2. cvičení

20.10.2015

Obsah

- Úvod
- Principy asociace ve vícerozměrném prostoru
- Euklidovská vzdálenost, Manhattan distance
  - Odvodit asociační matici 5x5
  - Pythagorova věta (excel, R)
  - Soubory z předchozího příkladu, ale oříznuté jen na dvě proměnné
  - Pomocí makra v excelu horní trojúhelníkovou matici zlinearizovat a vykreslit do histogramu
- Soubor s množstvím bodů (opět např. města)
  - Odvodit asociační matici nxn vzdušnou čarou
  - Odvodit asociační matici nxn po silnici
  - Ukázat opět xy graf a komentář, že jde o značně obtížnější problém
- Horní trohúhelníkové matice zlinerizovat a dát do xy grafu proti sobe

Úvod do vícerozměrných metod I.

- **Vícerozměrné metody**: Název vícerozměrné vychází z typu vstupních dat, tato data jsou tvořena jednotlivými objekty (i.e. klienti) a každý z nich je charakterizován svými parametry (věk, příjem atd.) a každý z těchto parametrů můžeme považovat za jeden rozměr objektu.
- **Maticová algebra**: Základem práce s daty a výpočtů vícerozměrných metod je maticová algebra, matice tvoří jak vstupní, tak výstupní data a probíhají na nich výpočty.
- **NxP matice**: N objektů s p parametry pak vytváří tzv. NxP matici, která je prvním typem vstupu dat do vícerozměrných analýz.
- **Asociační matice**: Na základě těchto matic jsou počítány matice asociační na nichž pak probíhají další výpočty, jde o čtvercové matice obsahující informace o podobnosti nebo rozdílnosti (tzv. **metriky**) buď objektů (Q mode analýza) nebo parametrů (R mode analýza).Měřítko podobnosti se liší podle použité metody a typu dat, některé metody umožňují použití uživatelských metrik.

Úvod do vícerozměrných metod II.

SHLUKOVÁ ANALÝZA

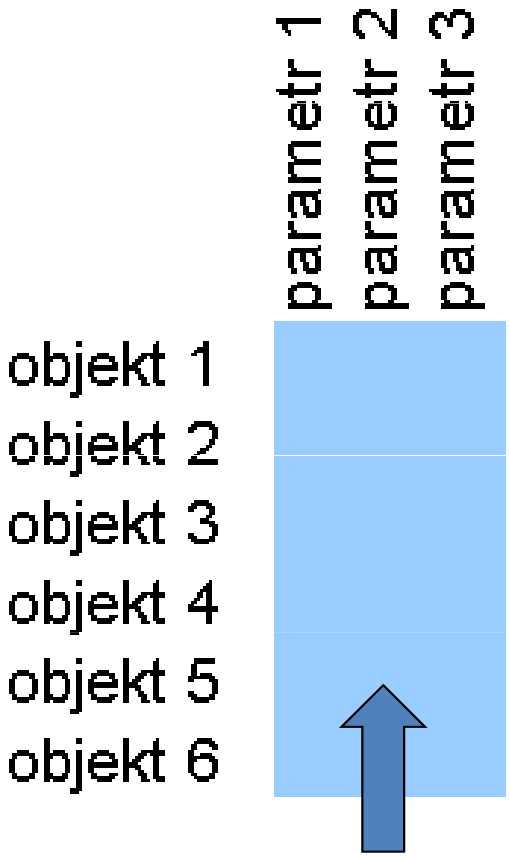
- vytváření shluků objektů na základě jejich podobnosti
- identifikace typů objektů

ORDINAČNÍ METODY

- zjednodušení vícerozměrného problému do menšího počtu rozměrů
- principem je tvorba nových rozměrů, které lépe vyčerpávají variabilitu dat

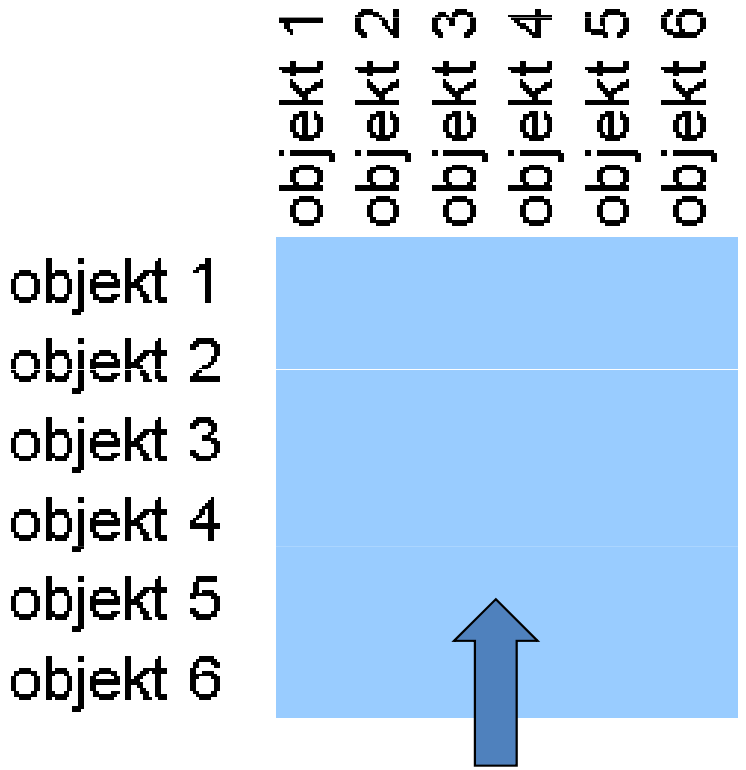
Vstupní matice vícerozměrných analýz

NxP MATICE



Hodnoty parametrů pro jednotlivé objekty

ASOCIAČNÍ MATICE



Korelace, kovariance, vzdálenost, podobnost

Asociace ve vícerozměrném prostoru

Data

STATISTICA - [Data: Adstudy* (24v by 50c)]

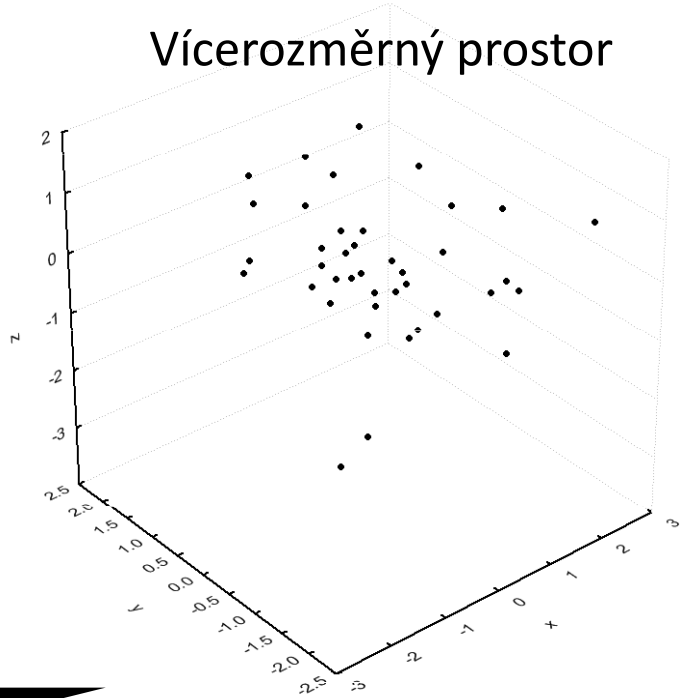
File Edit View Insert Format Statistics Data Mining Graphs Tools Data Window

Add to Workbook Add to Report

Arial 10 B I U

	Advertising Effectiveness Study.				
	1 ADVERT	2 MEASURE01	3 MEASURE02	4 MEASURE03	5 MEASU
R. Rafuse	id_1	9	1	6	
T. Leiker	id_2	6	7	1	
E. Bizot	id_3	9	8	2	
K. French	id_4	7	9	0	
E. Van Landuyt	id_5	7	1	6	
K. Harrell	id_6	6	0	0	
W. Noren	id_7	7	4	3	
W. Willden	id_8	9	9	2	
S. Kohut	id_9	7	8	2	
B. Madden	id_10	6	6	2	

Vícerozměrný prostor



Asociační matice

Case No.	Euclidean distances (multidinensional_normal_distribution)											
	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_10	C_11	C_12
C_1	0.00	2.58	3.73	0.95	1.46	1.85	3.78	1.85	1.59	1.80	0.46	2.82
C_2	2.58	0.00	2.17	2.58	1.90	0.82	2.32	1.99	1.14	1.57	2.26	1.73
C_3	3.73	2.17	0.00	3.15	3.54	2.19	1.66	2.49	2.34	2.58	3.53	2.88
C_4	0.95	2.58	3.15	0.00	1.85	1.77	3.45	1.31	1.53	1.56	1.01	3.00
C_5	1.46	1.90	3.54	1.85	0.00	1.50	3.88	1.58	1.58	1.12	1.02	2.90
C_6	1.85	0.82	2.19	1.77	1.50	0.00	2.40	1.38	0.41	1.07	1.56	1.81
C_7	3.78	2.32	1.66	3.45	3.88	2.40	0.00	3.31	2.36	3.28	3.70	1.86
C_8	1.85	1.99	2.49	1.31	1.58	1.38	3.31	0.00	1.48	0.59	1.53	3.14
C_9	1.59	1.14	2.34	1.53	1.58	0.41	2.36	1.48	0.00	1.27	1.39	1.68

Příklad výpočtu asociační matice

STATISTICA - [Data: Irisdat* (5v by 150c)]

File Edit View Insert Format Statistics Data Mining Graphs Tools Data Win

Add to Workbook Add to Rep

Arial 10 B I U

Fisher (1936) iris data: length & width of sepals and petals, 3 types of I

	1 SEPALLEN	2 SEPALWID	3 PETALLEN	4 PETALWID	5 IRISTYPE
1	5.0	3.3	1.4	0.2	SETOSA
2	6.4	2.8	5.6	2.2	VIRGINIC
3	6.5	2.8	4.6	1.5	VERSICO
4	6.7	3.1	5.6		
5	6.3	2.8	5.1		
6	4.6	3.4	1.4		
7	6.9	3.1	5.1	2.3	VIRGINIC
8	6.2	2.2	4.5	1.5	VERSICO
9	5.9	3.2	4.8	1.8	VERSICO
10	4.6	3.6	1.0	0.2	SETOSA
11	6.1	3.0	4.6	1.4	VERSICO
12	6.0	2.7	5.1	1.6	VERSICO
13	6.5	3.0	5.2	2.0	VIRGINIC
14	5.6	2.5	3.9	1.1	VERSICO
15	6.5	3.0	5.5	1.8	VIRGINIC
16	5.8	2.7	5.1	1.9	VIRGINIC
17	6.8	3.2	5.9	2.3	VIRGINIC
18	5.1	3.3	1.7	0.5	SETOSA
19	5.7	2.8	4.5	1.3	VERSICO
20	6.2	3.4	5.4	2.3	VIRGINIC
21	7.7	3.8	6.7	2.2	VIRGINIC
22	6.3	3.3	4.7	1.6	VERSICO
23	6.7	3.3	5.7	2.5	VIRGINIC
24	7.6	3.0	6.6	2.1	VIRGINIC
25	4.9	2.5	4.5	1.7	VIRGINIC
26	5.5	3.5	1.3	0.2	SETOSA
27	6.7	3.0	5.2	2.3	VIRGINIC
28	7.0	3.2	4.7	1.4	VERSICO
29	6.4	3.2	4.5	1.5	VERSICO
30	6.1	2.8	4.0	1.3	VERSICO
31	4.8	3.1	1.6	0.2	SETOSA
32	5.9	3.0	5.1	1.8	VIRGINIC
33	5.5	2.4	3.8	1.1	VERSICO
34	6.3	2.5	5.0	1.9	VIRGINIC

Euclidean distances (Irisdat)

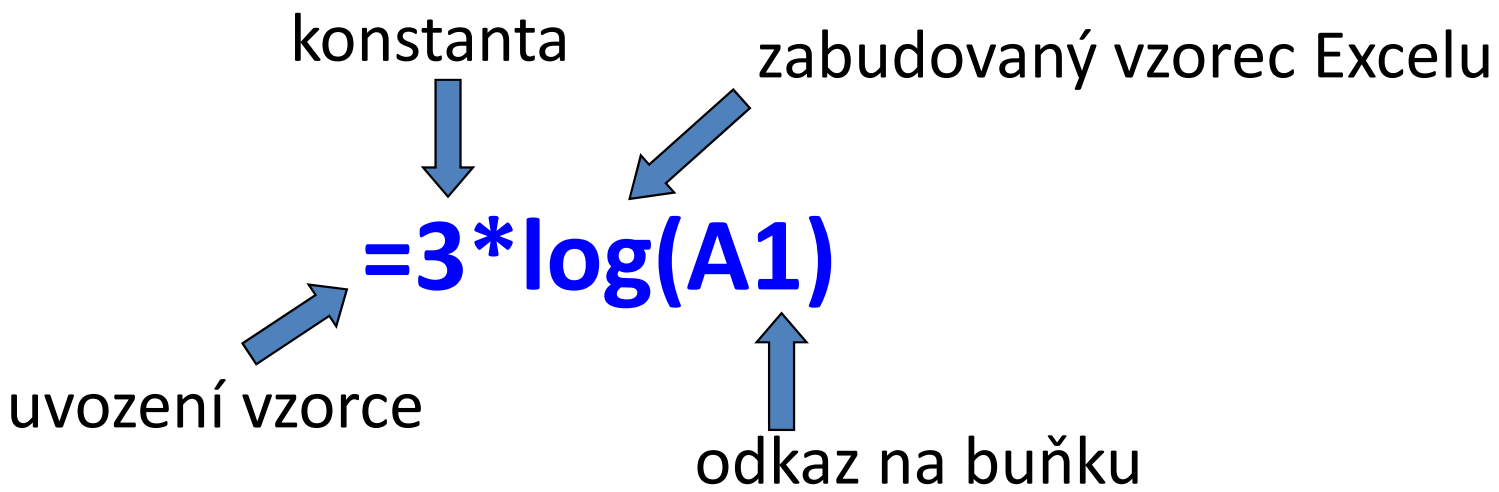
Case No.	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_10	C_11	C_12	C_13	C_14	C_15	C_16	C_17
C_1	0.00	4.88	3.80	5.04	4.16	0.42	4.66	3.73	3.87	0.64	3.60	4.12	4.47	2.84	4.66	4.19	5.28
C_2	4.88	0.00	1.22	0.47	0.87	4.98	0.77	1.45	1.10	5.39	1.33	0.88	0.50	2.20	0.47	0.84	0.65
C_3	3.80	1.22	0.00	1.39	0.54	3.96	1.07	0.68	0.81	4.35	0.46	0.72	0.81	1.24	0.97	0.95	1.61
C_4	5.04	0.47	1.39	0.00	1.14	5.15	0.55	1.75	1.28	5.54	1.54	1.24	0.61	2.48	0.65	1.21	0.35
C_5	4.16	0.87	0.54	1.14	0.00	4.29	1.04	0.85	0.71	4.69	0.58	0.33	0.58	1.48	0.57	0.65	1.30
C_6	0.42	4.98	3.96	5.15	4.29	0.00	4.80	3.88	3.94	0.46	3.72	4.22	4.59	2.95	4.78	4.26	5.40
C_7	4.66	0.77	1.07	0.55	1.04	4.80	0.00	1.52	1.16	5.17	1.31	1.21	0.52	2.22	0.76	1.24	0.81
C_8	3.73	1.45	0.68	1.75	0.85	3.88	1.52	0.00	1.13	4.30	0.82	0.81	1.21	0.98	1.35	0.96	1.99
C_9	3.87	1.10	0.81	1.28	0.71	3.94	1.16	1.13	0.00	4.34	0.53	0.62	0.77	1.37	0.94	0.60	1.51
C_10	0.64	5.39	4.35	5.54	4.69	0.46	5.17	4.30	4.34	0.00	4.12	4.64	4.98	3.38	5.17	4.69	5.78
C_11	3.60	1.33	0.46	1.54	0.58	3.72	1.31	0.82	0.53	4.12	0.00	0.62	0.94	1.04	1.06	0.82	1.74
C_12	4.12	0.88	0.72	1.24	0.33	4.22	1.21	0.81	0.62	4.64	0.62	0.00	0.71	1.37	0.73	0.36	1.42
C_13	4.47	0.50	0.81	0.61	0.58	4.59	0.52	1.21	0.77	4.98	0.94	0.71	0.00	1.89	0.36	0.77	0.84
C_14	2.84	2.20	1.24	2.48	1.48	2.95	2.22	0.98	1.37	3.38	1.04	1.37	1.89	0.00	2.03	1.47	2.71
C_15	4.66	0.47	0.97	0.65	0.57	4.78	0.76	1.35	0.94	5.17	1.06	0.73	0.36	2.03	0.00	0.87	0.73
C_16	4.19	0.84	0.95	1.21	0.65	4.26	1.24	0.96	0.60	4.69	0.82	0.36	0.77	1.47	0.87	0.00	1.43
C_17	5.28	0.65	1.61	0.35	1.30	5.40	0.81	1.99	1.51	5.78	1.74	1.42	0.84	2.71	0.73	1.43	0.00
C_18	0.44	4.48	3.41	4.63	3.77	0.62	4.25	3.36	3.46	0.96	3.21	3.73	4.07	2.47	4.26	3.79	4.88
C_19	3.40	1.58	0.83	1.87	0.87	3.49	1.70	0.81	0.73	3.91	0.47	0.74	1.29	0.71	1.39	0.86	2.08
C_20	4.68	0.67	1.32	0.62	1.05	4.75	0.82	1.70	0.86	5.14	1.27	1.05	0.62	2.20	0.71	0.95	0.81

Asociační matice euklidovských vzdáleností mezi rostlinami

Vzdálenosti měst ČR

Vzorce v Excelu

- vpisují se do buněk sešitu
- vzorce jsou vždy uvozeny **=** (lze též **+ -**)
- aritmetické operátory + zabudované funkce Excelu
- pro „sčítání“ nečíselných položek se používá **&**
- výpočet je založen buď na číselných konstantách nebo odkazech na buňky



Vzorce v Excelu – odkazy na buňku – styl A1

Relativní odkazy

- **A1** = buňka 1. řádku sloupci A
- **A1:B6** = blok buněk – levý horní roh je v 1. řádku, sloupec A, pravý dolní na řádku 6, sloupec B
- relativní odkaz se při automatickém vyplnění buněk vzorcem posune

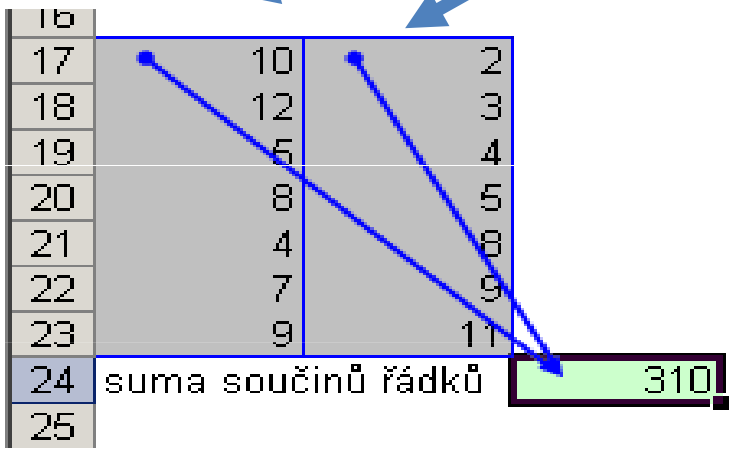
Absolutní odkaz – odkaz na buňku je pevně dán, při kopírování nebo automatickém vyplnění se nemění, lze uzamknout jak řádky, tak sloupce samostatně

uzamčení řádku

uzamčení sloupce → **A1**

Maticové vzorce v Excelu

- výpočty z matic dat
- zadávání je ukončeno stiskem CTRL+SHIFT+ENTER



17	10	2
18	12	3
19	5	4
20	8	5
21	4	8
22	7	9
23	9	11
24	suma součinů řádků	310
25		

Vzorec je založen na těchto dvou maticích dat

{=SUMA(A17:A23*B17:B23)}

Násobení řádků matic

Celkové sečtení

Nezbytné pro operace s **maticemi.**

Měření vzdálenosti objektů

Euklidovská vzdálenost

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

Vážená euklidovská vzdálenost

$$d_{ij} = \sqrt{\sum_{k=1}^{p} w_k^2 (x_{ik} - x_{jk})^2}$$

i,j – označení objektů
d_{ij} – vzdálenost objektů i a j
p – počet parametrů
k – k-tý parametr
w_k – váha parametru k

Minkowski (power distance)

$$d_{ij} = \sqrt[\lambda]{\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^{\lambda}}$$

- - celé číslo
- =1 Manhattan (city block)
- = 2 Euklidovská vzdálenost

Chebychev

$$d_{ij} = \max \left| x_{ik} - x_{jk} \right|$$

IBA MU

Měření podobnosti objektů

Binární koeficienty podobnosti

Objekt 2 \ Objekt 1	1	0
1	a	b
0	c	d

a, b, c, d = počet případů, kdy souhlasí binární charakteristika objektu 1 a 2

a+b+c+d=p

Symetrické binární koficienty - není rozdíl mezi případem 1-1 a 0-0

Simple matching coefficient

$$S(x_1 x_2) = \frac{a+d}{p}$$

Hamman, Yule coefficient, Pearson's ↗ (phi) a další koeficienty

Asymetrické binární koeficienty – odstranění double zero

Jaccard`s coefficient

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

Sorensen`s coefficient

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

Řada dalších koeficientů dávajících různou váhu jednotlivým kombinacím parametrů

Kvantitativní koeficienty

Obdoby binárních koeficientů pro více parametrů než 0/1

Simple matching coefficient pro více parametrů

$$S(x_1 x_2) = \frac{souhlas}{p}$$

p=počet parametrů

Gowerův koeficient

Zahrnutí podobnosti podle různých typů parametrů – binární, kvalitativní a semikvantitativní i kvantitativní (odlišný výpočet pro jednotlivé typy). Celkový součet podobností je podělen počtem parametrů. Může zahrnovat podmínku nepočítat s chybějícími parametry – Kronecker`s delta.

Více informací a další měření vzdáleností a podobností najdete v knize **LEGENDRE, P. & LEGENDRE, L. (1998). *Numerical ecology*. Elseviere Science BV, Amsterodam.**