}$$

where:

- x_i are the data points,
- $\bar{x}$ is the sample mean,
- n is the sample size.

3.2.3 Standard Deviation

Standard deviation is the square root of variance:

$$\sigma = \sqrt{\sigma^2}, \quad s = \sqrt{s^2} \tag{3}$$

It provides a measure of dispersion in the same units as the original data.

3.2.4 Quartiles

Quartiles divide a dataset into four equal parts:

- Q_1 (First Quartile): 25% of data falls below this value.
- Q_2 (Median): 50% of data falls below this value.
- Q_3 (Third Quartile): 75% of data falls below this value.

3.2.5 Outliers

Outliers are data points that lie significantly outside the typical range. A common rule to identify outliers is:

$$\text{Lower Bound} = Q_1 - 1.5 \times IQR, \quad \text{Upper Bound} = Q_3 + 1.5 \times IQR \tag{4}$$

where the Interquartile Range (IQR) is defined as:

$$IQR = Q_3 - Q_1 \tag{5}$$

Any data point outside these bounds is considered an outlier.

3.2.6 Boxplots

A boxplot is a graphical representation of data dispersion, displaying:

- Minimum (excluding outliers)
- Q_1 (First Quartile)
- Median (Q_2)
- Q_3 (Third Quartile)
- Maximum (excluding outliers)
- Outliers as separate points

4 Similarity and Dissimilarity Measures

Similarity and dissimilarity measures are fundamental in data analysis, clustering, and machine learning. They quantify the relationship between data objects.

4.1 Similarity Measure

- Numerical measure of how alike two data objects are.
- Is higher when objects are more alike.
- Often falls in the range $[0, 1]$.

4.2 Dissimilarity Measure

- Numerical measure of how different two data objects are.
- Lower when objects are more alike.
- Minimum dissimilarity is often 0.
- Upper limit varies depending on the measure.

4.3 Proximity

Proximity refers to either similarity or dissimilarity, depending on the context.

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } x = y \\ 1 & \text{if } x \neq y \end{cases}$	$s = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y \end{cases}$
Ordinal	$d = \frac{\lvert x-y \rvert}{(n-1)}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - d$
Interval or Ratio	$d = \lvert x - y \rvert$	$s = -d, \quad s = \frac{1}{1+d}, \quad s = e^{-d}, \quad s = 1 - \frac{d - min_d}{max_d - min_d}$

Table 1: Dissimilarity and Similarity Measures for Different Attribute Types

4.4 Distance Measures

Distance measures quantify the dissimilarity between data points in a given space. Some commonly used distance measures include:

4.4.1 Euclidean Distance

The Euclidean distance between two points $x = (x_1, x_2, ..., x_n)$ and $y = (y_1, y_2, ..., y_n)$ in an n-dimensional space is given by:

$$d_E(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{6}$$

This is the most commonly used distance metric in geometric space.

4.4.2 Minkowski Distance

Minkowski distance generalizes Euclidean and Manhattan distances and is defined as:

$$d_M(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}} \tag{7}$$

Special cases:

- $p = 1$ results in the Manhattan Distance.
- $p = 2$ results in the Euclidean Distance.

4.4.3 Mahalanobis Distance

Mahalanobis distance accounts for correlations between variables and is defined as:

$$d_{Mah}(x, y) = \sqrt{(x - y)^T \sum^{-1} (x - y)} \tag{8}$$

where $\sum$ is the covariance matrix of the dataset. It is useful in identifying outliers and handling correlated features.

4.5 Similarity Between Binary Vectors

Binary vectors represent objects using binary values (0 or 1), where similarity measures determine how alike two binary vectors are. Common measures include:

4.5.1 Simple Matching Coefficient (SMC)

The Simple Matching Coefficient (SMC) measures the proportion of matching attributes between two binary vectors:

$$SMC = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}} \tag{9}$$

where:

- f_{11} = Number of attributes where both vectors have 1s.
- f_{00} = Number of attributes where both vectors have 0s.
- f_{10} = Number of attributes where the first vector is 1 and the second is 0.
- f_{01} = Number of attributes where the first vector is 0 and the second is 1.

4.5.2 Jaccard Similarity

The Jaccard similarity coefficient measures the proportion of shared attributes where at least one of the vectors has a 1:

$$J = \frac{f_{11}}{f_{11} + f_{10} + f_{01}} \tag{10}$$

This measure ignores cases where both values are 0.

4.5.3 Cosine Similarity

Cosine similarity measures the cosine of the angle between two binary vectors:

$$s_{\cos}(x, y) = \frac{x \cdot y}{\|x\|\|y\|} \tag{11}$$

where:

- $x \cdot y$ is the inner product of the two binary vectors.
- $\|x\|$ and $\|y\|$ are the magnitudes (L2 norm) of the vectors.

These similarity measures are widely used in text analysis, clustering, and classification tasks involving binary data.

5 Correlation Measures the Linear Relationship Between Objects

Correlation quantifies the strength and direction of a linear relationship between two variables. It is computed using the formula:

$$\text{corr}(\mathbf{x}, \mathbf{y}) = \frac{\text{covariance}(\mathbf{x}, \mathbf{y})}{\text{standard\_deviation}(\mathbf{x}) \times \text{standard\_deviation}(\mathbf{y})} = \frac{s_{xy}}{s_x s_y} \tag{12}$$

where we use the following standard statistical notation and definitions:

$$\text{covariance}(\mathbf{x}, \mathbf{y}) = s_{xy} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})(y_k - \overline{y}) \tag{13}$$

$$\text{standard\_deviation}(\mathbf{x}) = s_x = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})^2} \tag{14}$$

$$\text{standard\_deviation}(\mathbf{y}) = s_y = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (y_k - \overline{y})^2} \tag{15}$$

The mean values of x and y are given by:

$$\overline{x} = \frac{1}{n} \sum_{k=1}^{n} x_k, \quad \overline{y} = \frac{1}{n} \sum_{k=1}^{n} y_k \tag{16}$$

The Pearson correlation coefficient is given by:

$$\rho_{X,Y} = \text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\mathbb{E}[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \tag{17}$$

where:

- $\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]$ is the expected value of the product of deviations.
- σ_X, σ_Y are the standard deviations of X and Y.

This measure is widely used in statistics and data analysis to assess relationships between numerical variables.

Correlation vs. Cosine vs. Euclidean Distance

Behavior under Transformations

- **Scaling (multiply all elements by a constant c):**
  - Correlation: invariant
  - Cosine: invariant (if $c > 0$)
  - Euclidean Distance: *changes* (multiplied by $|c|$)
- **Translation (add a constant a to all elements):**
  - Correlation: invariant
  - Cosine: *changes*
  - Euclidean Distance: *changes*

Choice of Proximity Measure

- **Correlation** is suitable when we care about the pattern or trend rather than absolute values (e.g., comparing temperature profiles).
- **Cosine Similarity** is suitable for comparing the orientation of vectors (e.g., text documents with word counts).
- **Euclidean Distance** is suitable when absolute numeric differences matter (e.g., physical distances, direct numeric comparisons).

6 Entropy and Mutual Information

Entropy and mutual information are fundamental concepts in information theory, used to measure uncertainty and shared information between variables.

6.1 Entropy

Entropy quantifies the amount of uncertainty or randomness in a dataset. Given a set of observations of some attribute X with n different possible values, the entropy is defined as:

$$H(X) = -\sum_{i=1}^{n} \frac{m_i}{m} \log_2 \frac{m_i}{m}$$

where m_i is the number of occurrences of the ith category and m is the total number of observations.

6.2 Mutual Information

Mutual information measures the amount of information one variable provides about another. It is defined as:

$$I(X, Y) = H(X) + H(Y) - H(X, Y)$$

where $H(X, Y)$ is the joint entropy of X and Y, given by:

$$H(X, Y) = -\sum_{i} \sum_{j} p_{ij} \log_2 p_{ij}$$

where p_{ij} is the probability that the ith value of X and the jth value of Y occur together.

For discrete variables, mutual information is straightforward to compute and is maximized as:

$$\log_2(\min(n_X, n_Y))$$

where n_X and n_Y are the number of values of X and Y, respectively.