

Similarity and Dissimilarity

- Similarity
 - Numerical measure of how alike two data objects are.
 - Is higher when objects are more alike.
 - Often falls in the range $[0,1]$
- Dissimilarity
 - Numerical measure of how different are two data objects
 - Lower when objects are more alike
 - Minimum dissimilarity is often 0
 - Upper limit varies
- Proximity refers to a similarity or dissimilarity

Similarity / Dissimilarity for Simple Attributes

p and q are the attribute values for two data objects.

| Attribute Type | Dissimilarity | Similarity |
|-------------------|---|---|
| Nominal | $d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$ | $s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$ |
| Ordinal | $d = \frac{ p-q }{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values) | $s = 1 - \frac{ p-q }{n-1}$ |
| Interval or Ratio | $d = p - q $ | $s = -d,$ |

Table 5.1. Similarity and dissimilarity for simple attributes

Euclidean Distance

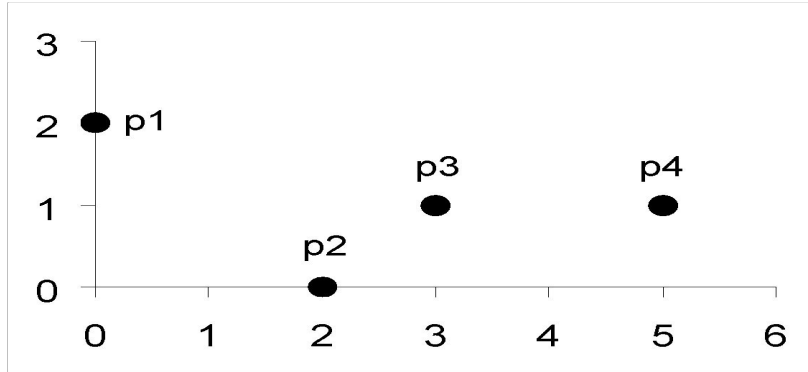
- Euclidean Distance

$$\mathbf{dist} = \sqrt{\sum_{k=1}^n (\mathbf{p}_k - \mathbf{q}_k)^2}$$

Where n is the number of dimensions (attributes) and p_k and q_k are, respectively, the k^{th} attributes (components) or data objects p and q .

- Standardization is necessary, if scales differ.

Euclidean Distance

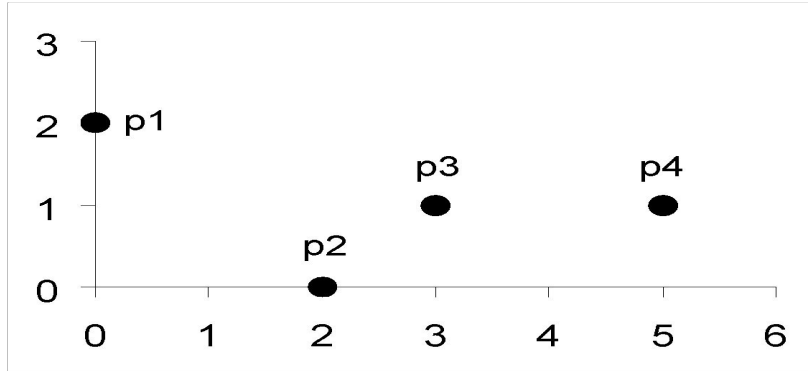


| point | x | y |
|-------|---|---|
| p1 | 0 | 2 |
| p2 | 2 | 0 |
| p3 | 3 | 1 |
| p4 | 5 | 1 |

| | p1 | p2 | p3 | p4 |
|----|----|----|----|----|
| p1 | | | | |
| p2 | | | | |
| p3 | | | | |
| p4 | | | | |

Distance Matrix

Euclidean Distance



| point | x | y |
|-------|---|---|
| p1 | 0 | 2 |
| p2 | 2 | 0 |
| p3 | 3 | 1 |
| p4 | 5 | 1 |

| | p1 | p2 | p3 | p4 |
|----|-------|-------|-------|-------|
| p1 | 0 | 2.828 | 3.162 | 5.099 |
| p2 | 2.828 | 0 | 1.414 | 3.162 |
| p3 | 3.162 | 1.414 | 0 | 2 |
| p4 | 5.099 | 3.162 | 2 | 0 |

Distance Matrix

Minkowski Distance

- Minkowski Distance is a generalization of Euclidean Distance

$$\mathbf{dist} = \left(\sum_{k=1}^n |p_k - q_k|^r \right)^{\frac{1}{r}}$$

Where r is a parameter, n is the number of dimensions (attributes) and p_k and q_k are, respectively, the k th attributes (components) or data objects p and q .

Minkowski Distance : Example

- $r = 1$. City block (Manhattan, taxicab, L_1 norm) distance.
 - A common example of this is the Hamming distance, which is just the number of bits that are different between two binary vectors
- $r = 2$. Euclidean distance
- $r \rightarrow \infty$. “supremum” (L_{\max} norm, L_{∞} norm) distance.
 - This is the maximum difference between any component of the vectors

$$d_S(P, Q) = \max\{|x_1 - x_2|, |y_1 - y_2|\}$$

- Do not confuse r with n , i.e., all these distances are defined for all numbers of dimensions.

Minkowski Distance

| point | x | y |
|-------|---|---|
| p1 | 0 | 2 |
| p2 | 2 | 0 |
| p3 | 3 | 1 |
| p4 | 5 | 1 |

| L1 | p1 | p2 | p3 | p4 |
|----|----|----|----|----|
| p1 | 0 | 4 | 4 | 6 |
| p2 | | | | |
| p3 | | | | |
| p4 | | | | |

| L2 | p1 | p2 | p3 | p4 |
|----|----|-------|-------|-------|
| p1 | 0 | 2.828 | 3.162 | 5.099 |
| p2 | | | | |
| p3 | | | | |
| p4 | | | | |

| L_{∞} | p1 | p2 | p3 | p4 |
|--------------|----|----|----|----|
| p1 | 0 | 2 | 3 | 5 |
| p2 | | | | |
| p3 | | | | |
| p4 | | | | |

Distance Matrix

Minkowski Distance

| point | x | y |
|-------|---|---|
| p1 | 0 | 2 |
| p2 | 2 | 0 |
| p3 | 3 | 1 |
| p4 | 5 | 1 |

| L1 | p1 | p2 | p3 | p4 |
|----|----|----|----|----|
| p1 | 0 | 4 | 4 | 6 |
| p2 | 4 | 0 | 2 | 4 |
| p3 | 4 | 2 | 0 | 2 |
| p4 | 6 | 4 | 2 | 0 |

| L2 | p1 | p2 | p3 | p4 |
|----|-------|-------|-------|-------|
| p1 | 0 | 2.828 | 3.162 | 5.099 |
| p2 | 2.828 | 0 | 1.414 | 3.162 |
| p3 | 3.162 | 1.414 | 0 | 2 |
| p4 | 5.099 | 3.162 | 2 | 0 |

| L_{∞} | p1 | p2 | p3 | p4 |
|--------------|----|----|----|----|
| p1 | 0 | 2 | 3 | 5 |
| p2 | 2 | 0 | 1 | 3 |
| p3 | 3 | 1 | 0 | 2 |
| p4 | 5 | 3 | 2 | 0 |

Distance Matrix

Common Properties of a Distance

- Distances, such as the Euclidean distance, have some well known properties.

1. $d(p, q) \geq 0$ for all p and q and $d(p, q) = 0$ only if $p = q$. (Positive definiteness)
2. $d(p, q) = d(q, p)$ for all p and q . (Symmetry)
3. $d(p, r) \leq d(p, q) + d(q, r)$ for all points p, q , and r . (Triangle Inequality)

where $d(p, q)$ is the distance (dissimilarity) between points (data objects), p and q .

Common Properties of a Similarity

- Similarities, also have some well known properties.
 - $s(p, q) = 1$ (or maximum similarity) only if $p = q$.
 - $s(p, q) = s(q, p)$ for all p and q . (Symmetry)

where $s(p, q)$ is the similarity between points (data objects), p and q .

Similarity between Binary Vectors

- Common situation is that objects, p and q , have only binary attributes
- Compute similarities using the following quantities

M_{01} = the number of attributes where p was 0 and q was 1

M_{10} = the number of attributes where p was 1 and q was 0

M_{00} = the number of attributes where p was 0 and q was 0

M_{11} = the number of attributes where p was 1 and q was 1

- Simple Matching and Jaccard Coefficients

SMC = number of matches / number of attributes

$$= (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$$

J = number of 11 matches / number of not-both-zero attributes values

$$= (M_{11}) / (M_{01} + M_{10} + M_{11})$$

SMC versus Jaccard : Example

$$p = 1000000000$$

$$q = 0000001001$$

$M_{01} = 2$ (the number of attributes where p was 0 and q was 1)

$M_{10} = 1$ (the number of attributes where p was 1 and q was 0)

$M_{00} = 7$ (the number of attributes where p was 0 and q was 0)

$M_{11} = 0$ (the number of attributes where p was 1 and q was 1)

$$SMC = (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00}) = ?$$

$$J = (M_{11}) / (M_{01} + M_{10} + M_{11}) = ?$$

SMC versus Jaccard : Example

$$p = 1000000000$$

$$q = 0000001001$$

$M_{01} = 2$ (the number of attributes where p was 0 and q was 1)

$M_{10} = 1$ (the number of attributes where p was 1 and q was 0)

$M_{00} = 7$ (the number of attributes where p was 0 and q was 0)

$M_{11} = 0$ (the number of attributes where p was 1 and q was 1)

$$SMC = (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00}) = (0+7) / (2+1+0+7) = 0.7$$

$$J = (M_{11}) / (M_{01} + M_{10} + M_{11}) = 0 / (2 + 1 + 0) = 0$$

Cosine Similarity

- If d_1 and d_2 are two document vectors, then

$$\cos(d_1, d_2) = (d_1 \cdot d_2) / \|d_1\| \|d_2\| ,$$

where \cdot indicates vector dot product and $\|d\|$ is the length of vector d .

- Example:

$$d_1 = 3205000200$$

$$d_2 = 1000000102$$

$$d_1 \cdot d_2 = ?$$

$$\|d_1\| = ?$$

$$\|d_2\| = ?$$

$$\cos(d_1, d_2) = ?$$

Cosine Similarity

- If d_1 and d_2 are two document vectors, then

$$\cos(d_1, d_2) = (d_1 \cdot d_2) / ||d_1|| ||d_2|| ,$$

where \cdot indicates vector dot product and $||d||$ is the length of vector d .

- Example:

$$d_1 = 3205000200$$

$$d_2 = 1000000102$$

$$d_1 \cdot d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$$

$$||d_1|| = (3*3 + 2*2 + 0*0 + 5*5 + 0*0 + 0*0 + 0*0 + 2*2 + 0*0 + 0*0)^{0.5} = (42)^{0.5} = 6.481$$

$$||d_2|| = (1*1 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 1*1 + 0*0 + 2*2)^{0.5} = (6)^{0.5} = 2.245$$

$$\cos(d_1, d_2) = .3150$$

Summary Statistics

- Summary statistics are numbers that summarize properties of the data
 - Summarized properties include frequency, location and spread
 - ◆ Examples: location - mean
spread - standard deviation
 - Most summary statistics can be calculated in a single pass through the data

Frequency and Mode

- The frequency of an attribute value is the percentage of time the value occurs in the data set
 - For example, given the attribute 'gender' and a representative population of people, the gender 'female' occurs about 50% of the time.
- The mode of an attribute is the most frequent attribute value
- The notions of frequency and mode are typically used with categorical data

Percentiles

- For continuous data, the notion of a percentile is more useful.

Given an ordinal or continuous attribute x and a number p between 0 and 100, the p th percentile is a value of x such that $p\%$ of the observed values of x are less than .

- For instance, the 50th percentile is the value such that 50% of all values of x are less than .

Measures of Location : Mean and Median

- The mean is the most common measure of the location of a set of points.
- However, the mean is very sensitive to outliers.
- Thus, the median or a trimmed mean is also commonly used.

$$\text{mean}(x) = \bar{x} = \frac{1}{m} \sum_{i=1}^m x_i$$

$$\text{median}(x) = \begin{cases} x_{(r+1)} & \text{if } m \text{ is odd, i.e., } m = 2r + 1 \\ \frac{1}{2}(x_{(r)} + x_{(r+1)}) & \text{if } m \text{ is even, i.e., } m = 2r \end{cases}$$

Measures of Spread : Range and Variance

- Range is the difference between the max and min
- The variance or standard deviation is the most common measure of the spread of a set of points.

$$\text{variance}(x) = s_x^2 = \frac{1}{m-1} \sum_{i=1}^m (x_i - \bar{x})^2$$

- However, this is also sensitive to outliers, so that other measures are often used.



Measures of Spread : Range and Variance

- Average Absolute Deviation

The average absolute deviation of a set $\{x_1, x_2, \dots, x_n\}$ is

$$\frac{1}{n} \sum_{i=1}^n |x_i - m(X)|.$$

For example, for the data set $\{2, 2, 3, 4, 14\}$:

| Measure of central tendency $m(X)$ | Average absolute deviation |
|------------------------------------|--|
| Mean = 5 | $\frac{ 2 - 5 + 2 - 5 + 3 - 5 + 4 - 5 + 14 - 5 }{5} = 3.6$ |
| Median = 3 | $\frac{ 2 - 3 + 2 - 3 + 3 - 3 + 4 - 3 + 14 - 3 }{5} = 2.8$ |
| Mode = 2 | $\frac{ 2 - 2 + 2 - 2 + 3 - 2 + 4 - 2 + 14 - 2 }{5} = 3.0$ |

Measures of Spread : Range and Variance

- Median Absolute Distance
 - D: 1, 1, 2, 2, 4, 6, 9
 - Median: 2
 - The absolute deviations about 2 are (1, 1, 0, 0, 2, 4, 7) which in turn have a median value of 1 (because the sorted absolute deviations are (0, 0, 1, 1, 2, 4, 7)).
 - So the median absolute deviation for this data is 1.

“

Questions ?