Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Measuring distances

Applied multivariate statistics – Spring 2012

Overview

- Distances between samples or variables?
- Scaling gives equal weight to all variables
- Dissimilarity is a generalization of Distance
- Dissimilarity for different data types:
  - interval scaled
  - binary (symmetric / asymmetric)
  - nominal
  - ordinal
  - mixed

Different perspective of one thing

- Data context (e.g. biologist, doctor, …) determines distance measure, not statistician
- In practice: Statistician has to offer choices with pros and cons

Between samples or variables?

Rest of this lecture

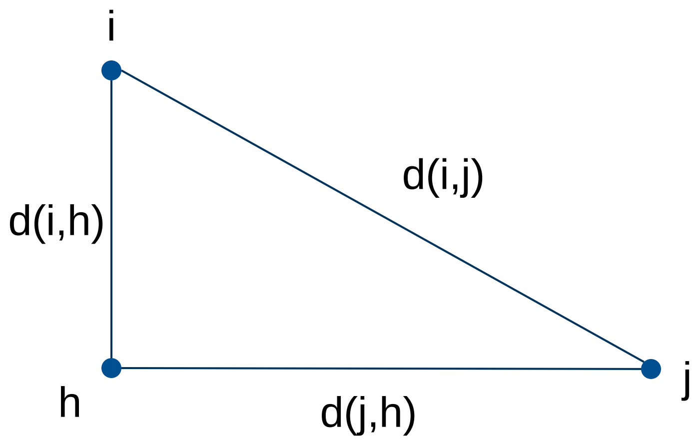
X_1	X_2	X_3
2.5	3.4	1.6
4.3	5.3	5.3
6.3	9.4	8.9

Use correlation

$$d(X_i, X_j) = \frac{1 - Cor(X_i, X_j)}{2}$$

Properties of distance measures

- D1: d(i,j) >= 0
- D2: d(i,i) = 0
- D3: d(i,j) = d(j,i)
- D4: d(i,j) <= d(i,h) + d(h,j) (triangle inequality)



Examples

- Euclidean distance:

$$d(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + ... + (x_{ip} - x_{jp})^2}$$

- Manhattan distance:

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + ... + |x_{ip} - x_{jp}|$$

- Maximum distance:

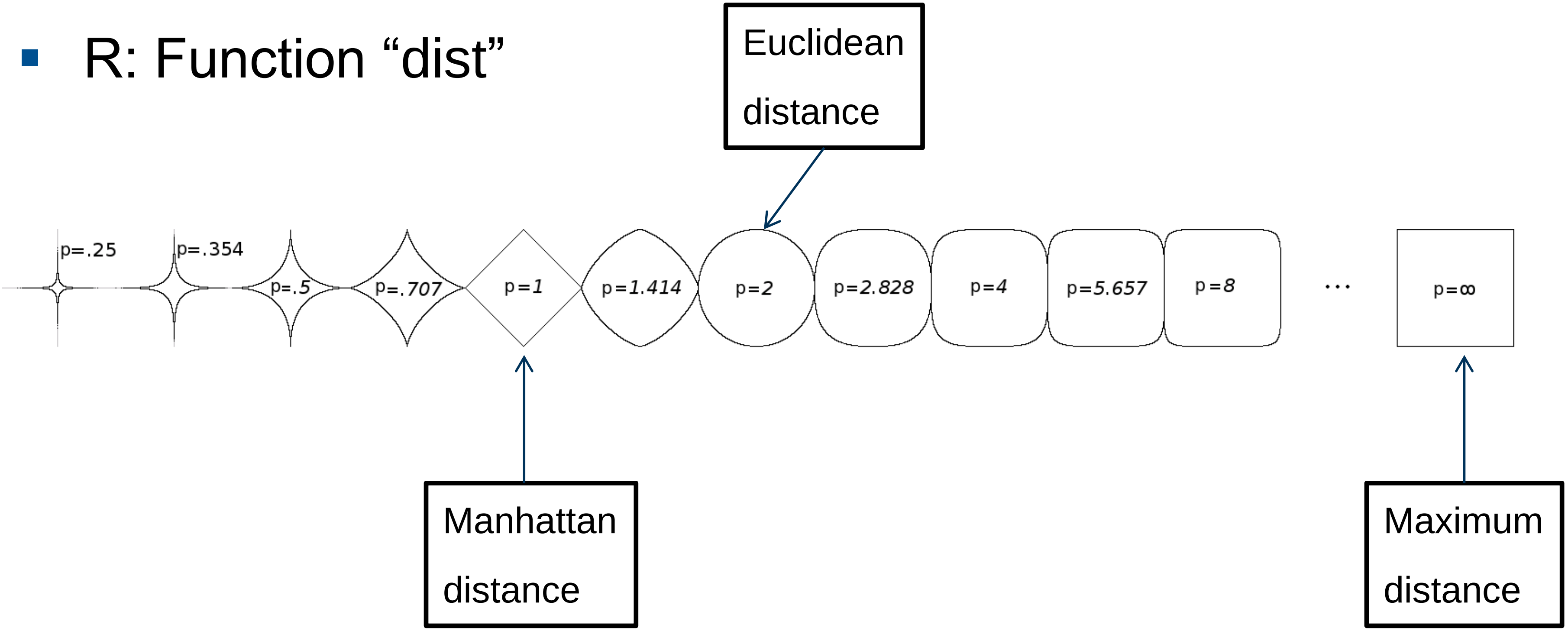
$$\begin{aligned} d(i,j) &= (|x_{i1} - x_{j1}|^{\infty} + |x_{i2} - x_{j2}|^{\infty} + ... + |x_{ip} - x_{jp}|^{\infty})^{\frac{1}{\infty}} = \\ &= max_{k=1}^{p} |x_{ik} - x_{jk}| \end{aligned}$$

- Special cases of Minkowski distance:

$$d(i,j) = (|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + ... + |x_{ip} - x_{jp}|^q)^{\frac{1}{q}}$$

Intuition for Minkowski Distance

- p: Index of Minkowski Distance
- Points on the line have equal Minkowski Distance from center
- R: Function “dist”



Distance metrics in practice

- Euclidean Distance: By far most common
  Our intuitive notion of distance
- Manhattan Distance: Sometimes seen

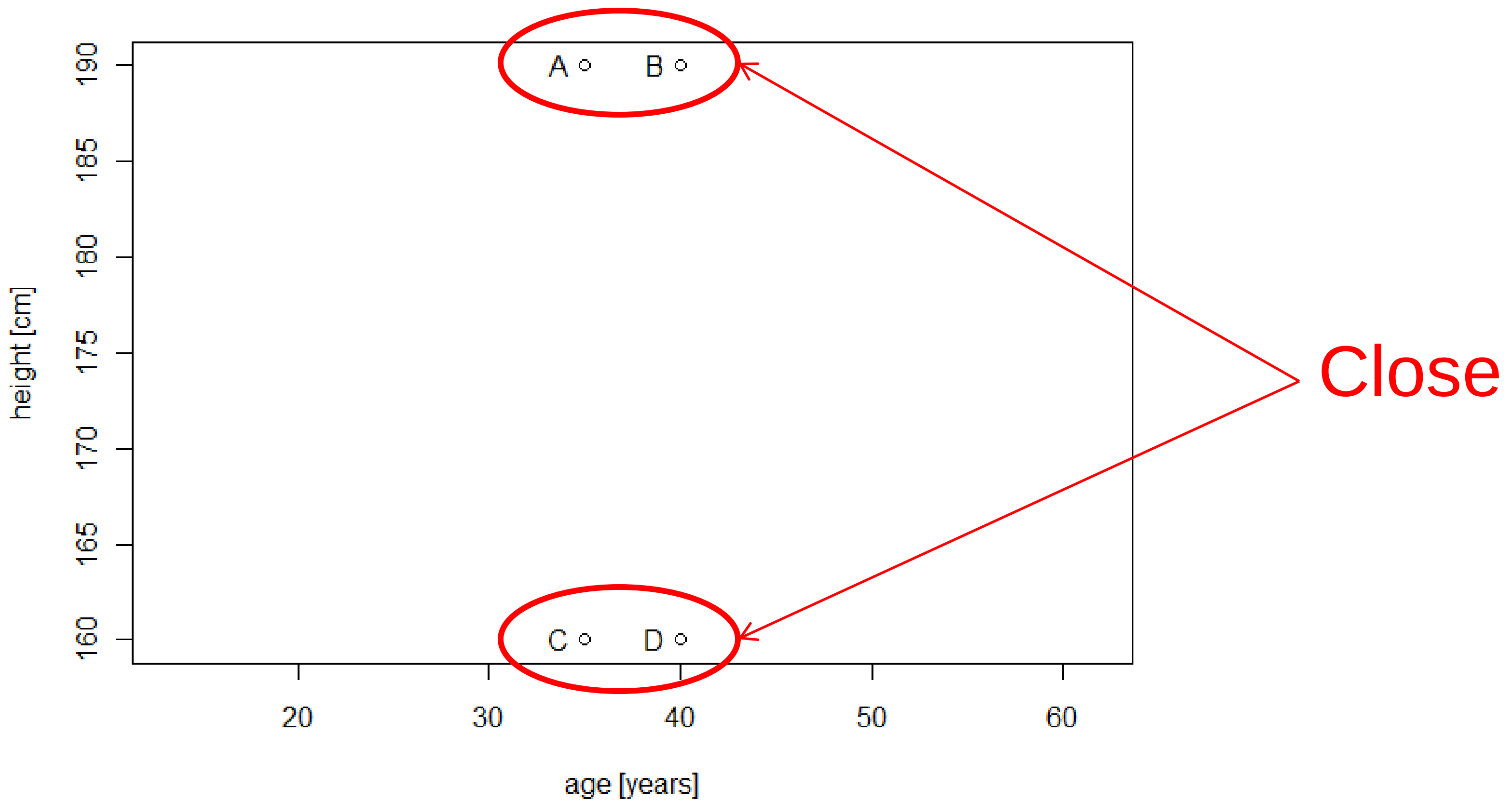
- Rest: Very rare

To scale or not to scale…

Example 1: cm

- 4 persons

Person	Age [years]	Height [cm]
A	35	190
B	40	190
C	35	160
D	40	160



ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Example 1: feet

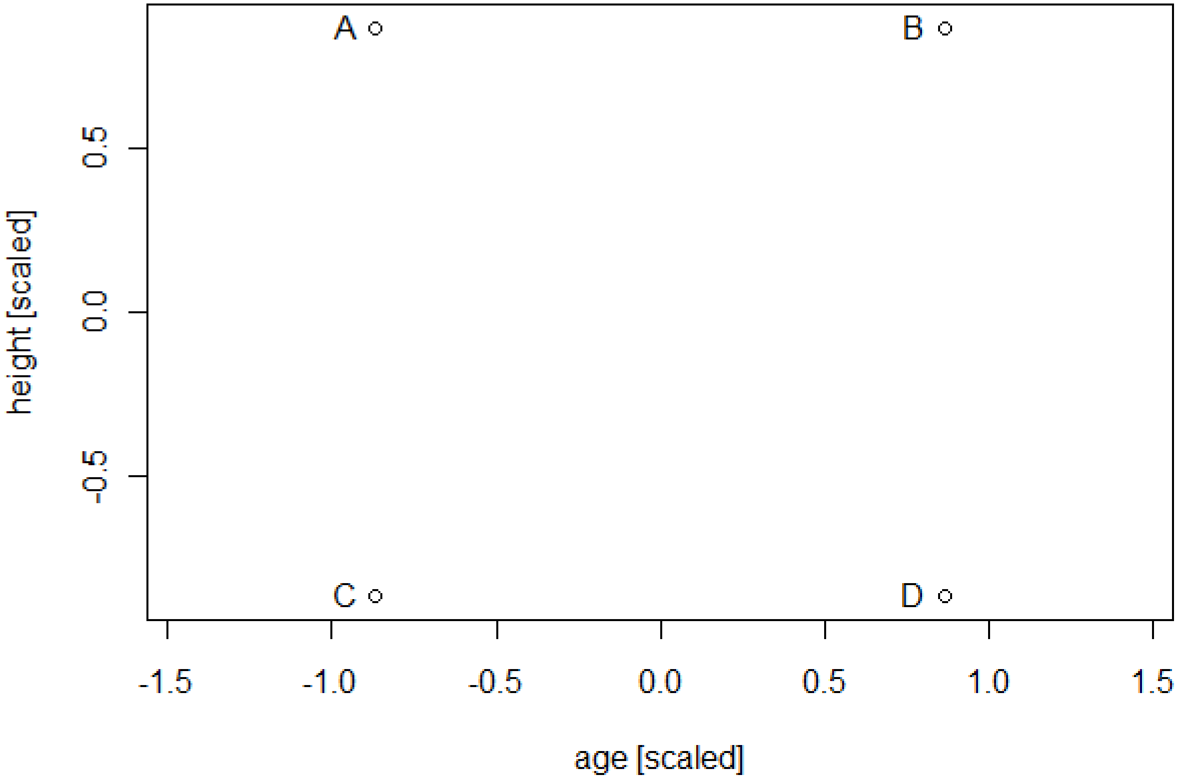
Person	Age [years]	Height [feet]
A	35	6.232
B	40	6.232
C	35	5.248
D	40	5.248

- 4 persons

Example 1: scaled

- 4 persons

Person	Age [scaled]	Height [scaled]
A	-0.87	0.87
B	0.87	0.87
C	-0.87	-0.87
D	0.87	-0.87



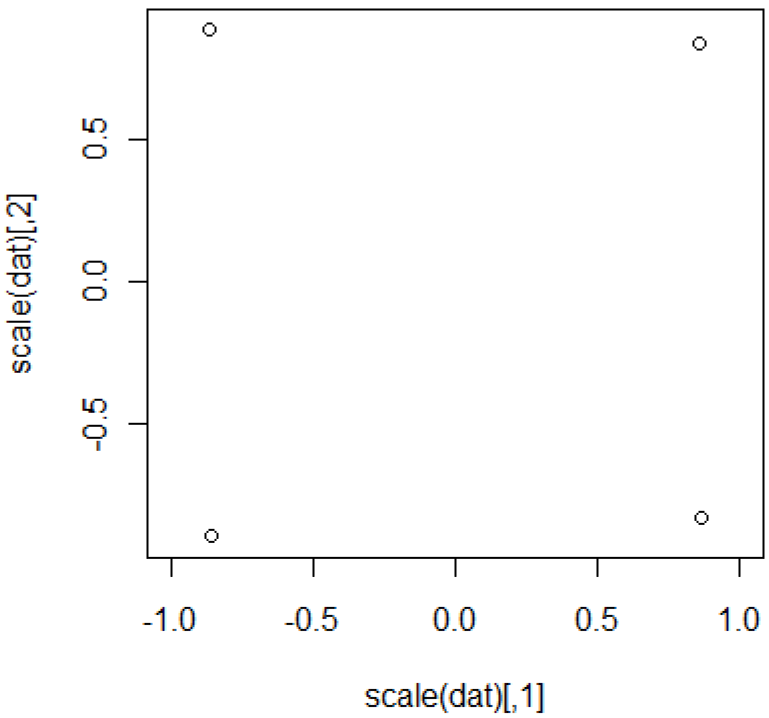
No subgroups anymore

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

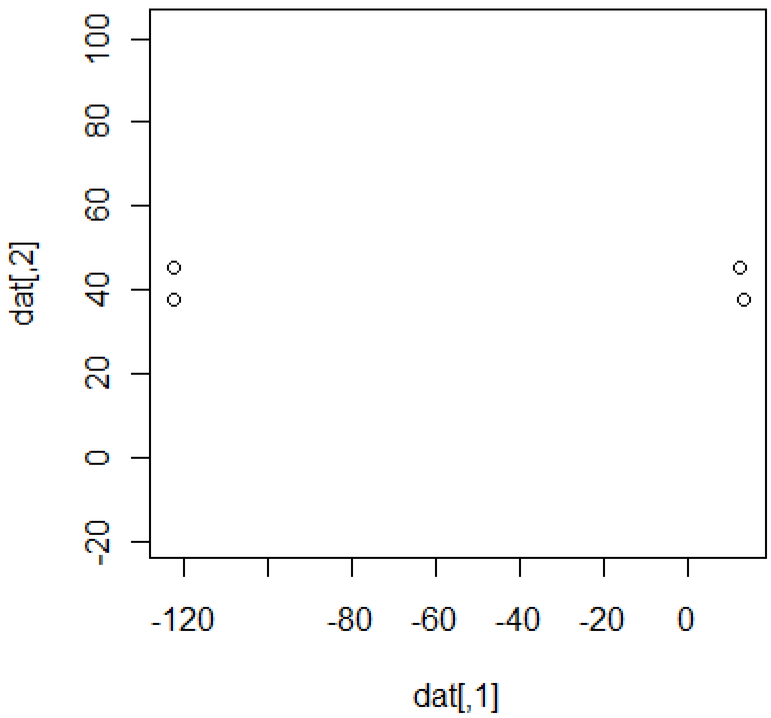
Example 2

- 4 objects

Object	x1	x2
A	13.3	38.0
B	12.4	45.4
C	-122.7	45.6
D	-122.4	37.7



OR

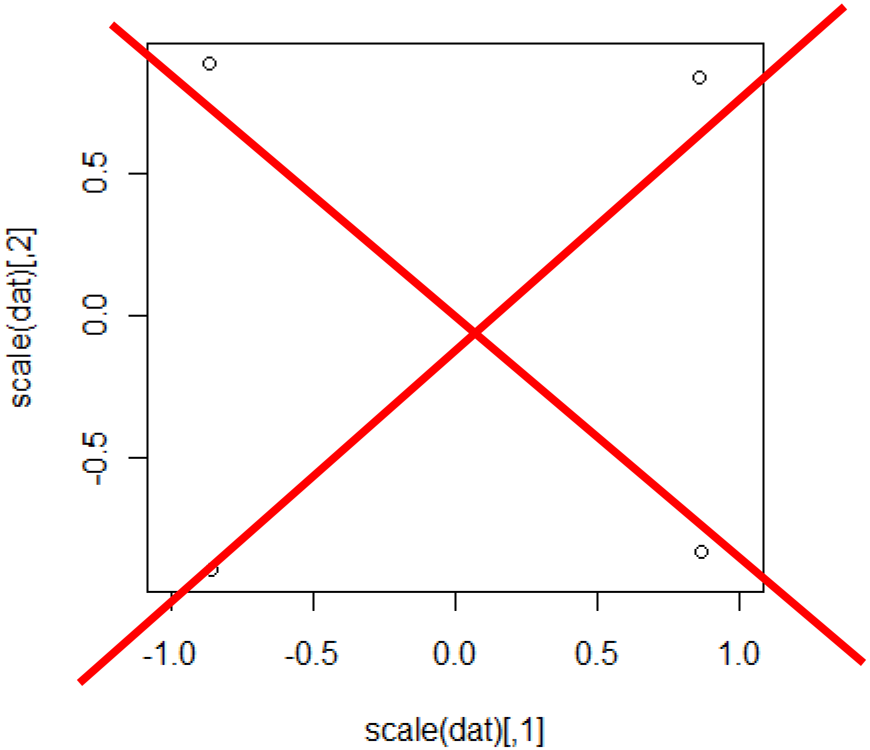


Need knowledge of context

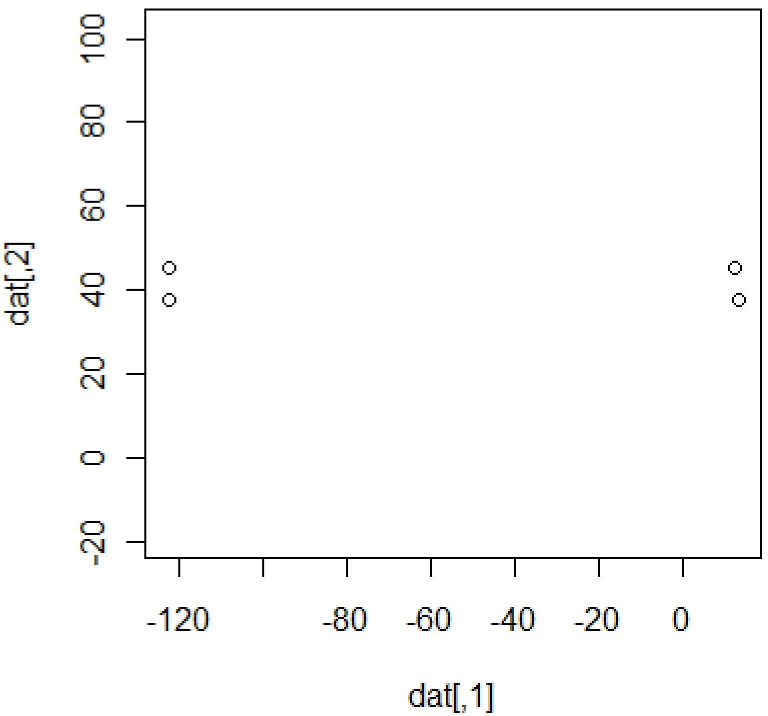
Example 2

- 4 objects

Object	Long.	Lat.
Palermo	13.3	38.0
Venice	12.4	45.4
Portland	-122.7	45.6
San Francisco	-122.4	37.7



OR



To scale or not to scale…

- If variables are not scaled
  - variable with largest range has most weight
  - distance depends on scale
- Scaling gives every variable equal weight
- Similar alternative is re-weighing:

$$d(i,j) = \sqrt{w_1(x_{i1} - x_{j1})^2 + w_2(x_{i2} - x_{j2})^2 + ... + w_p(x_{ip} - x_{jp})^2}$$

- Scale if,
  - variables measure different units (kg, meter, sec,…)
  - you explicitly want to have equal weight for each variable
- Don't scale if units are the same for all variables
- Most often: Better to scale.

Dissimilarities

- More flexible than distances

  D1: d(i,j) >= 0

  D2: d(i,i) = 0

  D3: d(i,j) = d(j,i)

	M	P	H
M	10	1	8
P		10	5
H			10

- Example: What do you think, how different are the topics Mathematics, Physics, History on a scale from 0 to 10 (very different)?
- Could also work with “Similarities” (e.g. 1-Dissimilarity)

Dissimilarities for different data types

- Interval-scaled:
  - continuous, positive or negative
  - examples: height, weight, temperature, age, cost,...
  Difference of values has a fixed interpretation
  - use metrics we just discussed
- Ratio-scaled:
  - continuous, positive
  - example: concentration
  Ratio of values has fixed interpretation
  - use log-transformation, then metrics we just discussed
- R:
  - Function “dist” in base distribution (includes Minkowski)
  - Function “daisy” in package “cluster”

Binary symmetric: Simple matching coefficient

- “Symmetric”: No clear asymmetry between group 0 and group 1
- Example: Gender, Right-handed
  Two right-handed people are as similar as two left-handed people
- Counter-example: Having AIDS, being Nobel Laureate
  Two Nobel Laureates are more similar than two non-Nobel-Laureates (e.g. Uni Prof at Harvard without Nobel Prize and baby from Sudan)

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Binary symmetric: Simple matching coefficient

Object j

Object i

	X=1	X=0
X=1	a	b
X=0	c	d

a+b+c+d = Number of variables

Simple matching coefficient

$$d(i,j) = \frac{b+c}{a+b+c+d}$$

Proportion of variables,
in which people disagree

Binary asymmetric: Jaccard distance

Object j

Object i

	X=1	X=0
X=1	a	b
X=0	c	d

a+b+c+d = Number of variables

Uninformative

Simple matching coefficient

$$d(i,j) = \frac{b+c}{a+b+c}$$

Proportion of variables,
in which people disagree
ignoring (0,0)

Nominal: Simple matching coefficient

Simple matching coefficient

- mm: Number of variables in which object i and j mismatch
- p: Number of variables

$$d(i,j) = \frac{mm}{p}$$

Proportion of variables, in which people disagree

Ordinal: Normalized ranks

- Rank outcome of variable f=1,2,…,M: r_{if}
- Normalize: $z_{if} = \frac{r_{if} - 1}{M_f - 1}$

- Treat z_{if} as interval-scaled

Mixed: Gower Distance

- Idea: Use distance measure between 0 and 1 for each variable: $d_{ij}^{(f)}$
- Aggregate: $d(i,j) = \frac{1}{p}\sum_{i=1}^{p} d_{ij}^{(f)}$

- Binary (a/s), nominal: Use methods discussed before
- Interval-scaled: $d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{R_f}$
  x_{if}: Value for object i in variable f
  R_f: Range of variable f for all objects
- Ordinal: Use normalized ranks; then like interval-scaled based on range

Concepts to know

- Effect of scaling / no scaling
- Distance measures for
  - interval scaled
  - binary (s/a)
  - nominal
  - categorical
  - mixed data

R functions to know

- dist, daisy