UNIVERSITAT DE BARCELONA

Distance in statistics

MULTIVARIATE ANALYSIS
MESIO (15-16)

PROF. SERGI CIVIT
PROF. MIQUEL SALICRÚ

Observations

Real-value attributes/variables

```
# e.g., salary, height
```

Binary attributes

```
# e.g., gender (M/F), has_cancer(T/F)
```

Nominal (categorical) attributes

```
# e.g., religion (Christian, Muslim, Buddhist, Hindu,...)
```

Ordinal/Ranked attributes

```
# e.g., military rank (soldier, sergeant, captain, etc.)
```

Variables of mixed types

```
# multiple attributes with various types
```

Distance functions

The distance **d(x,y)** between two objects **x** and **y** is a metric if

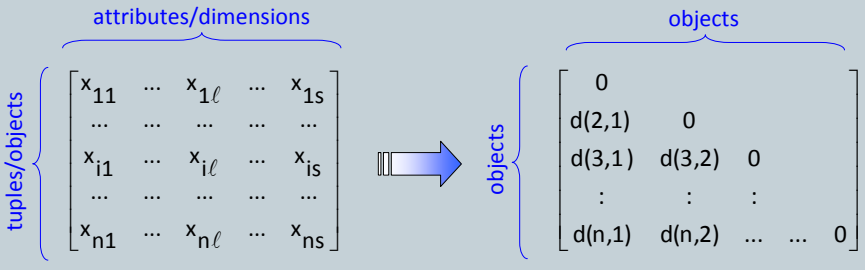
- **d(i,j)≥0** (non-negativity)
- **d(i,i)=0** (isolation)
- **d(i,j)= d(j, i)** (symmetry)
- **d(i,j) ≤ d(i,h) + d(h,j)** (triangular inequality)

The definitions of distance functions are usually different for real, boolean, categorical, and ordinal variables.

Weights may be associated with different variables based on applications and data semantics.

Data Structures

Data matrix **Distance matrix**



Distance measures 1/7

L_p norms or *Minkowski distance* (geometrical distance)

$$L_p(x_1,x_2)=\left(\sum_{j=1}^{s}|x_{1j}-x_{2j}|^p\right)^{1/p}$$

where:

```
# p is a positive value
# x_i=(x_i1,….,x_is), i=1,2
```

p=1: L_1 (Manhattan or city block) distance

$$d_{Man}(x_1,x_2)=L_1(x_1,x_2)=\sum_{j=1}^{s}|x_{1j}-x_{2j}|$$

#p=2: L_2 (Euclidean) distance

$$d_{EC}(x_1,x_2)=L_2(x_1,x_2)=\sqrt{\sum_{j=1}^{s}(x_{1j}-x_{2j})^2}$$

Distance measures 2/7

L_p weighted distance

$$L_p(x_1,x_2)=\left(\sum_{j=1}^{s}w_j|x_{1j}-x_{2j}|^p\right)^{1/p}$$

p=1

$$d_{ManW}(x_1,x_2)=\sum_{j=1}^{s}w_j|x_{1j}-x_{2j}|$$

#p=2

$$d_{ECW}(x_1,x_2)=\sqrt{\sum_{j=1}^{s}w_j(x_{1j}-x_{2j})^2}$$

Distance measures 3/7

L_p related distance

Absolute value distance

$$d_{ABS}(x_1, x_2) = \sqrt{\sum_{j=1}^{s} |x_{1j} - x_{2j}|}$$

Squared ED

$$d_{SQEC}(x_1, x_2) = \sum_{j=1}^{s} (x_{1j} - x_{2j})^2$$

Mahalanobis distance (transform $S^{-1/2}$)

$$d_{Mah}(x_1, x_2) = d_{12}' S^{-1} d_{12}$$

where $d_{12} = (x_{11} - x_{21}, \ldots, x_{1s} - x_{2s})'$ (S:var-covar matrix)

Distance measures 4/7

Helliger distance

$$d_{Hell}(x_1, x_2) = \sqrt{\sum_{j=1}^{s} \left[\sqrt{\frac{x_{1j}}{x_{1+}}} - \sqrt{\frac{x_{2j}}{x_{2+}}} \right]^2}$$

where $x_{i+} = x_{i1} + \ldots + x_{is}$, $i = 1, 2$

Chord distance

$$d_{Chord}(x_1, x_2) = \sqrt{\sum_{j=1}^{s} \left[\frac{x_{1j}}{\|x_1\|} - \frac{x_{2j}}{\|x_2\|} \right]^2}$$

where $\|x_i\| = \sqrt{\sum_{j=1}^{s} x_{ij}^2}$

Distance measures 5/7

Species profiles distance

$$d_{SpProf}(x_1, x_2) = \sqrt{\sum_{j=1}^{s} \left[\frac{x_{1j}}{x_{1+}} - \frac{x_{2j}}{x_{2+}} \right]^2}$$

Bray-Curtis distance

$$d_{BC}(x_1, x_2) = \sum_{j=1}^{s} \frac{|x_{1j} - x_{2j}|}{x_{1+} + x_{2+}}$$

Distance measures 6/7

χ^2 metric distance

$$d_{m\chi^2}(x_1, x_2) = \sqrt{\sum_{j=1}^{s} \left[\frac{x_{1j}}{x_{1+}x_{+j}} - \frac{x_{2j}}{x_{2+}x_{+j}} \right]^2}$$

χ^2 distance

$$d_{m\chi^2}(x_1, x_2) = \sqrt{\sum_{j=1}^{s} \left[\frac{x_{1j}}{(x_{1+}x_{+j})/x_{++}} - \frac{x_{2j}}{(x_{2+}x_{+j})/x_{++}} \right]^2}$$

where $x_{+j}=x_{ij}+x_{2j}$, $j=1,\ldots,s$ and $x_{++}=\sum_{ij}(x_{ij})$

Distance measures 7/7

Binary ED (binary variables)

$$d_{BinEC}(x_1, x_2) = \sqrt{\sum_{j=1}^{s} [I(x_{1j} > 0) - I(x_{2j} > 0)]^2}$$

Squared binary ED (binary variables)

$$d_{BinEC}(x_1, x_2) = \sum_{j=1}^{s} [I(x_{1j} > 0) - I(x_{2j} > 0)]^2$$

where $I(x_{ij}>0)=1$ if $x_{ij}>0$ and $I(x_{ij}>0)=0$ if $x_{ij}<0$

Pearson correlation component (for variables)

$$d_{CP}(x_1, x_2) = 1 - r^2_{Perarson}(x_1, x_2)$$

Comment

Group 1 (Pythagorean Euclidean Distance for quantitative data and related): Euclidean Distance; Squared Euclidean Distance; Manhattan or city block metric; Absolute Value Distance; Mahalanobis Distance (Like Euclidean distance but it takes into account variable correlations and "extracts" them); Lp distance; Binary Euclidean Distance and Binary Squared Euclidean Distance (Pythagorean Euclidean Distance for binary data)

Group 2 (Angle between Profiles/Euclidean distance computed after transforming data): Chord Distance; Hellinger Distance; Species Profiles Distance; Bray-Curtis Distance (same sampling effort)

Group 3 (Statistical distances): Chi-square distance and Chi-square metric (Contingency tables, Correspondence Analysis,....).

Distance?

	C1	C2	C3	C4
P1	0	3	0	2
P2	2	0	1	0
P3	0	20	0	12

$$d_{EC} = \begin{pmatrix} 0 & 4.24 & 19.72 \\ 4.24 & 0 & 23.43 \\ 19.72 & 23.43 & 0 \end{pmatrix}$$

$$d_{Man} = \begin{pmatrix} 0 & 8 & 27 \\ 8 & 0 & 35 \\ 27 & 35 & 0 \end{pmatrix}$$

$$d_{BC} = \begin{pmatrix} 0 & 1 & 0.73 \\ 1 & 0 & 1 \\ 0.73 & 1 & 0 \end{pmatrix}$$

$$d_{Hell} = \begin{pmatrix} 0 & 1.41 & 0.03 \\ 1.41 & 0 & 1.41 \\ 0.03 & 1.41 & 0 \end{pmatrix}$$

Similarity measures 1/3

Contingency table for binary data

		Object *j*		
		1	0	sum
Object *i*	1	a	b	a+b
	0	c	d	c+d
	sum	a+c	b+d	

where

a = number of double presence (occurrence)

b = number of presence-absence

c = number of absence-presence

d = number of double absence

Similarity measures 2/3

Simple Matching Coefficient

$$s_1(x_1,x_2)=\frac{a+d}{a+b+c+d}$$

Jacard index

$$s_2(x_1,x_2)=\frac{a}{a+b+c}$$

Sorensen index

$$s_3(x_1,x_2)=\frac{2a}{2a+b+c}$$

distance: $d(x_1,x_2)=[1-s(x_1,x_2)]^{\alpha}$

Similarity measures 3/3

Motyka index

$$s_4(x_1,x_2)=\frac{\sum_{j=1}^{s}\min(x_{1j},x_{2j})}{\sum_{j=1}^{s}x_{1j}+\sum_{j=1}^{s}x_{2j}}=\frac{\sum_{j=1}^{s}\min(x_{1j},x_{2j})}{x_{1+}+x_{2+}}$$

Contingency table: $s_4(x_1,x_2)=\frac{1}{2}s_3(x_1,x_2)$

Kulczynski index

$$s_5(x_1,x_2)=\frac{1}{2}\left(\frac{\sum_{j=1}^{s}\min(x_{1j},x_{2j})}{\sum_{j=1}^{s}x_{1j}}+\frac{\sum_{j=1}^{s}\min(x_{1j},x_{2j})}{\sum_{j=1}^{s}x_{2j}}\right)$$

Contingency table: $s_5(x_1,x_2)=\frac{1}{2}\left[\frac{a}{a+b}+\frac{a}{a+c}\right]$

Mixed variables: Gower's distance

$$s_{ij}(x_1, x_2) = \frac{\sum_{h=1}^{p_1} (1 - \frac{|x_{ih} - x_{jh}|}{R_h}) + a + \alpha}{p_1 + (p_2 - d) + p_3}$$

being:

- $\mathbf{p_1}$, $\mathbf{p_2}$ and $\mathbf{p_3}$**:** ordinal and continuous numeric variables, binary variables and nominal variables
- $\mathbf{R_h}$**:** rank $\mathbf{X_h}$
- **a:** number of double presence 1/1 (binary) – Jacard index
- **α:** number of coincidences (nominal/categorical)
- **d:** number of double absence 0/0 (binary)

and many special distances.....

Measuring distance between two sequences: linguistics, genomics,...

Operations: insertion, deletion, substitution

Example: HARINA → AHORA

Insert +A	AHARINA
Replace A/O	AHORINA
Delete –I	AHORNA
Delete –N	AHORA

Levenshtein distance = 4

Para comenzar,....

¿Qué distancia (similitud)

caracteriza las analogías y diferencias

en la situación experimental que estudiamos?

Hay muchas más medidas?

economía, biología, electrónica,.....