University of Reading

CSMDM21 - Data Analytics and Mining

Proximity Measures

Module convenor
Dr. Carmen Lam
carmen.lam@reading.ac.uk
Department of Computer Science

Lecture notes and videos created by
Prof. Giuseppe Di Fatta

Overview

- Proximity measures: similarity and dissimilarity
- Transformations
- Proximity between objects with a single attribute
- Proximity between objects with multiple attributes
- Useful proximity measurements
  – Dense data: correlation, Euclidian distance
  – Sparse data: Jaccard and cosine similarity measures

Similarity and Dissimilarity

- Similarity
  - A numerical measure of how alike two data objects are.
  - Higher values indicate the objects are more alike.
  - The values are often defined in the range [0,1]
    - 0: totally different data objects
    - 1: identical data objects

- Dissimilarity
  - A numerical measure of how different two data objects are.
  - Lower values indicate the objects are more alike.
  - Minimum dissimilarity is often 0: identical data objects
  - Upper limit varies (usually 1 or ∞): totally different data objects

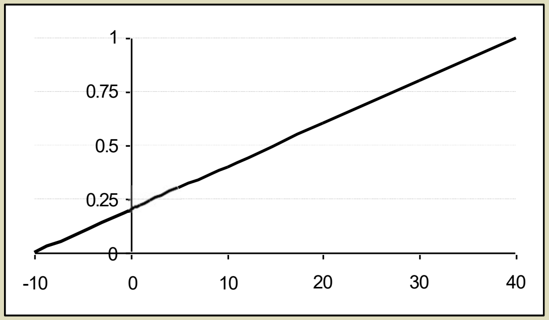
- Proximity generically refers to either similarity or dissimilarity.

Transformations

- Convert similarity to a dissimilarity, or vice versa. If both in the interval [0,1] it is straightforward: s=1-d

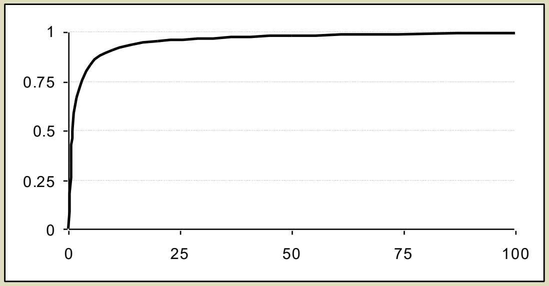
- Transform a proximity measure to fall within a particular range. For example [0,1]:

$$m' = \frac{(m - m_{\min})}{(m_{\max} - m_{\min})}$$



- If the proximity measure is defined in [0, ∞] then a non-linear transformation is needed. For example, for dissimilarity d:

$$d' = \frac{d}{1+d}$$



Similarity/Dissimilarity

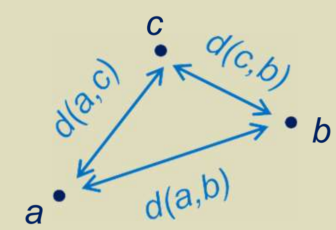
Similarity/Dissimilarity for objects with a single attribute

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	$d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Interval or Ratio	$d = \lvert p - q \rvert$	$s = -d$, $s = \frac{1}{1+d}$ or $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

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

Distance as Dissimilarity Between Objects

- Distances are normally used to measure the dissimilarity between two data objects.
- A distance that satisfies specific properties is a 'metric'.
- A metric function d(a,b) on the pair of points (a,b) has the following properties:
  1. $d(a,b) \geq 0$, $d(a,b) = 0$ if $a=b$ (positive definiteness)
  2. $d(a,b) = d(b,a)$ *(symmetry)*
  3. $d(a,b) \leq d(a,c) + d(c,b)$ *(triangular inequality)*

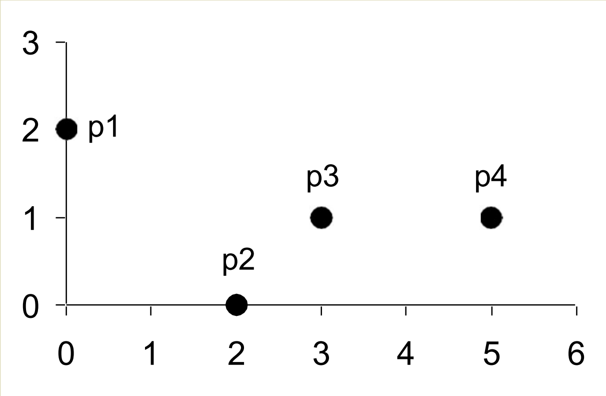


- For example, the Euclidian distance is a metric and is defined as:

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

- 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

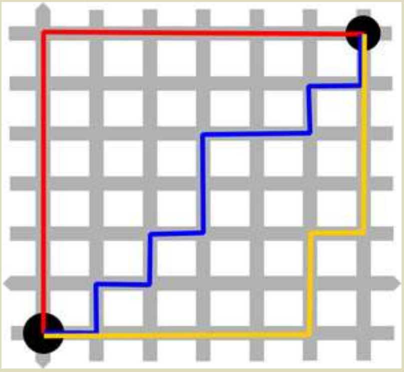
- The Minkowski Distance is a generalization of the Euclidean Distance

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

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

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 (L_2 norm)
- $r \to \infty$. "supremum" (L_{max} norm, L_∞ norm) distance.
  - This is the maximum difference between any component of the vectors
- Do not confuse *r* with *n.*

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

Standardization and Correlation

- Issues with distance measures:
  - Attributes may not have same range of values: “the variables have different scales”. In this case, attributes can be/need to be standardized.
  - Some of the attributed may also be correlated.
- If the data distribution are approximately Gaussian, then the Mahalanobis distance can be used. It is a generalization of the Euclidian distance which takes these issues into account.

Mahalanobis Distance

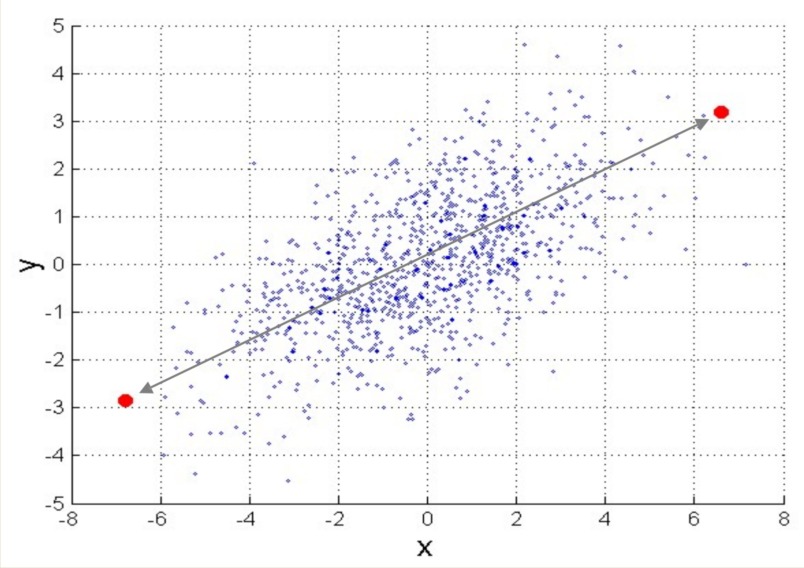
The Statistical, or Mahalanobis, Distance:

$$d(\vec{x},\vec{y})=\sqrt{(\vec{x}-\vec{y})^T\Sigma^{-1}(\vec{x}-\vec{y})}$$

- It is the distance between two multi-dimensional points scaled by the statistical variation in each component.
- Useful for comparing feature vectors whose elements are quantities having different ranges and amounts of variation.
- It also takes into account the correlation among components.

where Σ is the covariance matrix of the input data *X:*

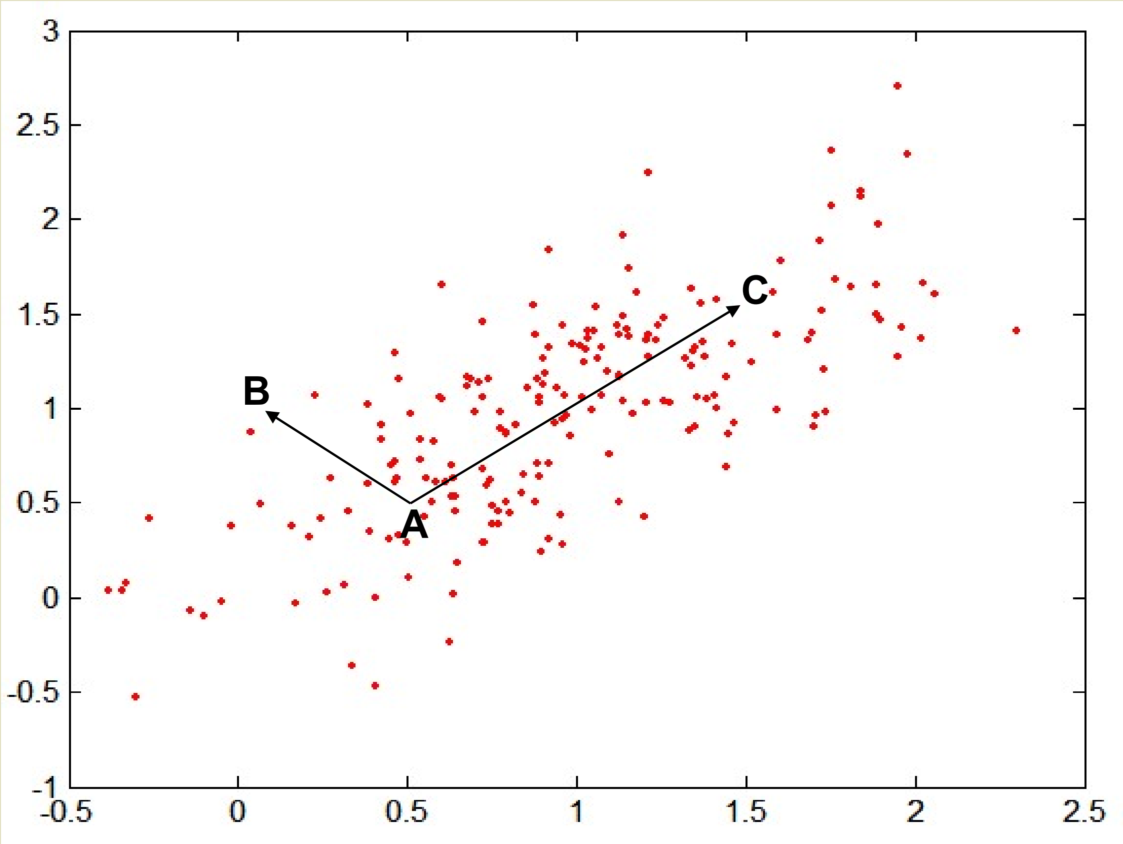
$$\Sigma_{j,k}=\frac{1}{N-1}\sum_{i=1}^{N}(X_{ij}-\overline{X}_j)(X_{ik}-\overline{X}_k)$$



Distance between the two red points:

- the Euclidean distance is 14.7
- the Mahalanobis distance is 6.

Mahalanobis Distance



Covariance Matrix:

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

A: (0.5, 0.5)

B: (0, 1)

C: (1.5, 1.5)

D_E(A,B) = 0.7

D_E(A,C) = 1.4

D_M(A,B) = 5

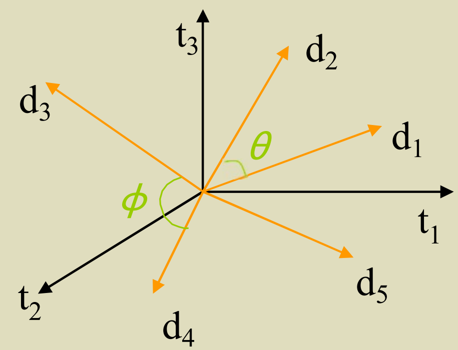
D_M(A,C) = 4

Cosine Similarity

- If d_1 and d_2 are two vectors, then

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

where "•" indicates **vector dot product** and $||d||$ is the length of vector d.



- Note: this is *similarity*, not distance. No triangle inequality for similarity.
- Example:

d_1 = **3 2 0 5 0 0 0 2 0 0**

d_2 = **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

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

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

$$\cos(d_1, d_2) = 0.3150$$

- Other (Dis-)Similarity measures: Pearson Correlation, Extended Jaccard Similarity, etc.

Common Properties of a Similarity

- Similarities also have some well-known properties:

  1. $s(p, q) = 1$ (or maximum similarity) only if $p = q$
  2. $s(p, q) = s(q, p)$ for all p and q (Symmetry)

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

Similarity Between Binary Vectors

- If data objects have only binary attributes, they are represented as binary vectors. For binary vectors *p* and *q, t*heir similarity is computed 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

- The Simple Matching Coefficient (SMC)
  SMC = number of matches / number of attributes = $(M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$

- The Jaccard index, aka the Jaccard Similarity Coefficient
  The Jaccard index (originally coined 'coefficient de communauté' by Paul Jaccard) is a statistic used for comparing the similarity between sample sets (e.g., A={a,b,c} and B={a,c,d,e}).
  The Jaccard coefficient is defined as the size of the intersection divided by the size of the union of the sample sets:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

  For two binary vectors p and q:
  J(p,q) = number of "1-1" 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
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)

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

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

Hamming Distance

- In **Information Theory**, the Hamming distance between two strings of equal length is **the number of positions at which the corresponding symbols are different**.
- It measures the minimum number of substitutions required to change one string into the other, or the number of errors that transformed one string into the other.

- For binary data it corresponds to the L1 distance.
  - Hamming = $(M_{10} + M_{01})/(M_{01} + M_{10} + M_{11} + M_{00})$

- It's a distance, while SMC and Jaccard coefficients are similarities:
  - d = 1-s
  - In particular, Hamming = 1-SMC

- Example: the Hamming distance can be used as a measure of genetic distance.

Extended Jaccard Coefficient (Tanimoto)

- **Tanimoto coefficient**: T(p, q)

$$T(p,q) = \frac{p \bullet q}{\|p\|^2 + \|q\|^2 - p \bullet q}$$

- It is the extension of the Jaccard coefficient for continuous or count attributes (it reduces to Jaccard for binary attributes).

Correlation

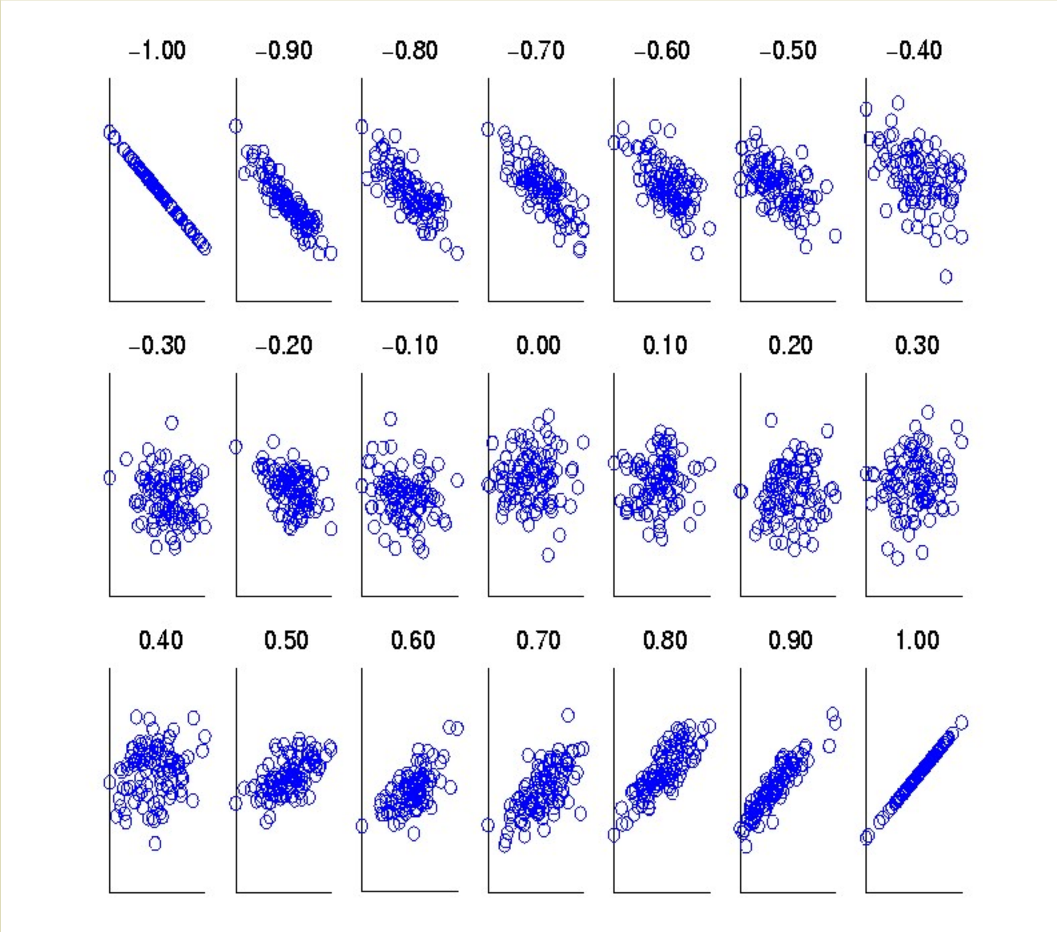
- Correlation measures the linear relationship between objects
- To compute correlation, we standardize data objects, p and q, and then take their dot product

$$p'_k = (p_k - mean(p)) / std(p)$$

$$q'_k = (q_k - mean(q)) / std(q)$$

$$correlation(p, q) = p' \bullet q'$$

Visually Evaluating Correlation



Scatter plots showing the similarity from –1 to 1.

Summary Table for Proximity Measures

Proximity measure	M_{00}	M_{11}	M_{10}	M_{01}	*Binary /continuous attributes*
SMC	√	√			b
Jaccard similarity coefficent		√			b
Hamming distance			√	√	b
Cosine similarity		√			c
Tanimoto coefficient		√			c

Next:

- P02: practical on Data Processing in KNIME

Next week:

- Clustering