Data Mining: Similarity Measures

Tom Claassen

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 two data objects are.
  - 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

- Nominal attributes p and q

$$d = \begin{cases} 0, & p = q \\ 1, & p \neq q \end{cases} \qquad s = 1 - d$$

- Ordinal attributes: map *n* distinct values to integers from 0 to $n - 1$

$$d = \frac{|p - q|}{n - 1} \qquad s = 1 - d$$

- Interval or ratio attributes: $d = |p - q| \qquad s = \frac{1}{1+d}$

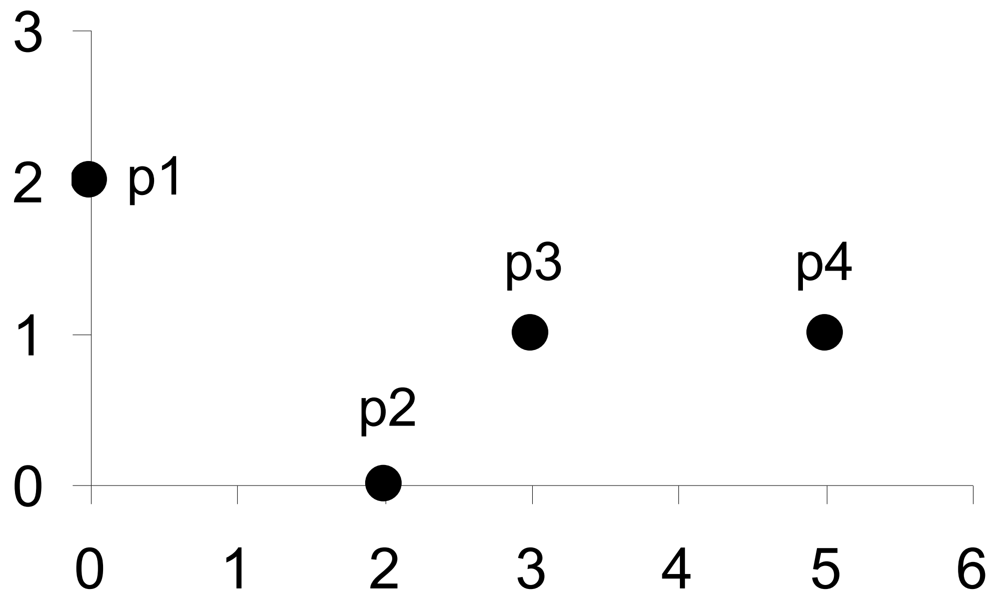
Euclidean distance

- Euclidean distance between two objects $\mathbf{p}$ and $\mathbf{q}$ with n attributes

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

- Distance of a ruler
- 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

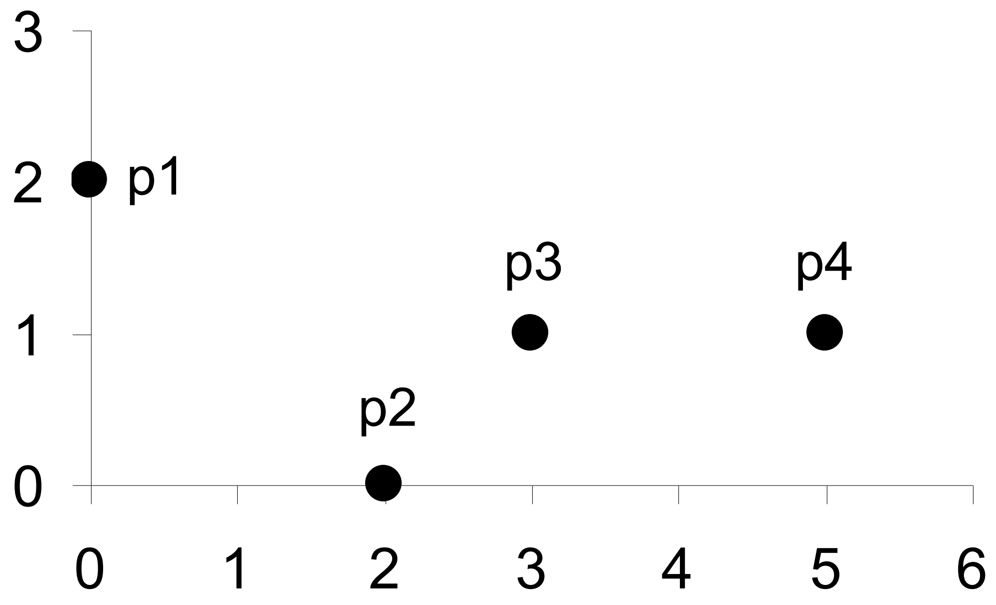
Minkowski distance

- Generalization of Euclidean distance:

$$d(\mathbf{p},\mathbf{q}) = \sqrt[r]{\sum_{k=1}^{n} |p_k - q_k|^r}$$

- $r = 1$: city block (Manhattan, taxicab, L_1 norm) distances
  - reduces to Hamming distance, which just counts the number of differences, in case of binary variables
- $r = 2$ corresponds to Euclidean distance
- $r = \infty$: supremum (L_{max}, L_∞) or Chebyshev distance
  - maximum distance between any component of the vectors
  - distance kings have to travel on a chess board

Manhattan Distance ($r = 1$)

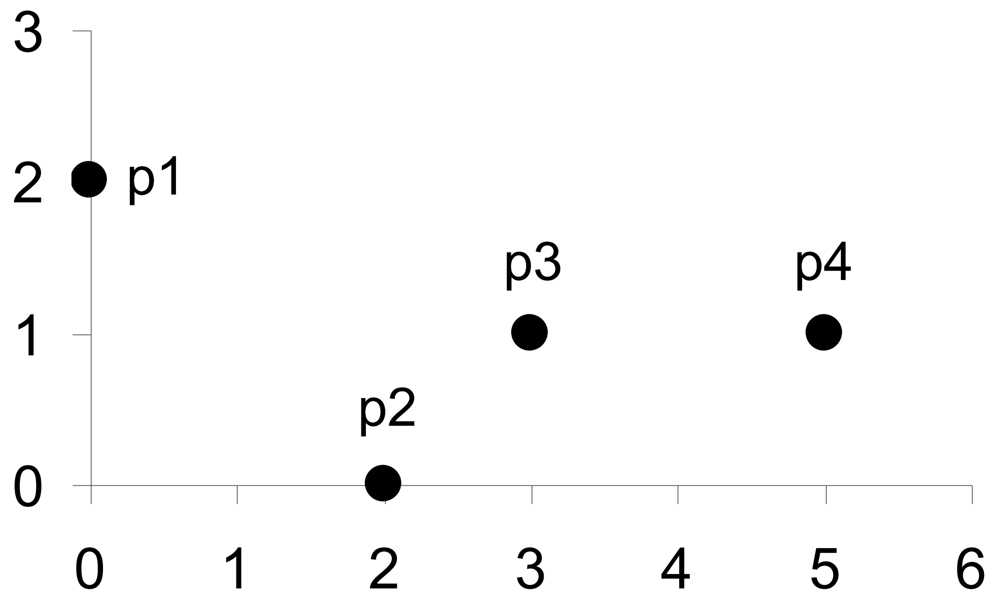


point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

distance matrix

Supremum Distance ($r = \infty$)

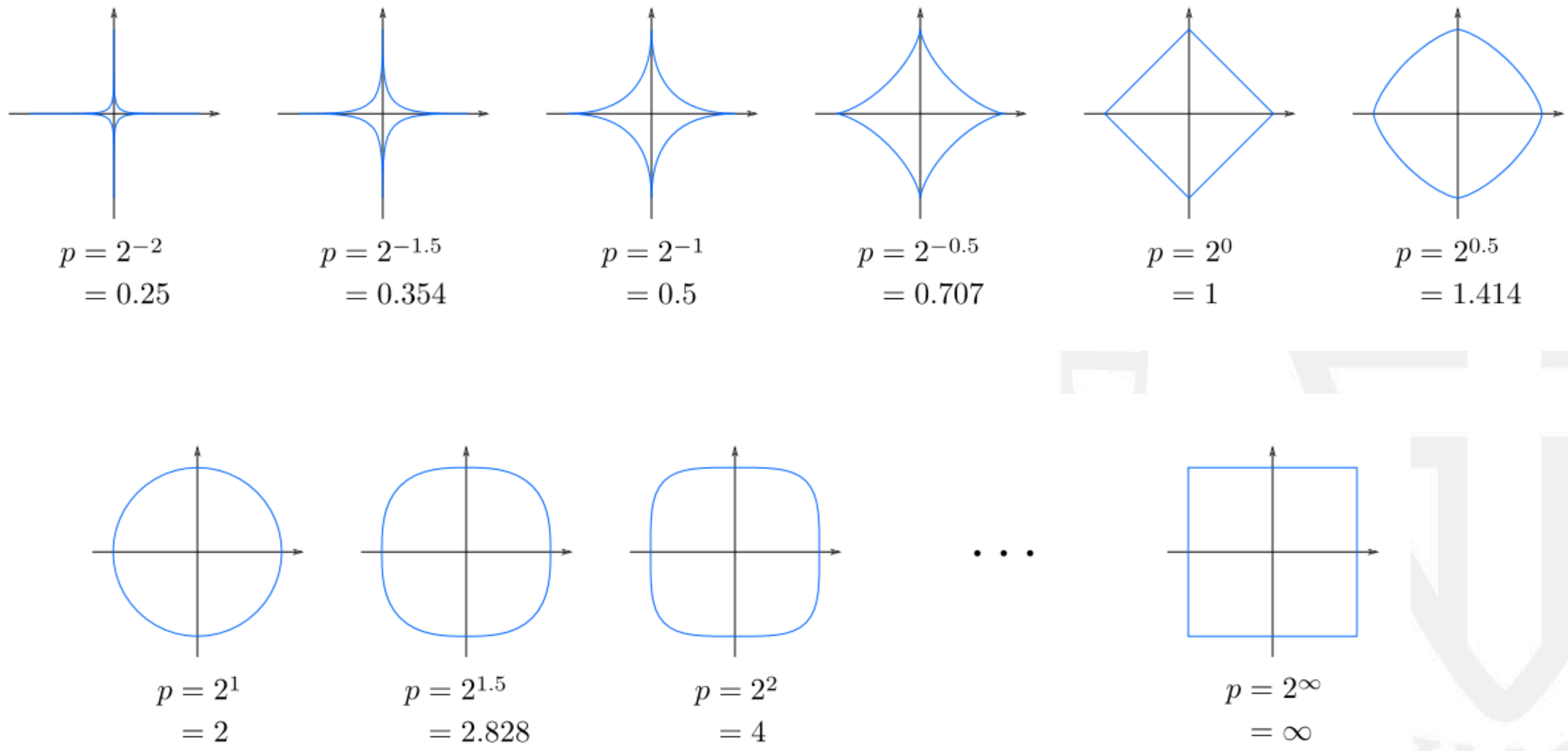


point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	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

Unit circle



$p = 2^{-2}$
$= 0.25$
$p = 2^{-1.5}$
$= 0.354$
$p = 2^{-1}$
$= 0.5$
$p = 2^{-0.5}$
$= 0.707$
$p = 2^{0}$
$= 1$
$p = 2^{0.5}$
$= 1.414$
$p = 2^{1}$
$= 2$
$p = 2^{1.5}$
$= 2.828$
$p = 2^{2}$
$= 4$
$\cdots$
$p = 2^{\infty}$
$= \infty$

Common properties of a distance

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

  1. Positive definiteness: $d(\mathbf{p}, \mathbf{q}) \geq 0$ for all $\mathbf{p}$ and $\mathbf{q}$ and $d(\mathbf{p}, \mathbf{q}) = 0$ iff $\mathbf{p} = \mathbf{q}$
  2. Symmetry: $d(\mathbf{p}, \mathbf{q}) = d(\mathbf{q}, \mathbf{p})$ for all $\mathbf{p}$ and $\mathbf{q}$
  3. Triangle inequality: $d(\mathbf{p}, \mathbf{r}) \leq d(\mathbf{p}, \mathbf{q}) + d(\mathbf{q}, \mathbf{r})$ for all $\mathbf{p}, \mathbf{q}$ and $\mathbf{r}$

- A distance that satisfies these properties is called a metric

Similarity Between Binary Vectors

		$\mathbf{p}$	
		0	1
$\mathbf{q}$	0	M_{00}	M_{10}
	1	M_{01}	M_{11}

- M_{00} = number of attributes with $p_k = 0$ and $q_k = 0$, etc.
- Simple matching coefficient (SMC):

$$s(\mathbf{p}, \mathbf{q}) = \frac{\#\text{ matches}}{\#\text{ attributes}} = \frac{M_{00} + M_{11}}{M_{00} + M_{01} + M_{10} + M_{11}}$$

- Jaccard coefficient:

$$s(\mathbf{p}, \mathbf{q}) = \frac{\#\text{ 11 matches}}{\#\text{ not−both−zero}} = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}$$

SMC versus Jaccard

		$\mathbf{p}$	
		0	1
$\mathbf{q}$	0	7	1
	1	2	0

- $\mathbf{p} = [1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0]$
- $\mathbf{q} = [0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 1]$

- Simple matching coefficient (SMC):

$$s(\mathbf{p}, \mathbf{q}) = \frac{\#\text{ matches}}{\#\text{ attributes}} = \frac{7}{10} = 0.7$$

- Jaccard coefficient:

$$s(\mathbf{p}, \mathbf{q}) = \frac{\#\ 11\text{ matches}}{\#\text{ not}-\text{both}-\text{zero}} = \frac{0}{3} = 0$$

Cosine Similarity

- Specifically for documents vectors

$$s(\mathbf{p}, \mathbf{q}) = \frac{\mathbf{p} \cdot \mathbf{q}}{\|\mathbf{p}\| \|\mathbf{q}\|}$$

- With inner product

$$\mathbf{p} \cdot \mathbf{q} = \sum_{k=1}^{n} p_k q_k$$

- And vector length

$$\|\mathbf{p}\| = \sqrt{\mathbf{p} \cdot \mathbf{p}}$$

Cosine Similarity Example

Document	team	coach	play	ball	score	game	win	lost	timeout	season
1	3	2	0	5	0	0	0	2	0	0
2	0	7	0	2	1	0	0	3	0	0
3	1	0	0	0	0	0	0	1	0	2
4	1	4	0	2	3	0	1	6	2	1
5	2	3	3	1	6	1	3	0	0	4

- $\mathbf{p} = [3\ 2\ 0\ 5\ 0\ 0\ 0\ 2\ 0\ 0]$
- $\mathbf{q} = [1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 2]$

- Inner product:

$$\mathbf{p} \cdot \mathbf{q} = 3 \times 1 + 2 \times 0 + 0 \times 0 + 5 \times 0 + 0 \times 0 + 0 \times 0 + 0 \times 0 + 2 \times 1 + 0 \times 0 + 0 \times 2 = 5$$

- Vector lengths: $\|\mathbf{p}\| = \sqrt{3^2 + 2^2 + 5^2 + 2^2} = \sqrt{42}$ and $\|\mathbf{q}\| = \sqrt{1^2 + 1^2 + 2^2} = \sqrt{6}$

- Cosine similarity:

$$s(\mathbf{p}, \mathbf{q}) = \frac{\mathbf{p} \cdot \mathbf{q}}{\|\mathbf{p}\| \|\mathbf{q}\|} = \frac{5}{\sqrt{252}}$$