

# DATA MINING 1

## Data Similarity

---

Dino Pedreschi, Riccardo Guidotti

*Revisited slides from Lecture Notes for Chapter 2 “Introduction to Data Mining”, 2nd  
Edition by Tan, Steinbach, Karpatne, Kumar*



UNIVERSITÀ  
DI PISA

# 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 one Attribute

$p$  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           | $d = \frac{ p-q }{n-1}$<br>(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 $   | $s = -d, s = \frac{1}{1+d} \text{ or } s = 1 - \frac{d - \min\_d}{\max\_d - \min\_d}$ |

**Table 5.1.** Similarity and dissimilarity for simple attributes

# Euclidean Distance

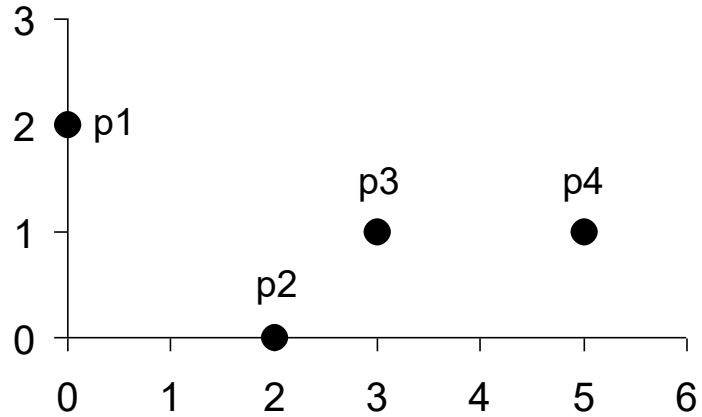
---

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

where  $n$  is the number of dimensions (attributes) and  $x_k$  and  $y_k$  are, respectively, the  $k^{th}$  attributes (components) or data objects  $\mathbf{x}$  and  $\mathbf{y}$ . Standardization is necessary, if scales differ.

- 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 | 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

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

Where  $r$  is a parameter,  $n$  is the number of dimensions (attributes) and  $x_k$  and  $y_k$  are, respectively, the  $k^{\text{th}}$  attributes (components) or data objects  $\mathbf{x}$  and  $\mathbf{y}$ .

# Minkowski Distance: Examples

---

- $r = 1$ . 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,  $L_{\infty}$  norm) distance.
  - This is the maximum difference between any component of the vectors
- 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 | 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     |

| $L_\infty$ | 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, have some well-known properties.

1.  $d(\mathbf{x}, \mathbf{y}) \geq 0$  for all  $\mathbf{x}$  and  $\mathbf{y}$  and  $d(\mathbf{x}, \mathbf{y}) = 0$  only if  $\mathbf{x} = \mathbf{y}$ . (Positive definiteness)
2.  $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$  for all  $\mathbf{x}$  and  $\mathbf{y}$ . (Symmetry)
3.  $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$  for all points  $\mathbf{x}$ ,  $\mathbf{y}$ , and  $\mathbf{z}$ . (Triangle Inequality)

where  $d(\mathbf{x}, \mathbf{y})$  is the distance (dissimilarity) between points (data objects),  $\mathbf{x}$  and  $\mathbf{y}$ .

- A distance that satisfies these properties is a **metric**

# Common Properties of a Similarity

---

Similarities, also have some well-known properties.

1.  $s(\mathbf{x}, \mathbf{y}) = 1$  (or maximum similarity) only if  $\mathbf{x} = \mathbf{y}$ .  
(does not always hold, e.g., cosine)
2.  $s(\mathbf{x}, \mathbf{y}) = s(\mathbf{y}, \mathbf{x})$  for all  $\mathbf{x}$  and  $\mathbf{y}$ . (Symmetry)

where  $s(\mathbf{x}, \mathbf{y})$  is the similarity between points (data objects),  $\mathbf{x}$  and  $\mathbf{y}$ .

# Binary Data

| Categorical | insufficient | sufficient | good | very good | excellent  |
|-------------|--------------|------------|------|-----------|------------|
| p1          | 0            | 0          | 1    | 0         | 0          |
| p2          | 0            | 0          | 1    | 0         | 0          |
| p3          | 1            | 0          | 0    | 0         | 0          |
| p4          | 0            | 1          | 0    | 0         | 0          |
|             |              |            |      |           |            |
| item        | bread        | butter     | milk | apple     | tooth-past |
| p1          | 1            | 1          | 0    | 1         | 0          |
| p2          | 0            | 0          | 1    | 1         | 1          |
| p3          | 1            | 1          | 1    | 0         | 0          |
| p4          | 1            | 0          | 1    | 1         | 0          |
|             |              |            |      |           |            |

# 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

---

$$p = 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0$$

$$q = 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 1$$

$M_{01} = 2$  (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)

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

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

# Document Data

---

|            | team | coach | play | ball | score | game | win | lost | timeout | season |
|------------|------|-------|------|------|-------|------|-----|------|---------|--------|
| Document 1 | 3    | 0     | 5    | 0    | 2     | 6    | 0   | 2    | 0       | 2      |
| Document 2 | 0    | 7     | 0    | 2    | 1     | 0    | 0   | 3    | 0       | 0      |
| Document 3 | 0    | 1     | 0    | 0    | 1     | 2    | 2   | 0    | 3       | 0      |

# Cosine Similarity

- If  $d_1$  and  $d_2$  are two document vectors, then

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

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

- Example:

$$d_1 = \mathbf{3\ 2\ 0\ 5\ 0\ 0\ 0\ 2\ 0\ 0}$$

$$d_2 = \mathbf{1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 2}$$

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

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

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

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

# Using Weights to Combine Similarities

---

- May not want to treat all attributes the same.
  - Use non-negative weights  $\omega_k$

- $similarity(\mathbf{x}, \mathbf{y}) = \frac{\sum_{k=1}^n \omega_k \delta_k s_k(\mathbf{x}, \mathbf{y})}{\sum_{k=1}^n \omega_k \delta_k}$

- Can also define a weighted form of distance

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



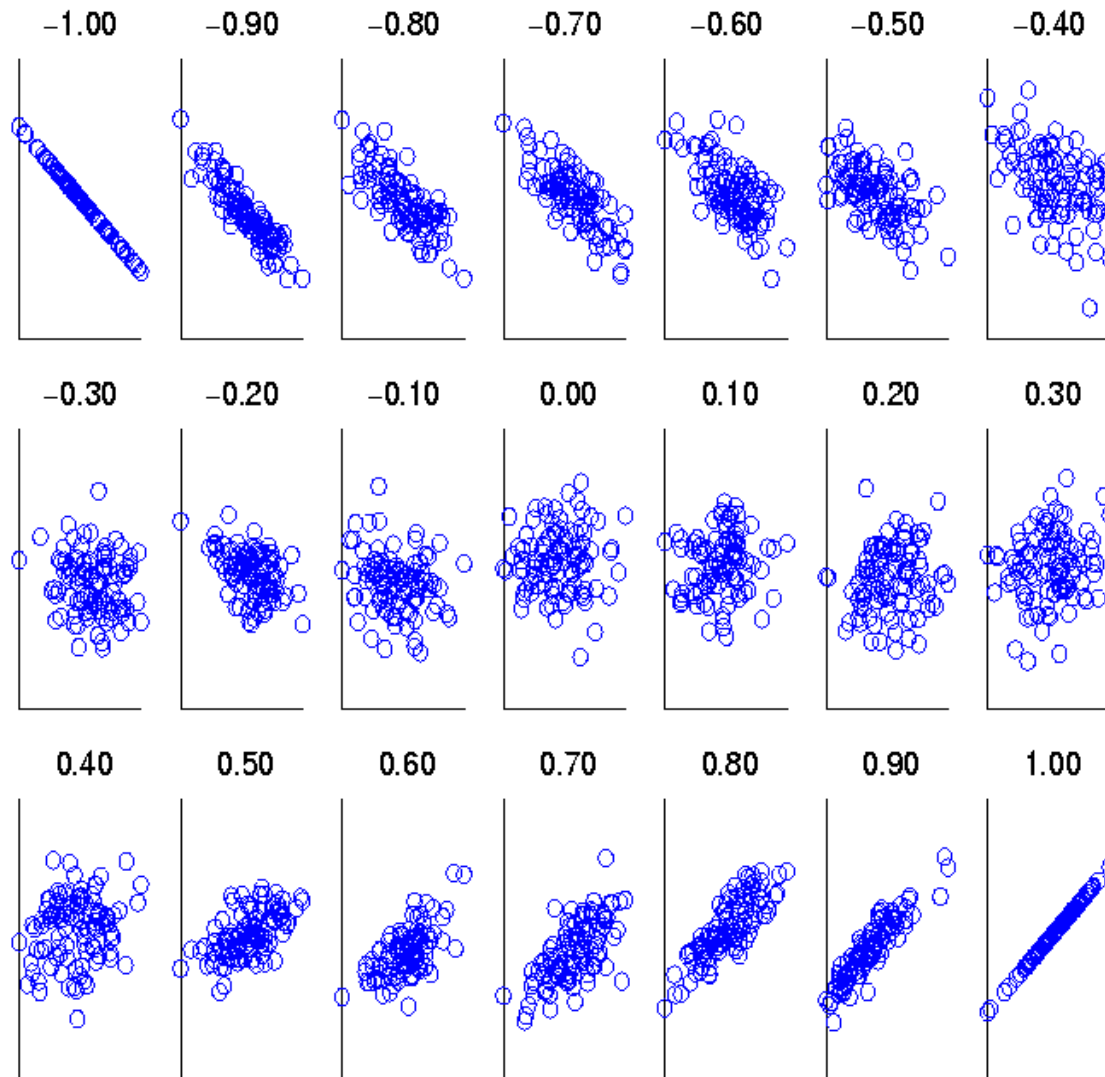
# Correlation

---

- Correlation measures the linear relationship between objects (binary or continuous)
- To compute correlation, we standardize data objects,  $\mathbf{p}$  and  $\mathbf{q}$ , and then take their dot product (covariance/standard deviation)

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

# Visually Evaluating Correlation



**Scatter plots  
showing the  
similarity from  
-1 to 1.**

# Mixed/Heterogenous Distances

---

- What happen if we have data with both continuous and categorical attributes?
- Option 1: discretize continuous attributes and use categorical distances like Jaccard, SMC, etc.
- ~~• Option 2: pretend that categorical attributes can be represented with values and use continuous distances like Euclidean, Manhattan, etc.~~
- Option 3: define a new heterogenous distance like:
- $d(x, y) = n_{\text{cat}}/n \ d_{\text{cat}}(x_{\text{cat}}, y_{\text{cat}}) + n_{\text{con}}/n \ d_{\text{con}}(x_{\text{con}}, y_{\text{con}})$