Vícerozměrné metody - cvičení

Mgr. Jan Zdražil

Shluková analýza – typy metod – opakování

Shluková analýza

Hierarchické
shluky jsou definovány postupným skládáním objektů

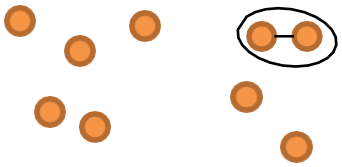
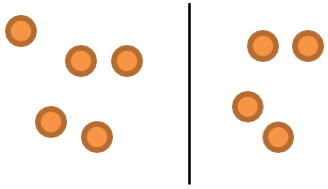
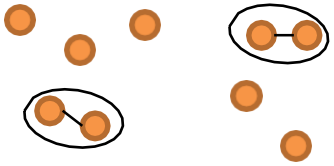
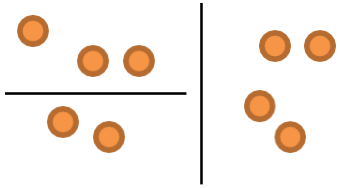
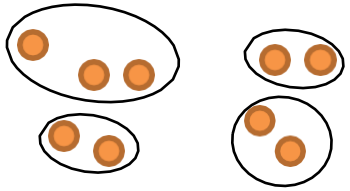
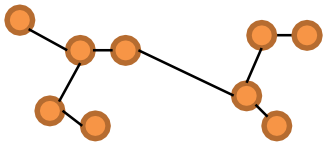
Nehierarchické
shluky jsou definovány v jednom kroku

Aglomerativní
Po spojení první dvojice objektů dochází k postupnému napojování dalších objektů.

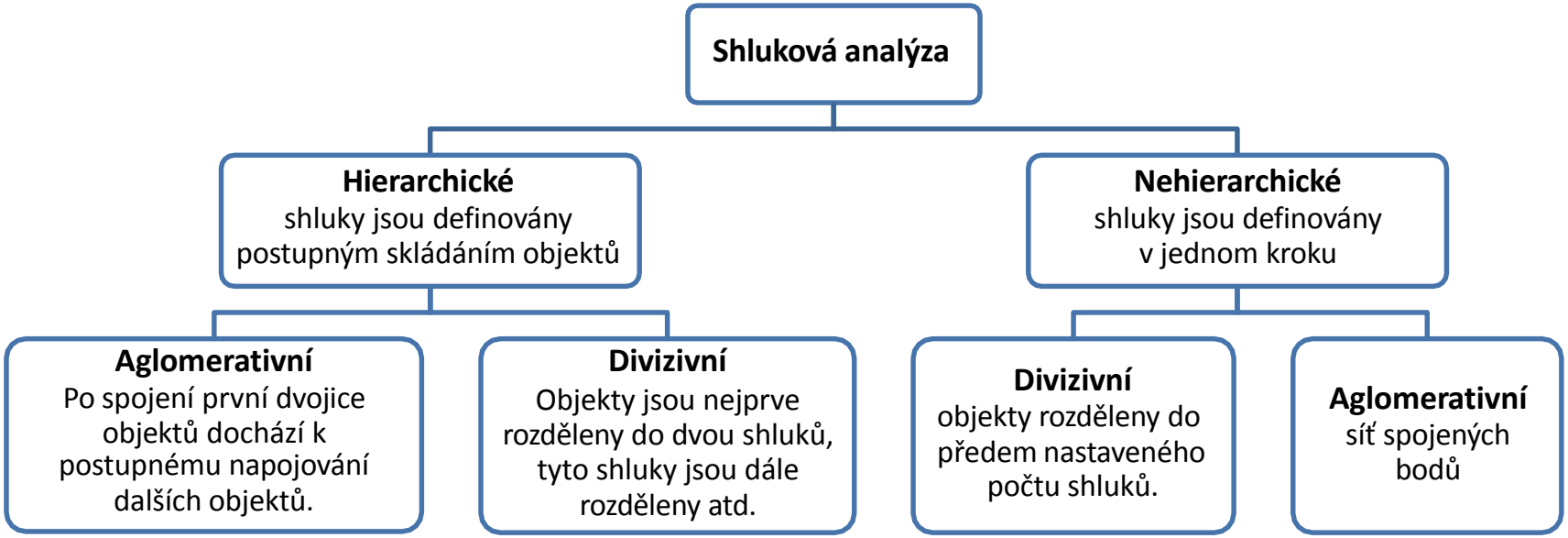
Divizivní
Objekty jsou nejprve rozděleny do dvou shluků, tyto shluky jsou dále rozděleny atd.

Divizivní
objekty rozděleny do předem nastaveného počtu shluků.

Aglomerativní
síť spojených bodů

1. Krok			Kolik shluků chceme definovat? Například 4	Minimum spanning tree, Prime network
2. Krok				
X. Krok	Atd.	Atd.	Výpočet ukončen	Výpočet ukončen

Shluková analýza – typy metod – opakování



Metoda nejbližšího souseda
Metoda nejvzdál. souseda
Metoda průměrné vazby
Centroidová metoda
Wardova metoda

Metody monotetické
Metody polytetické – TWINSPAN

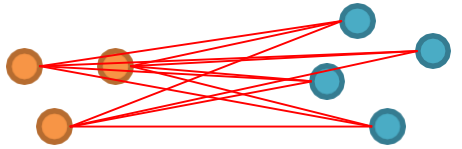
Metoda *k*-průměrů
Metoda *x*-průměrů
Metoda *k*-medoidů

Minimum spanning tree

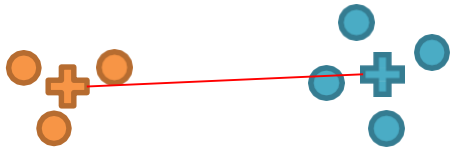
Shlukovací algoritmy hierarchického aglomerativního shlukování

- **Metoda nejbližšího souseda** (jednospojná metoda, metoda jediné vazby, metoda krátké ruky, *nearest neighbour, simple linkage*) – spojení dle nejmenší vzdálenosti mezi objekty shluků

- **Metoda průměrné vazby** (středospojná metoda, *average linkage*) – spojení dle průměrné vzdálenosti mezi objekty shluků
  - Nevážená (*unweighted, UPGMA*) – výpočet spojovací vzdálenosti je ovlivněn velikostí spojovaných shluků
  - Vážená (*weighted, WPGMA*) – odstranění vlivu velikosti shluků, shluky bez ohledu na velikost přispívají k výpočtu spojovací vzdálenosti stejnou vahou



- **Centroidová metoda** (centroidní metoda, metoda středospojné vzdálenosti, Gowerova metoda, *centroid method*) – spojení dle vzdálenosti centroidů shluků
  - Nevážená (*unweighted, UPGMC*) – výpočet spojovací vzdálenosti je ovlivněn velikostí spojovaných shluků
  - Vážená (*weighted, WPGMC*, mediánová metoda, *median method*) – odstranění vlivu velikosti shluků



- **Metoda nejvzdálenějšího souseda** (všespojná metoda, metoda dlouhé ruky, *furthest neigbour, complete linkage*) – spojení dle největší vzdálenosti mezi objekty shluků

Příklad 3

Ve studii byl u 6 osob zjišťován systolický tlak a hladina celkového cholesterolu v krvi:

Pac	Systolický tlak (mmHg)	Celkový cholesterol (mmol/l)
A	165	4,5
B	125	4,7
C	160	7,5
D	170	7,0
E	130	4,0
F	165	6,5

A) Asociační matice počítaná na původních datech

	A	B	C	D	E	F
A	0	40,00	5,83	5,59	35,00	2,00
B	40,00	0	35,11	45,06	5,05	40,04
C	5,83	35,11	0	10,01	30,20	5,10
D	5,59	45,06	10,01	0	40,11	5,02
E	35,00	5,05	30,20	40,11	0	35,09
F	2,00	40,04	5,10	5,02	35,09	0

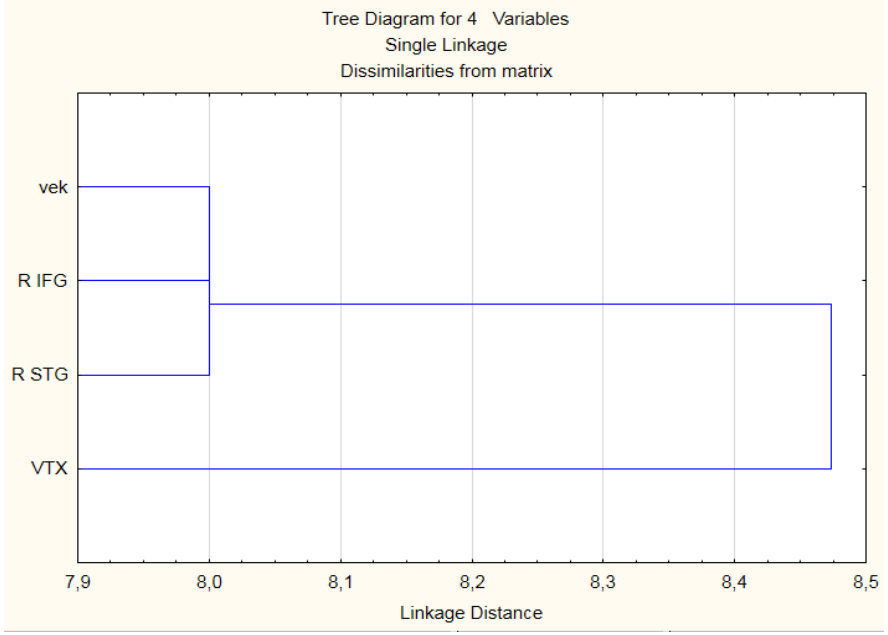
B) Asociační matice počítaná na standardizovaných datech

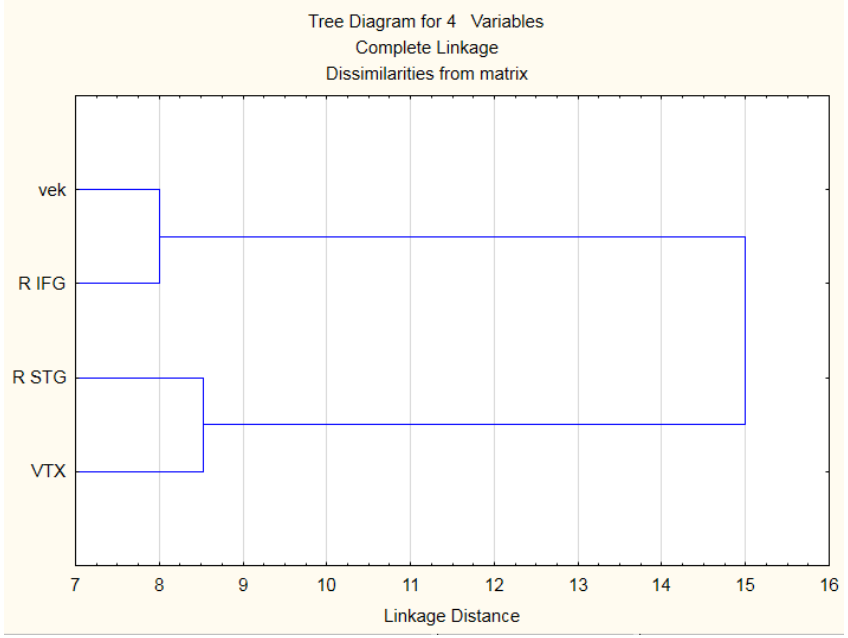
	A	B	C	D	E	F
A	0	2,04	2,05	1,71	1,81	1,35
B	2,04	0	2,60	2,77	0,54	2,37
C	2,05	2,60	0	0,61	2,82	0,72
D	1,71	2,77	0,61	0	2,87	0,42
E	1,81	0,54	2,82	2,87	0	2,46
F	1,35	2,37	0,72	0,42	2,46	0

1. Jedná se o podobnost nebo vzdálenost?
2. Který koeficient použit? Jaccardův, Sokalům-Michenerův, Euklidova vzdálenost nebo Manhattenská?
3. Použili bychom asociační matici?
4. Můžeme použít asociační matici počítanou na standardizovaných datech?

Doplnění - co se stane, když jsou v datech stejné vzdálenosti

	1 vek	2 R IFG	3 R STG	4 VTX
vek	0,00000	8,00000	8,00000	15,00000
R IFG	8,00000	0,00000	9,00000	8,47351
R STG	8,00000	9,00000	0,00000	8,52360
VTX	15,00000	8,47351	8,52360	0,00000
Means	71,03563	0,13179	0,56080	0,24312
Std.Dev.	7,26298	1,00000	1,00000	0,99999
No.Cases	34,00000			
Matrix	3,00000			



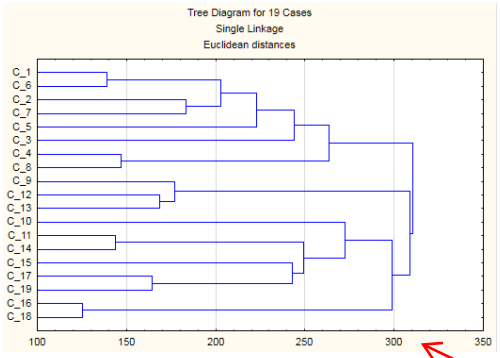


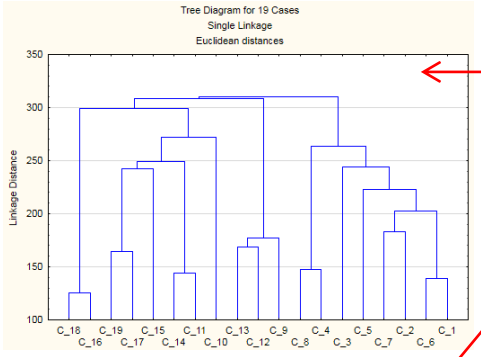
Výpočet shlukové analýzy v softwarech

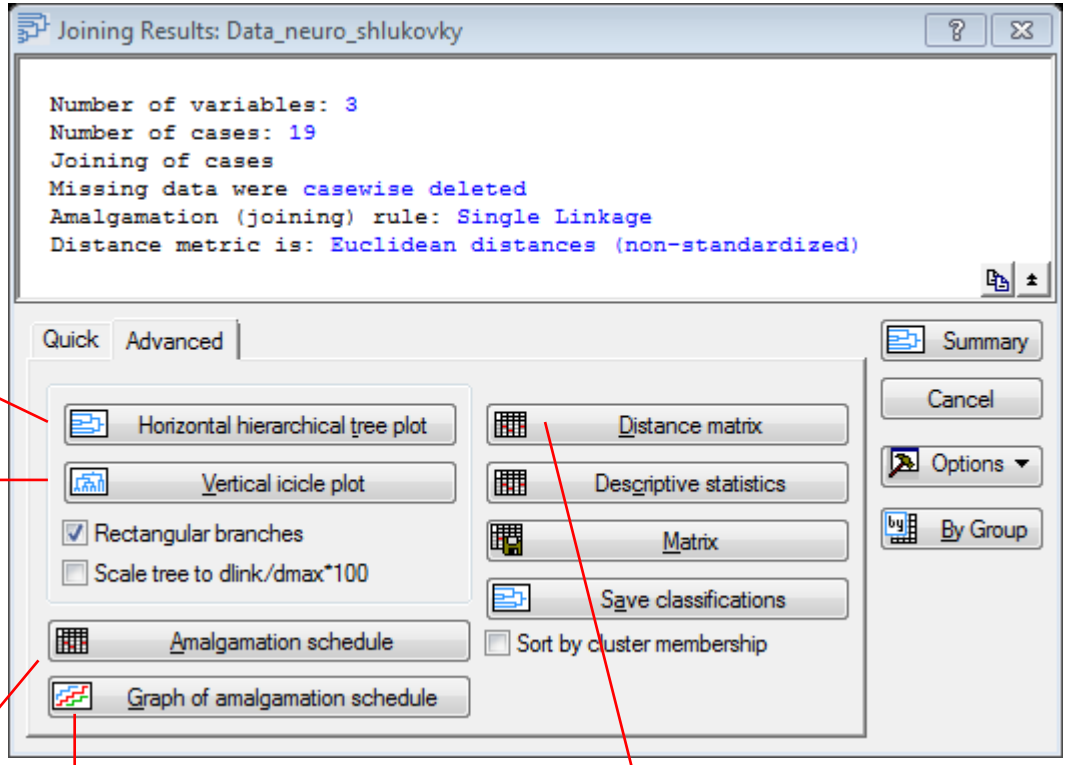
STATISTICA – hierarchické aglomerativní shlukování

- Statistics – Mult/Exploratory – Cluster – Joining (tree clustering) – OK – přepnout se na záložku Advanced
- Variables: výběr proměnných (např. objem hipokampu, amygdaly a pallida)
- Cluster: zvolit, zda chceme shlukovat proměnné (Variables (columns)) či subjekty (Cases (rows))
- Amalgamation (linkage) rule = volba shlukovacího algoritmu:
  - Single Linkage – metoda nejbližšího souseda
  - Complete Linkage – metoda nejvzdálenějšího souseda
  - Unweighted pair-group average – metoda průměrné vazby (nevážená)
  - Weighted pair-group average – metoda průměrné vazby (vážená)
  - Unweighted pair-group centroid – centroidová metoda (nevážená)
  - Weighted pair-group centroid (median) – centroidová metoda (vážená) = mediánová metoda
  - Ward's method – Wardova metoda
- Distance measure = volba metrik vzdáleností objektů (subjektů):
  - Squared Euclidean distances – čtverec Euklidovy vzdálenosti
  - Euclidean distances – Euklidova metrika
  - City-block (Manhattan) distances – Hammingova (manhattanská) metrika
  - Chebychev distance metric – Čebyševova metrika
  - Power: SUM(ABS(x-y)**p)**1/r – pokud r=p, jde o Minkovského metriku
  - Percent disagreement
  - 1-Pearson r – jedna mínus Pearsonův korelační koeficient

STATISTICA – hierarch. aglom. shluk. – pokračování

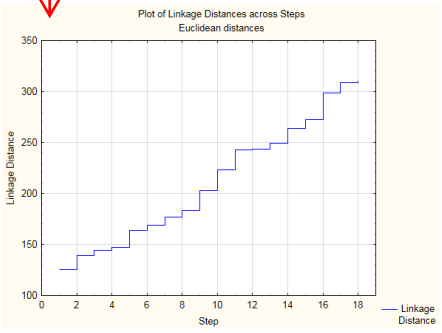






Amalgamation Schedule (Data_neuro_shlukovky)
Single Linkage
Euclidean distances

linkage distance	Obj. No. 1	Obj. No. 2	Obj. No. 3	Obj. No. 4	Obj. No. 5	O
125.1972	C_16	C_18				
139.2640	C_1	C_6				
143.9270	C_11	C_14				
147.0873	C_4	C_8				
164.1363	C_17	C_19				
168.6528	C_12	C_13				
176.9954	C_9	C_12	C_13			
183.2707	C_2	C_7				
202.7584	C_1	C_6	C_2	C_7		
223.0460	C_1	C_6	C_2	C_7	C_5	



asociační matice Euklidových vzdáleností

Euclidean distances (Data_neuro_shlukovky)

Case No.	C_1	C_2	C_3	C_4	C_5	C_6	C
C_1	0	291	299	490	271	139	
C_2	291	0	244	264	454	251	
C_3	299	244	0	500	527	311	
C_4	490	264	500	0	535	410	
C_5	271	454	527	535	0	223	
C_6	139	251	311	410	223	0	
C_7	307	183	262	328	399	203	

STATISTICA – nehierarchické shlukování

- Statistics – Mult/Exploratory – Cluster – K-means clustering – OK – přepnout se na záložku Advanced
- Variables: výběr proměnných (např. objem hipokampu, amygdaly a pallida)
- Cluster: zvolit, zda chceme shlukovat proměnné (Variables (columns)) či subjekty (Cases (rows))
- Number of clusters: zvolit počet shluků (např. 3)
- Number of iterations: volba počtu iterací (metoda *k*-průměrů je iterativní metoda)
- Initial cluster centers: volba počátečních středů shluků

- příslušnost jednotlivých subjektů do shluků nalezneme na záložce Advanced v „Members of each cluster & distances“

SPSS – hierarchické aglomerativní shlukování

- Analyze – Classify – Hierarchical Cluster...
- Cluster: zvolit, zda chceme shlukovat proměnné (Variables) či subjekty (Cases)
- Statistics...: zatrhnout Proximity matrix (= asociační matice vzdáleností či podobností)
- Plots...: zatrhnout Dendrogram (možnost volby Vertical či Horizontal)
- Method...:
  - Cluster Method = volba shlukovacího algoritmu:
    - Between-groups linkage – metoda průměrné vazby mezi skupinami
    - Within-groups linkage – metoda průměrné vazby uvnitř skupin
    - Nearest neighbor – metoda nejbližšího souseda
    - Furthest neighbor – metoda nejvzdálenějšího souseda
    - Centroid clustering – centroidová metoda (nevážená)
    - Median clustering – centroidová metoda (vážená) = mediánová metoda
    - Ward's method – Wardova metoda
  - Distance measure: volba metrik vzdáleností objektů (subjektů):
    - Euclidean distance – Euklidova metrika
    - Squared Euclidean distance – čtverec Euklidovy vzdálenosti
    - Cosine – kosinová metrika
    - Pearson correlation – Pearsonův korelační koeficient
    - Chebychev – Čebyševova metrika
    - Block – Hammingova (manhattanská) metrika
    - Minkowski – Minkovského metrika
    - Customized – výpočet pomocí SUM(ABS(x-y)**p)**1/r
  - Transform Values, Transform Measure – je možno transformovat původní data nebo vypočtené vzdálenosti
- Save: zatrhnout Single solution a zvolit Number of clusters: 3
- Pozor! Při vykreslování dendrogramu SPSS nezachovává původní vzdálenosti, ale přeškálovává je na škálu od 0 do 25!!!

SPSS – nehierarchické shlukování

- Analyze – Classify – K-Means Cluster...
- Variables: výběr proměnných (např. objem hipokampu, amygdaly a pallida)
- Number of clusters: zvolit počet shluků (např. 3)
- Method: přepnout na „Classify only“ v případě, že známe středy shluků, které můžeme načíst pomocí „Read initial“
- Iterate... – Maximum Iterations (volba počtu iterací – metoda *k*-průměrů je iterativní metoda)
- Options... – zatrhnout „Cluster information for each case“, abychom získali tabulku, do kterého shluku patří který subjekt

Software R – hierarchické aglomerativní shlukování

- funkce *dist* na výpočet vzdáleností objektů (či subjektů) :
  - „euclidean“ – Euklidovska metrika
  - „maximum“ – Čebyševova metrika
  - „manhattan“ – Hammingova (manhattanská) metrika
  - „canberra“ – Canberrská metrika
  - „minkowski“ – Minkovského metrika
- funkce *hclust* na výpočet shlukové analýzy:
  - „ward.D“ a „ward.D2“ – dva algoritmy pro Wardovu metodu
  - „single“ – metoda nejbližšího souseda (single linkage)
  - „complete“ – metoda nejvzdálenějšího souseda (complete linkage)
  - „average“ – metoda průměrné vazby (nevážená) (average linkage)
  - „mcquitty“ – metoda průměrné vazby (vážená)
  - „median“ – centroidová metoda (vážená) = mediánová metoda
  - „centroid“ – centroidová metoda (nevážená)
- podrobná ukázka v souboru Shlukovky_skript.R

Software R – nehierarchické shlukování

- funkce *kmeans*
- ukázka:

```
cl <- kmeans(data.vyber, 3)     # provedeni shlukove analyzy
table(cl$cluster,groupCodes) # zjisteni, kolik subjektu bylo spatne zarazenych
```

Matlab – výpočet vzdáleností

Funkce:

- pdist (vzdálenost mezi páry objektů matice X či páry proměnných matice X^T)
- pdist2 (vzdálenost mezi maticemi X a Y)

Výběr metrik vzdáleností u obou těchto funkcí:

- ‘euclidean’ – Euklidova metrika vzdálenosti
- ‘squaredeuclidean’ – čtverec Euklidovy metriky vzdálenosti
- ‘seuclidean’ – standardizovaná Euklidova metrika vzdálenosti
- ‘cityblock’ – Hammingova (manhattanská) metrika vzdálenosti
- ‘minkowski’ – Minkovského metrika vzdálenosti
- ‘chebychev’ – Čebyševova metrika vzdálenosti
- ‘mahalanobis’ – Mahalanobisova metrika vzdálenosti
- ‘cosine’ – 1 mínus kosinová podobnost
- ‘correlation’ – 1 mínus Pearsonův korelační koeficient
- ‘spearman’ – 1 mínus Spearmanův korelační koeficient
- ‘hamming’ – Hamminova vzdálenost (pro kvalitativní proměnné)
- ‘jaccard’ – 1 mínus Jaccardův koeficient
- lze případně nadefinovat i jinou metriku

Matlab – hierarchické aglomerativní shlukování

- funkce *linkage*, která umožňuje volbu shlukovacího algoritmu i volbu metriky vzdálenosti mezi objekty (subjekty)
- volba shlukovacího algoritmu:
  - „average“ – metoda průměrné vazby (nevážená) (average linkage)
  - „centroid“ – centroidová metoda (nevážená)
  - „complete“ – metoda nejvzdálenějšího souseda (complete linkage)
  - „median“ – centroidová metoda (vážená) = mediánová metoda
  - „single“ – metoda nejbližšího souseda (single linkage)
  - „ward“ – Wardova metoda
  - „weighted“ – metoda průměrné vazby (vážená)
- volba metriky vzdáleností – stejná nabídka jako u funkce *pdist*
- ukázka:

```
[num, txt] = xlsread('Data_neuro_shlukovky.xlsx',1);
data=num(:,[23,24,26]);

Z=linkage(data,'complete','euclidean');     % provedeni shlukove analyzy
dendrogram(Z)                                % vykresleni dendrogramu

c=cluster(Z,'maxclust',3); % vytvoreni definovaneho poctu shluku
crosstab(c,num(:,3))          % zjisteni, kolik subjektu bylo spatne zarazenych
```

Matlab – nehierarchické shlukování

- funkce *kmeans*
- ukázka:

```
[idx,C]=kmeans(data,3);    % provedeni shlukove analyzy (matice C – centroidy skupin)
crosstab(idx,num(:,3)) % zjisteni, kolik subjektu bylo spatne zarazenych
```

- funkce *kmedoids*
- bohužel není ve starých verzích Matlabu
- ukázka:

```
[idx,C]=kmedoids(data,3);    % provedeni shlukove analyzy (matice C – medoidy skupin)
crosstab(idx,num(:,3)) % zjisteni, kolik subjektu bylo spatne zarazenych
```