Scales

scale	operations		example	statistics
ratio	·	/	21 years, $273°K$ _(defined zero)_	generalized mean
interval	+	−	2015 A.D., $20°C$ _(distances, no-defined zero)_	mean
ordinal	>	<	A, B, C, D, F _(grades)_	median
nominal	=	≠	Alice, Bob, Carol	mode _(names, id ...)_

4 different scales

equality / inequality

names that occurs more often (frequency)

Prof. Dr. Thomas A. Runkler

Matrix Representation of Numerical Data

- numerical data set $X = \{x_1, \ldots, x_n\} \subset \mathrm{IR}^p$

feature vector x_k

data point x_k

data (set) X

i^{th} feature $x^{(i)}$

i^{th} component $x^{(i)}$

datum $x_k^{(i)}$

value $x_k^{(i)}$
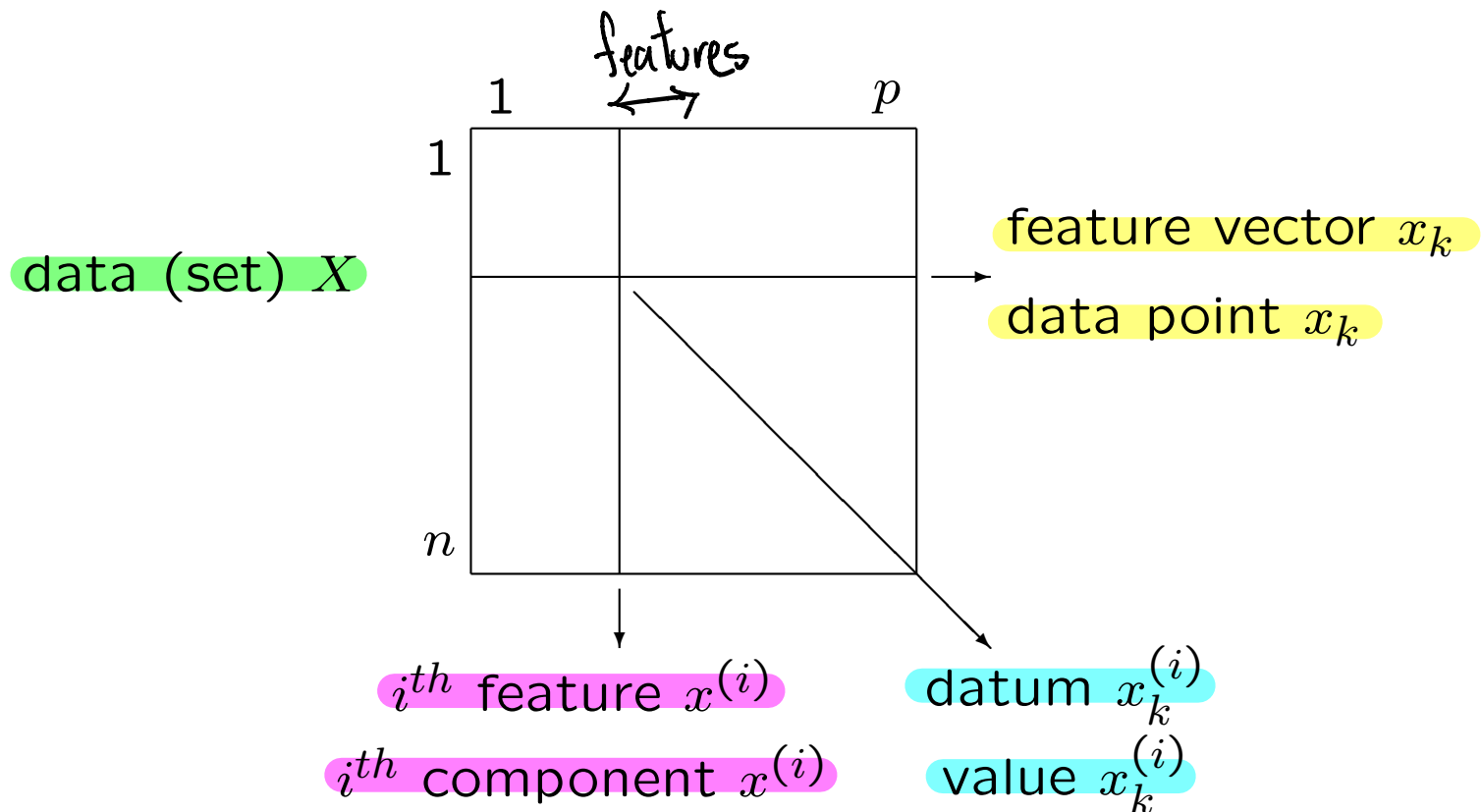
Relations on Object Data

- object data $\qquad O = \{o_1, \ldots, o_n\}$

- relational data $\qquad R = \begin{pmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{n1} & \cdots & r_{nn} \end{pmatrix} \in \mathrm{IR}^{n \times n}$

- semantics r_{ij}: similarity / dissimilarity

- symmetric relation: $r_{ij} = r_{ji} \ \forall i, j \in \{1, \ldots, n\}$

- for $O \subset \mathrm{IR}^p$: R can be obtained by norm on IR^p

Dissimilarity/Distance Measures

- Properties:

$$
\begin{aligned}
d(x,y) &= d(y,x) \\
d(x,y) &= 0 \quad \Leftrightarrow \quad x = y \\
d(x,z) &\leq d(x,y) + d(y,z)
\end{aligned}
$$

$$d \longrightarrow \text{dissimilarity}$$

Prof. Dr. Thomas A. Runkler

Norms

- $\|.\| : \mathrm{IR}^p \to \mathrm{IR}^+$ is a norm, iff

  1. $\|x\| = 0 \Leftrightarrow x = (0, \dots, 0)$
  2. $\|a \cdot x\| = |a| \cdot \|x\| \quad \forall a \in \mathrm{IR}, x \in \mathrm{IR}^p$
  3. $\|x + y\| \leq \|x\| + \|y\| \quad \forall x, y \in \mathrm{IR}^p$

- "hyperbolic norm"

$$\|x\|_h = \prod_{i=1}^{p} x^{(i)}$$

  ist not a norm!

Prof. Dr. Thomas A. Runkler

Matrix Norms

$$\|x\|_A = \sqrt{xAx^T}$$

Examples:

- Euclidean:
$$A = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

- diagonal:
$$A = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_p \end{pmatrix}$$

- Mahalanobis:
$$A = \mathrm{cov}^{-1}(X) = \left(\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})^T (x_k - \bar{x}) \right)^{-1}$$
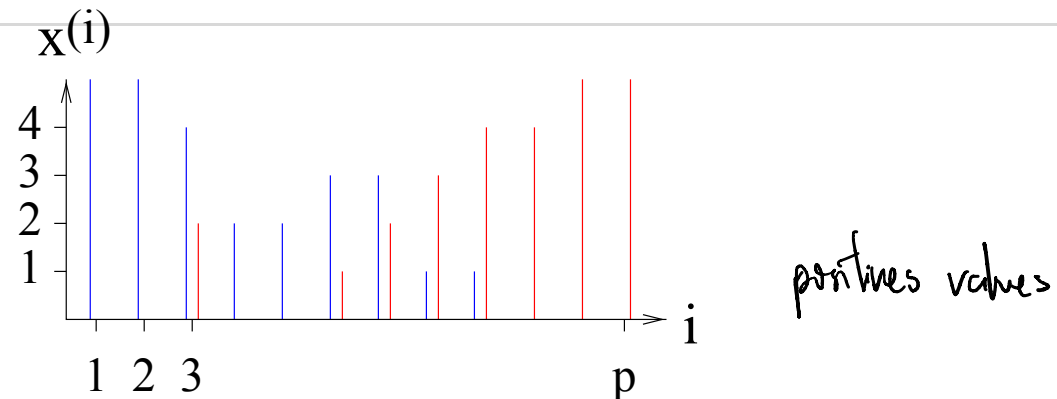
Minkowski/Lebesgue Norms

$$\|x\|_\alpha = \sqrt[\alpha]{\sum_{j=1}^{p} \left|x^{(j)}\right|^\alpha}$$

Examples:

- Manhattan or city block: $\alpha = 1$

- Euclidean: $\alpha = 2$

- sup or max: $\alpha \to \infty$

$$\lim_{\alpha \to \infty} \sqrt[\alpha]{\sum_{j=1}^{p} \left|x^{(j)}\right|^\alpha} = \max_{j=1,\dots,p} |x^{(j)}|$$

Similarity/Proximity Measures

$x^{(i)}$

positives values

- Properties:

$$s(x, y) = s(y, x) \quad \text{simetry}$$
$$s(x, y) \leq s(x, x)$$
$$s(x, y) \geq 0$$

- normalized similarity measures:

$$s(x, x) = 1$$

Similarity Measures $(x^{(i)}, y^{(i)} \geq 0)$

- cosine

$$s(x,y) = \frac{\sum\limits_{i=1}^{p} x^{(i)} y^{(i)}}{\sqrt{\sum\limits_{i=1}^{p} \left(x^{(i)}\right)^2 \sum\limits_{i=1}^{p} \left(y^{(i)}\right)^2}}$$

- overlap

$$s(x,y) = \frac{\sum\limits_{i=1}^{p} x^{(i)} y^{(i)}}{\min\left(\sum\limits_{i=1}^{p} \left(x^{(i)}\right)^2, \sum\limits_{i=1}^{p} \left(y^{(i)}\right)^2\right)}$$

Similarity Measures $(x^{(i)}, y^{(i)} \geq 0)$

- Dice

$$s(x, y) = \frac{2 \sum\limits_{i=1}^{p} x^{(i)} y^{(i)}}{\sum\limits_{i=1}^{p} \left(x^{(i)}\right)^2 + \sum\limits_{i=1}^{p} \left(y^{(i)}\right)^2}$$

- Jaccard/Tanimoto

$$s(x, y) = \frac{\sum\limits_{i=1}^{p} x^{(i)} y^{(i)}}{\sum\limits_{i=1}^{p} \left(x^{(i)}\right)^2 + \sum\limits_{i=1}^{p} \left(y^{(i)}\right)^2 - \sum\limits_{i=1}^{p} x^{(i)} y^{(i)}}$$