FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

SHLUKOVÁ ANALÝZA

Autor textu:
Ing. Petr Honzík, Ph.D.

Květen 2014

Komplexní inovace studijních programů a zvyšování kvality výuky na FEKT VUT v Brně
OP VK CZ.1.07/2.2.00/28.0193

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Obsah přednášky

1. Shluková analýza
2. Podobnost objektů
3. Hierarchické shlukování
4. Nehierarchické shlukování
5. Optimální počet shluků
6. Další metody

Učení bez učitele

- **není dána výstupní klasifikace** (veličina *G)*
- cíl: vytvoření shluků, tedy nalezení takových skupin záznamů, které jsou si navzájem podobnější než záznamy ostatní
- cíl: určení (vytvoření) výstupního vektoru *G*
- aplikace: analýza DNA (hledání charakteristických sekvencí); analýza trhu, rozpoznání nových segmentů; analýza obrazu – rozpoznání hran a objektů; web-mining – rozlišování různých typů dokumentů; bezdrátová komunikace – nalezení center skupin čidel
- základní metody jsou **shluková analýza** a **asociační pravidla**

Formulace úlohy

- χ je množina N objektů. Rozklad $\Omega=\{C_1,C_2,\ldots,C_m\}$ množiny χ je množina disjunktních, neprázdných podmnožin, které dohromady tvoří χ. Pro $i \neq j$:

$$C_i \cap C_j = 0 \quad C_1 \cup C_2 \cup \ldots \cup C_m = \chi$$

- každá množina C_i se nazývá **kmponentou rozkladu**
- shlukování je pak takový rozklad množiny χ, který maximalizuje vzájemnou mezishlukovou nepodobnost (nebo minimalizuje podobnost)

Metody a nástroje

- shluková analýza **hierarchická** vs. **nehierarchická**
- hierarchická
  - systém navzájem různých neprázdných podmnožin, přičemž průnik libovolných dvou podmnožin je jedna z nich nebo množina prázdná
  - v systému existuje alespoň jedna dvojice podmnožin, jejíž průnik je jedna z těchto množin
  - graf (binární strom) znázorňující hierarchické shlukování se nazývá **dendrogram**
- nehierarchická
  - různé neprázdné podmnožiny s prázdným průnikem

? jak se dělí medoty shlkuování

Vzdálenost mezi instancemi, metrika

- Normalizace veličin
  - aby byly vstupní veličiny souměřitelné, jsou normalizovány
  - typickou normalizací je standardizované normální rozložení se střední hodnotou 0 a rozptylem 1 (ze *z* na *x*)

$$\bar{z}_j = \frac{1}{N}\sum_{i=1}^{N} z_{ij} \qquad s_j = \left[\frac{1}{N}\sum_{i=1}^{N}\left(z_{ij} - \bar{z}_j\right)^2\right]^{\frac{1}{2}} \qquad x_{ij} = \frac{z_{ij} - \bar{z}_j}{s_j}$$

- Metriky
  - Minkowského (p-norma)
  - Manhattanská (též City block)
  - Euklidovská
  - Hammingova
  - …

$$d_p(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/p}$$

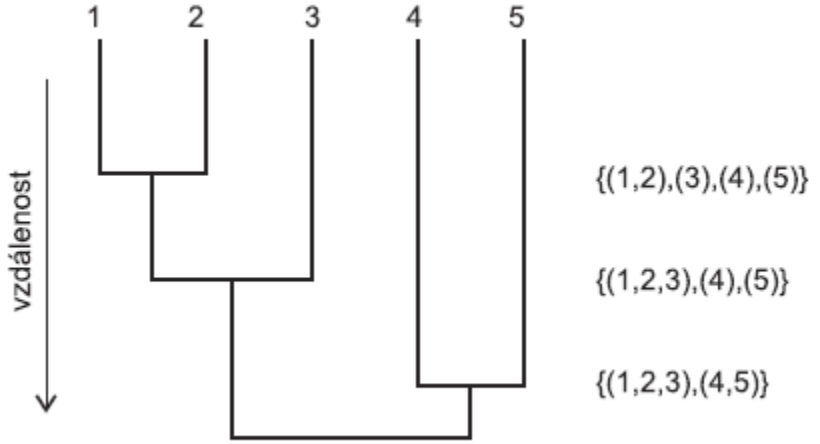
? z jakého důvodu se prování normalizace veličin

Vzdálenost mezi shluky

- Hierarchické metody shlukování dle metriky určující vzdálenost mezi dvěma shluky
  - metoda průměrová (průměr ze všech vzájemných dvojic)
  - metoda centroidní (vzdálenost aritmetických průměrů prvků z každého shluku)
  - metoda nejbližšího souseda (dva nejbližší z obou shluků)
  - metoda nejvzdálenějšího souseda (2 nejvzdálenější ze shluků)
  - metoda mediánová (vzdálenost středního z prvků obou shluků)

Hierarchické metody - princip

- Dělení podle postupu tvorby
  - aglomerativní (od jednotlivých prvků=shluků postupným slučováním až k jednomu shluku)
  - divizní (na počátku jeden velký shluk, postupný rozklad až po jednotlivé prvky, jen první dělení $2^{N-1}-1$ možností)
- dendrogram – binární strom znázorňující hierarchické shlukování; každý uzel představuje shluk; řezy stromu představují jednotlivé rozklady



Co je to dendrogram? Co je divizní shlukování, jakou má nevýhodu?

Hierarchické metody - vlastnosti

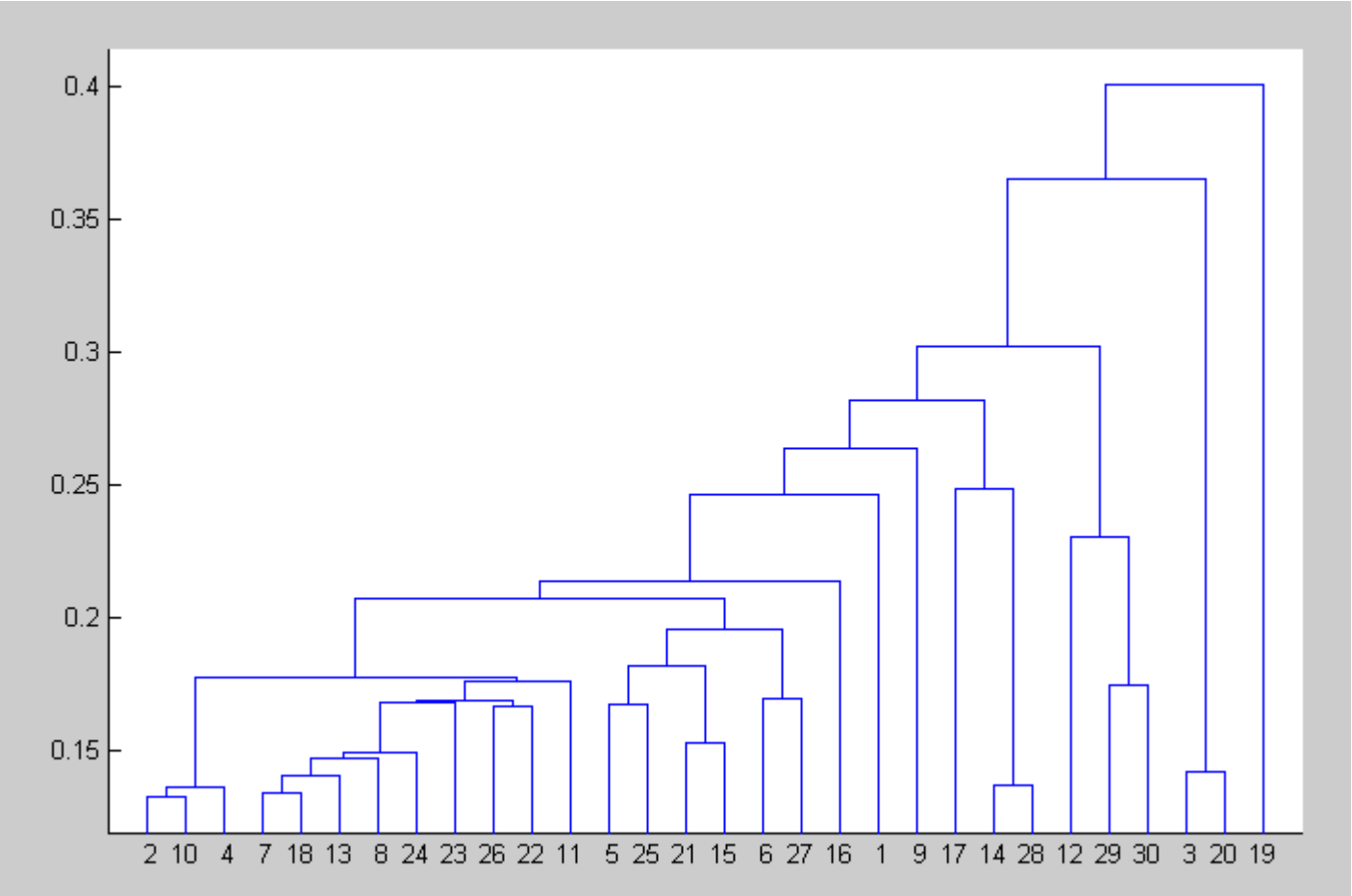
- při opakovaném pokusu nachází stejné řešení
- schopnost sledovat i „linii“, nejen podobu co do „oblastí“ (závisí na typu metriky měřící vzájemnou vzdálenost shluků)
- výstupem je dendrogram
- aglomertaivní metoda – jednoduchý algoritmus

Jsou hierarchické metody deterministické nebo stochastické?

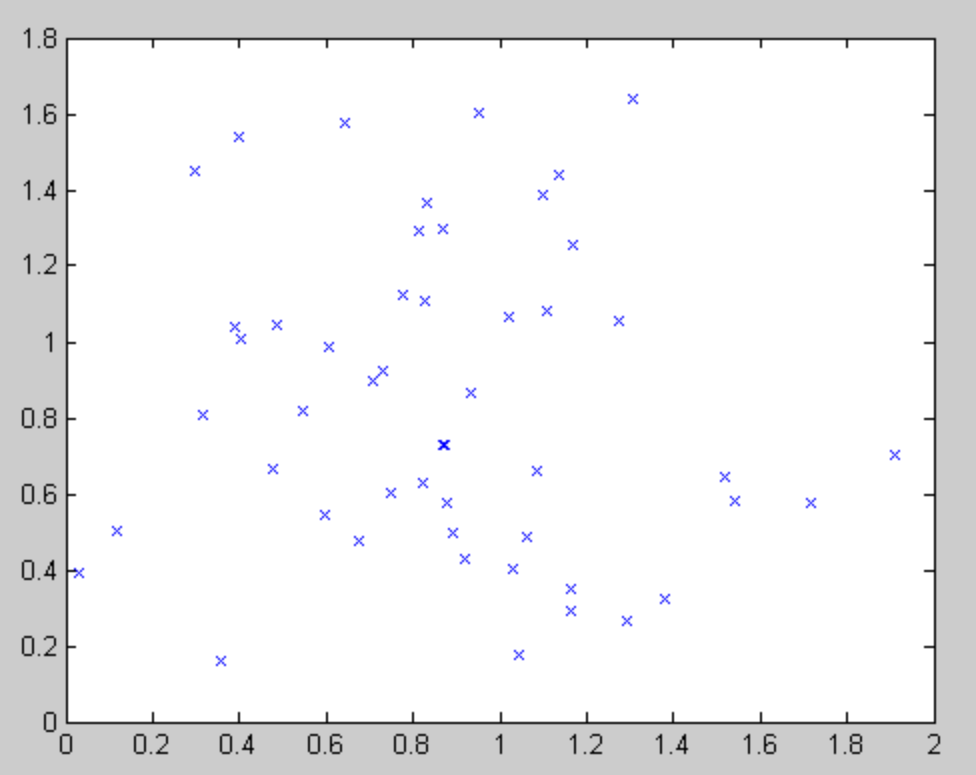
Hierarchické metody - algoritmus

- Aglomerativní metoda
  - nultý rozklad, každý jednotlivý prvek je považován za samostatný shluk
  - v i-tém kroku najdeme takovou dvojici shluků, jejíž vzdálenost je nejmenší (nepodobnost je nejmenší) a tyto shluky sjednotíme
  - v posledním kroku jsou shluky spojeny a vzniká jeden shluk odpovídající celé definiční množině

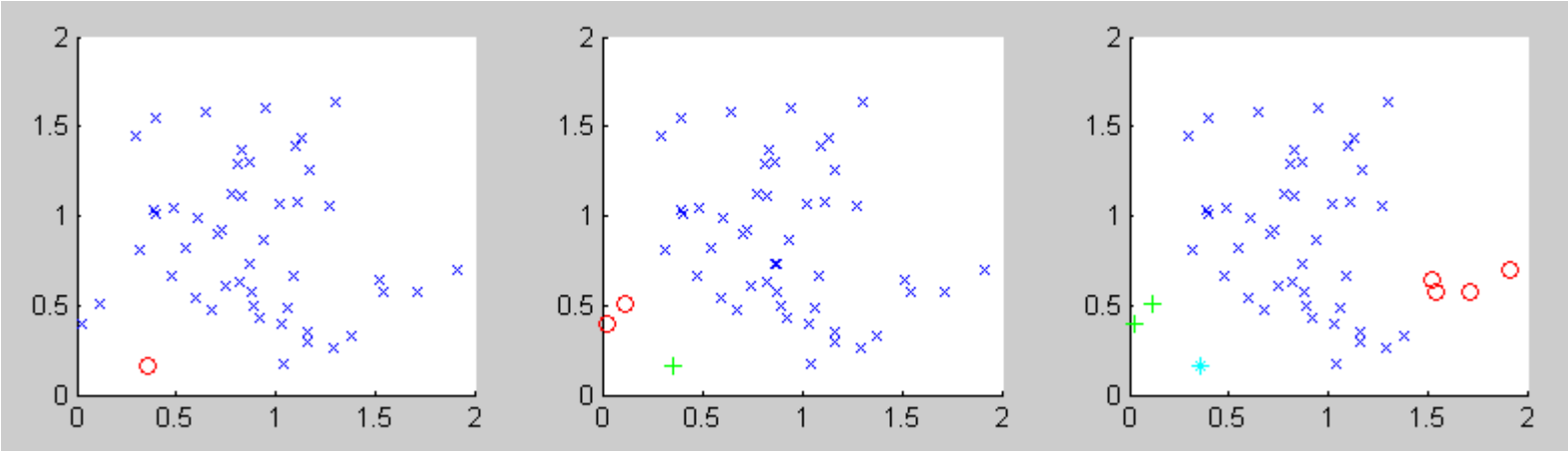
Hierarchické metody (centroidní) – příklad 1



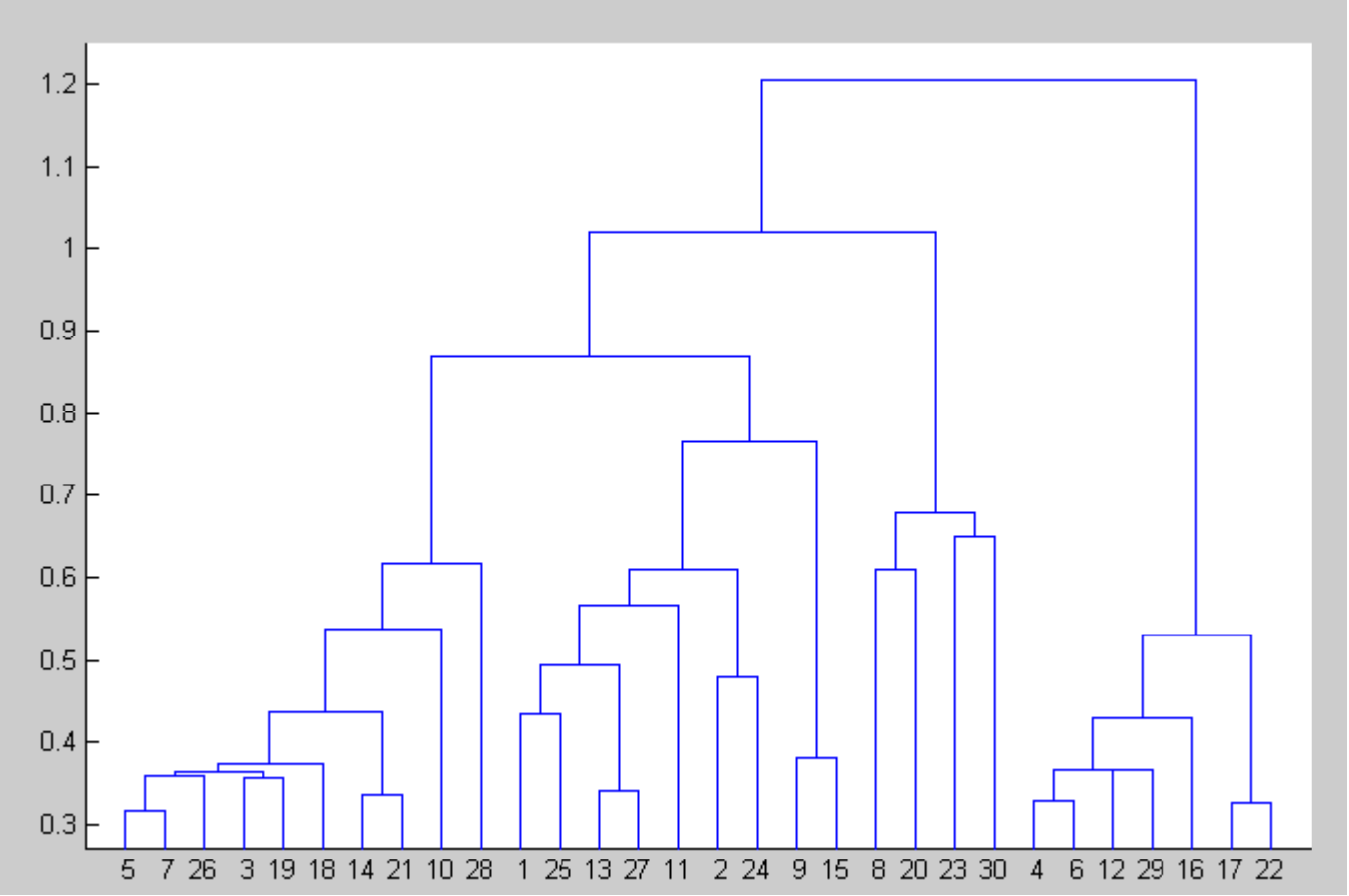
Hierarchické metody – příklad 1



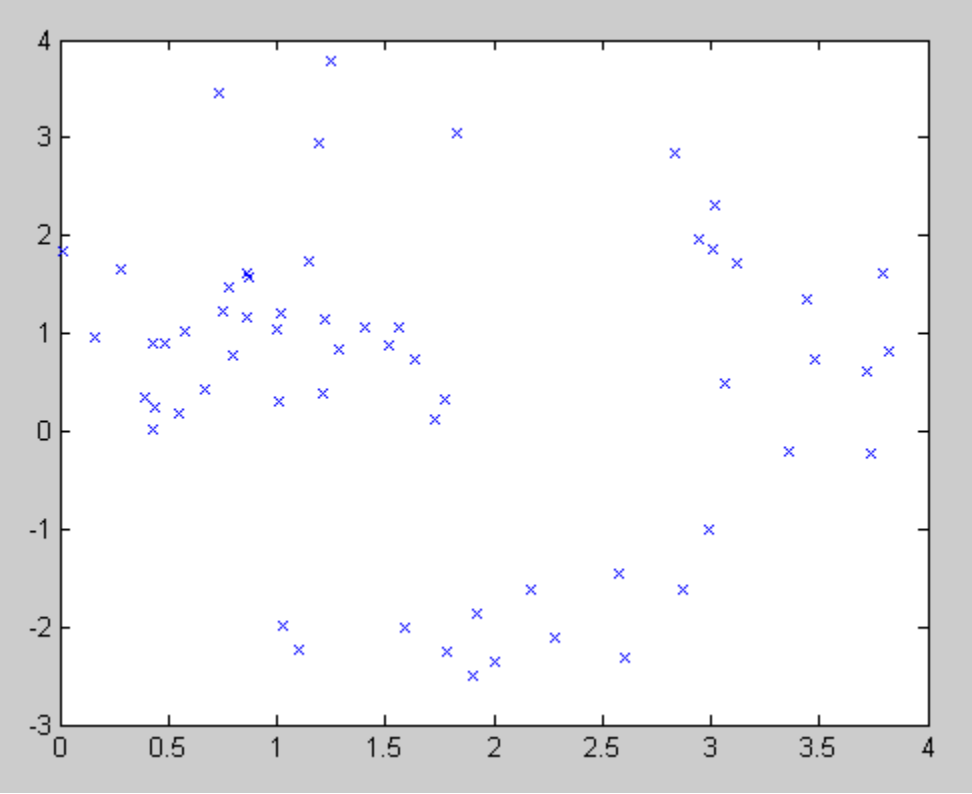
Hierarchické metody – příklad 1



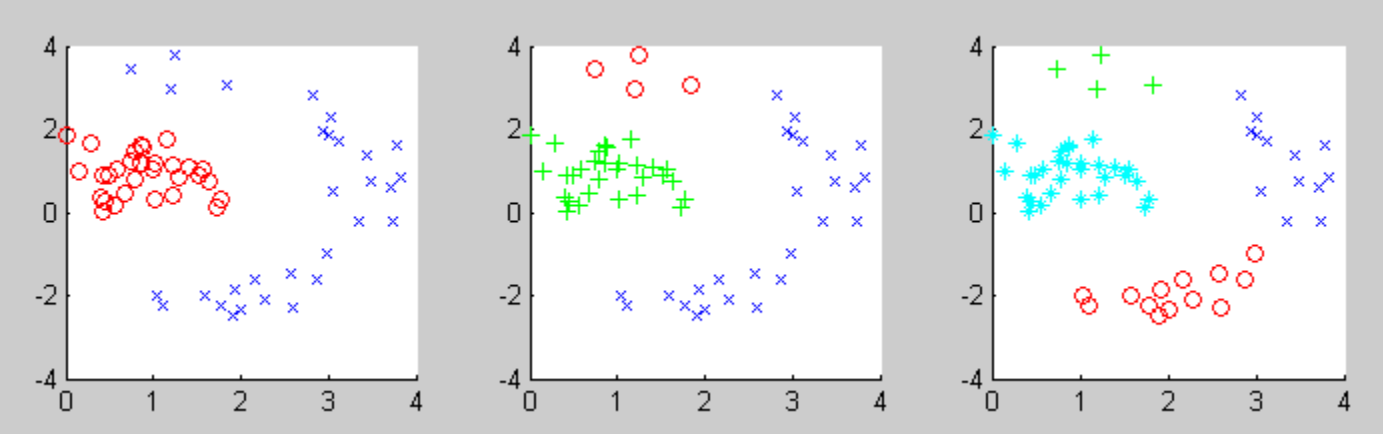
Hierarchické metody (nejbližší s.) – příklad 2



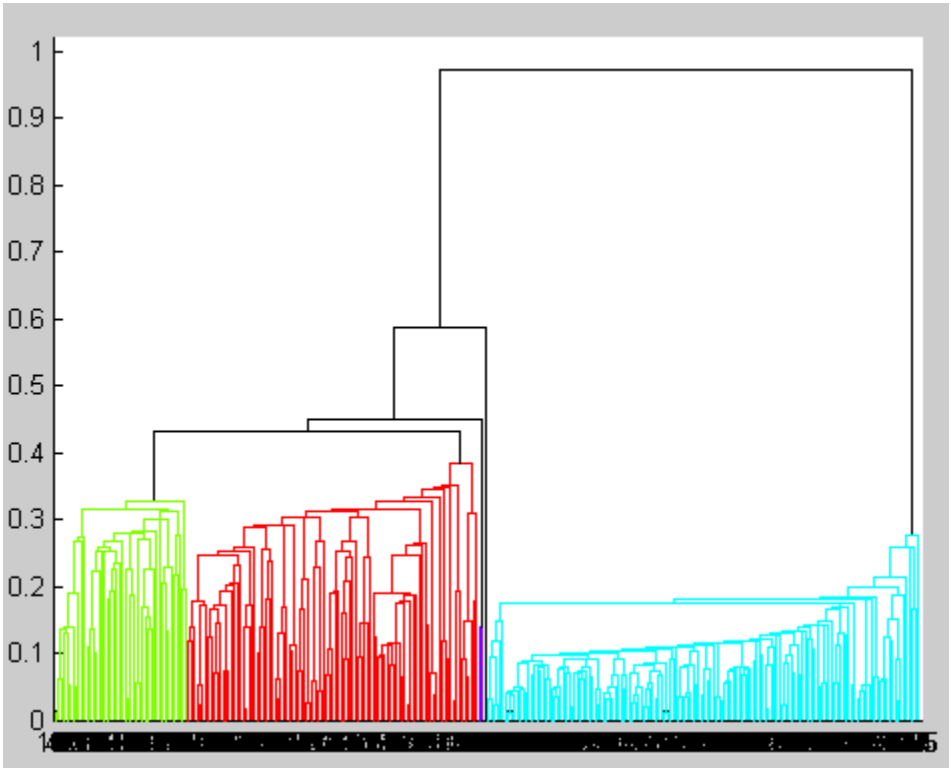
Hierarchické metody – příklad 2

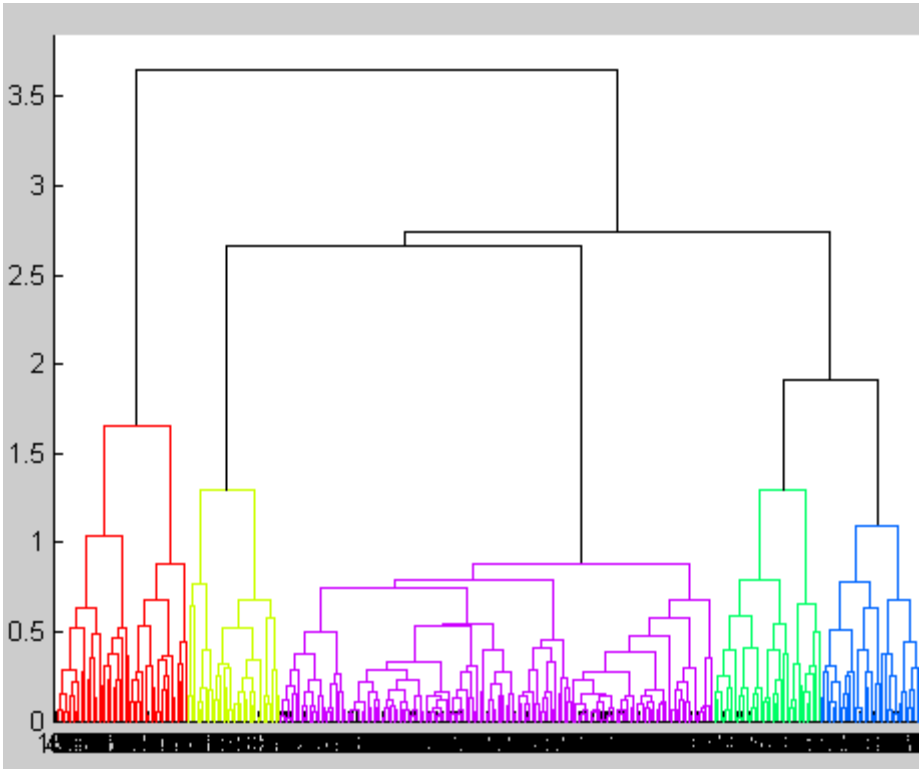


Hierarchické metody – příklad 2

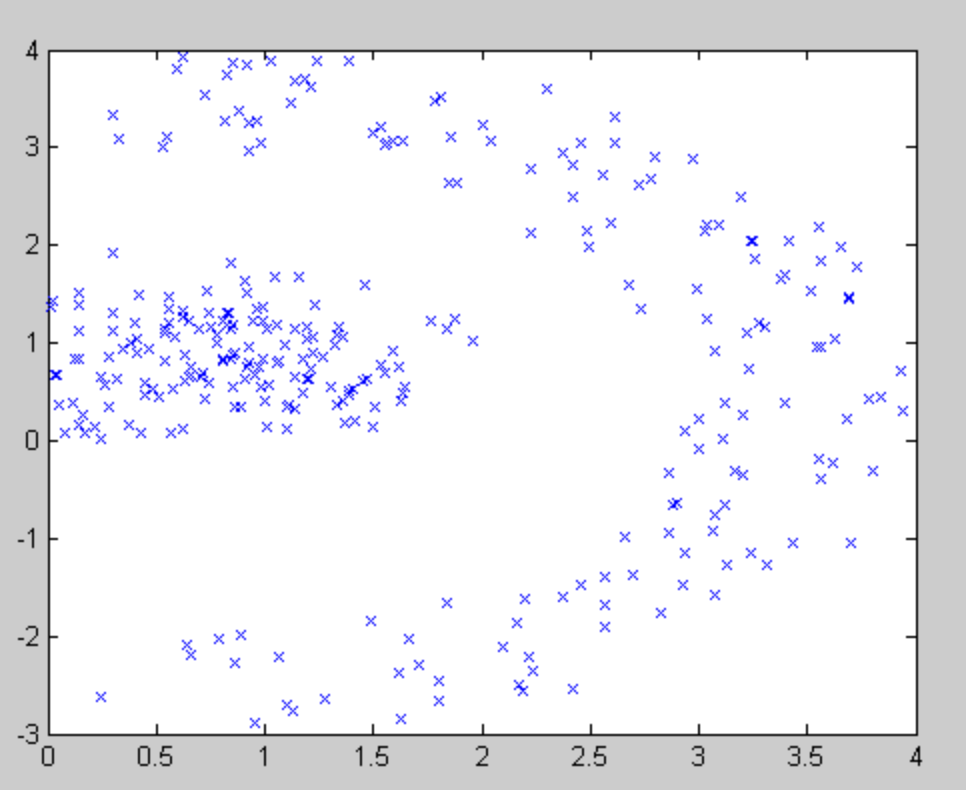


Hierarchické metody (NS,C) – příklad 3

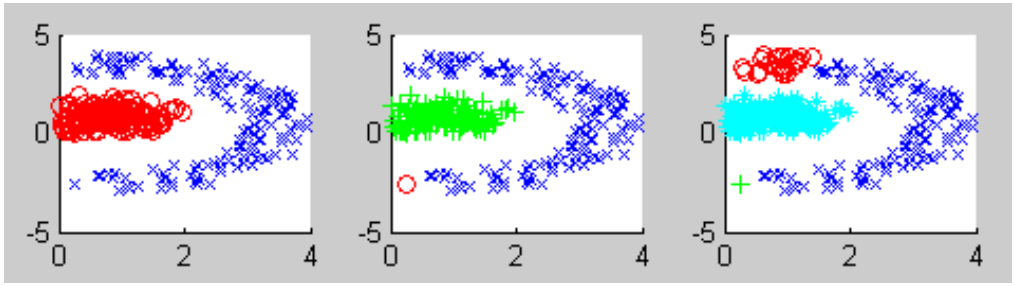


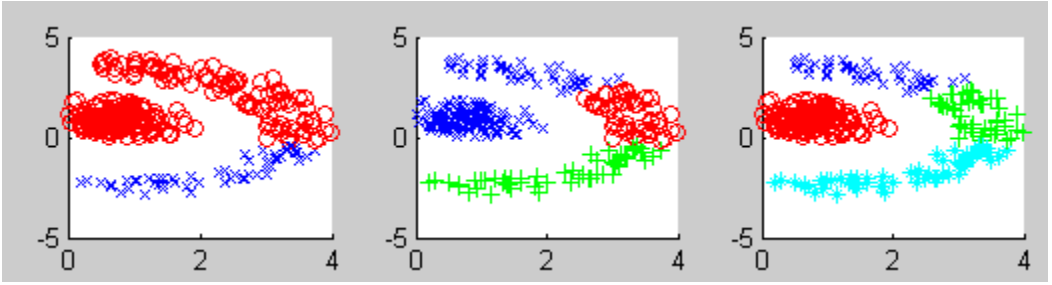


Hierarchické metody – příklad 3



Hierarchické metody – příklad 3





K-means –vlastnosti

- při vytváření malého počtu shluků z velkého počtu dat
- pro data kvantitativní bez odlehlých hodnot (je třeba normalizovat jednotlivé veličiny)
- pokaždé vede k jinému řešení
- rychle iteruje i při velkém počtu dat, nalezení zpravidla lokálního optima
- vhodné pro větší počet dat

Je metoda K-means deterministická?

K-means – princip

- dáno: data ***X,*** počet shluků K (cíleně, náhodně, heuristiky)
- cíl: nalezení K shluků tak, že mezishluková suma čtverců bude minimalizována
- princip
  - určení počátečních K charakteristických vektorů μ
  - v cyklu opakuj:
    - podle charakteristckého vektoru μ přiřaď všem bodům jejich třídu
    - spočítej z bodů jednoho shluku jejich nový charakteristický vektor μ podle těžiště nebo průměru
  - ukončovací podmínka:
    - konec, pokud 1 (nebo 2) nová iterace nezpůsobí změnu klasifikace ani jednoho z prvků

K-means – algoritmus

- *K* shluků, vstupní data $x_1,\ldots,x_N$

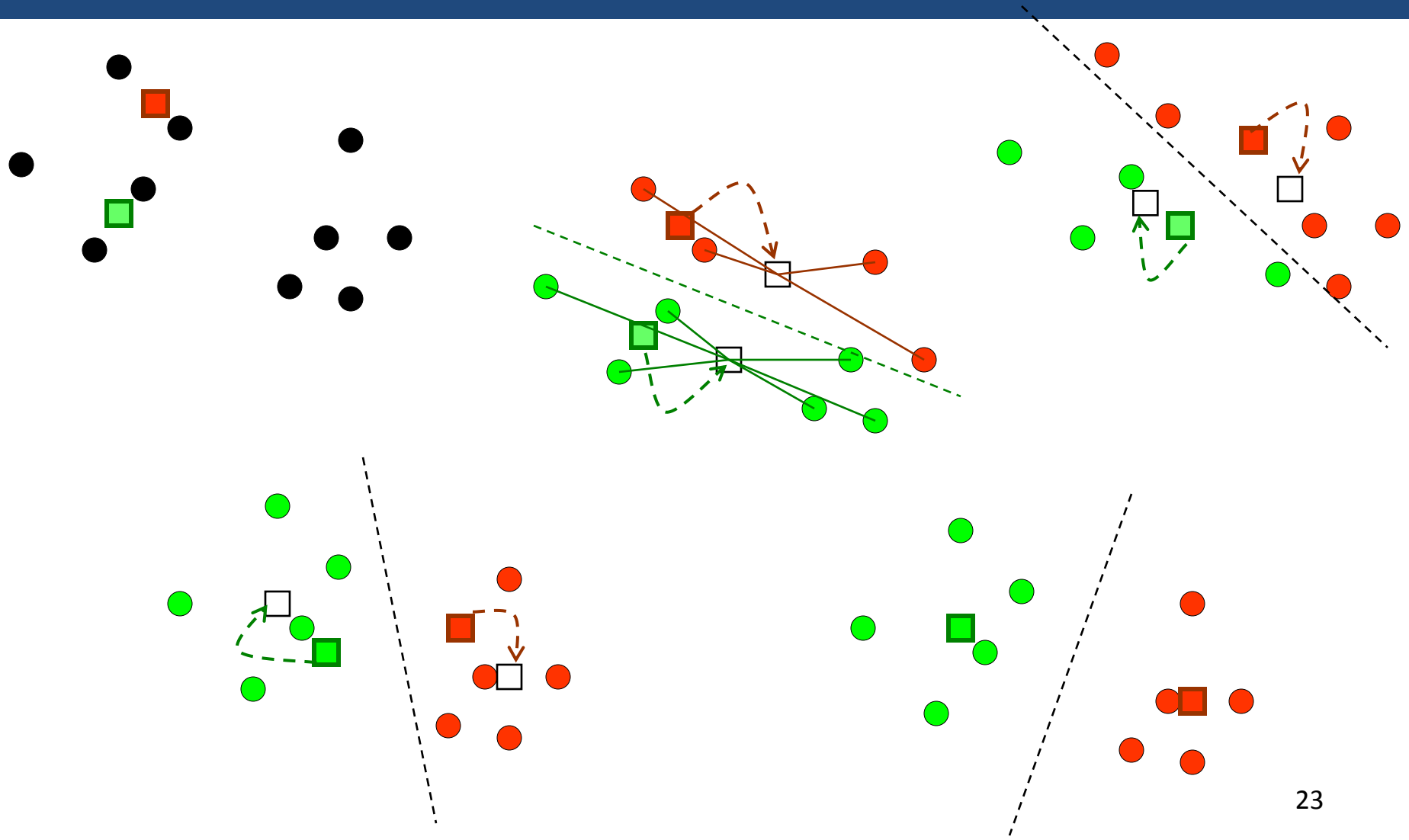
1. Počáteční určení K center (náhodné, apriorní znalost, těžiště všech bodů a extrémy, atd.) – μ_j
2. Klasifikuj do tříd $C_1,\ldots,C_K$ podle minima euklidovské vzdálenosti od vektoru μ_j
3. Opakuj v cyklu
   - každému prvku x_i je přiřazena klasifikace g_i

$$g_i = \arg\min_{j=1,\ldots,K} \left\| x_i - \mu_j \right\|$$

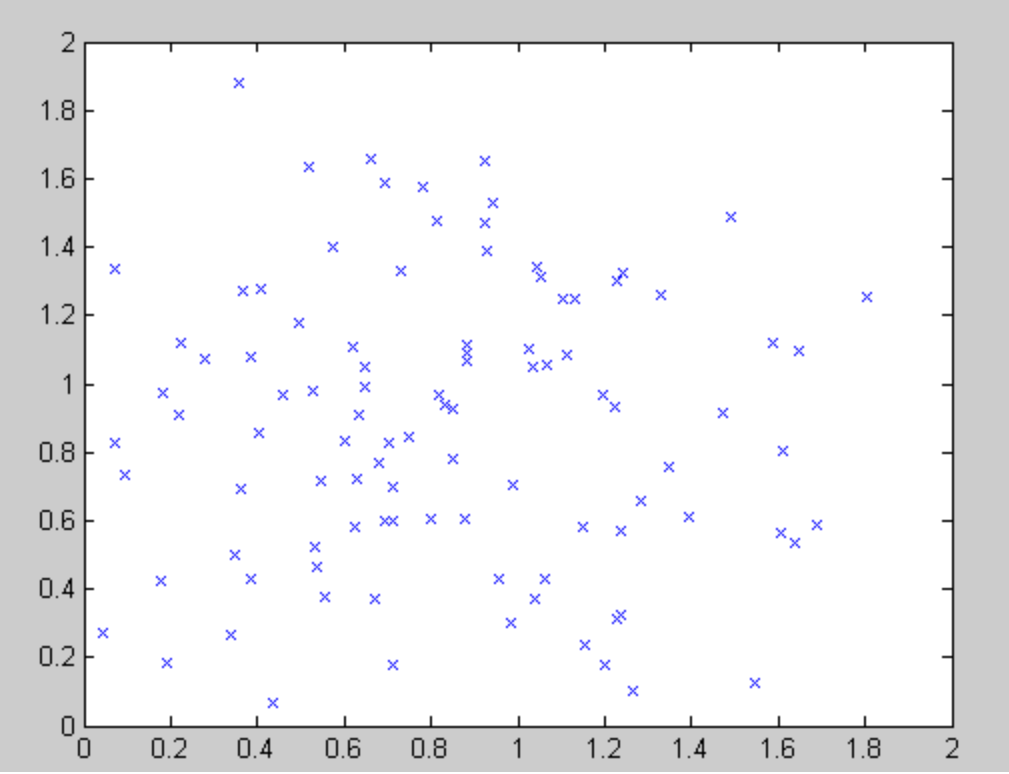
   - přepočítej vektory μ_j

$$\mu_j = \frac{1}{\left| C_j \right|} \sum_{i=1}^{N} \begin{cases} x_i \,, & g_i = C_j \\ 0 \,, & g_i \neq C_j \end{cases}$$

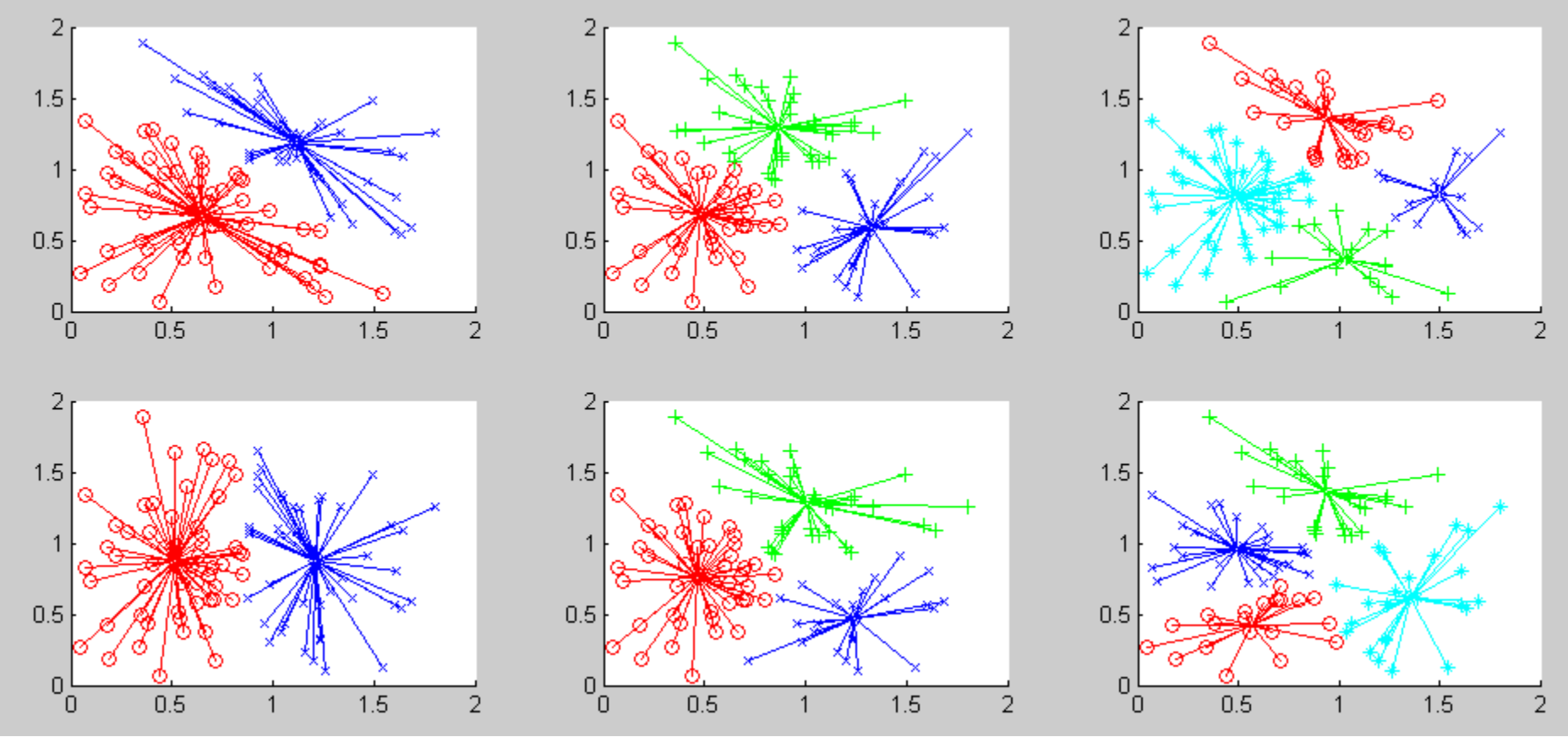
K-means – příklad



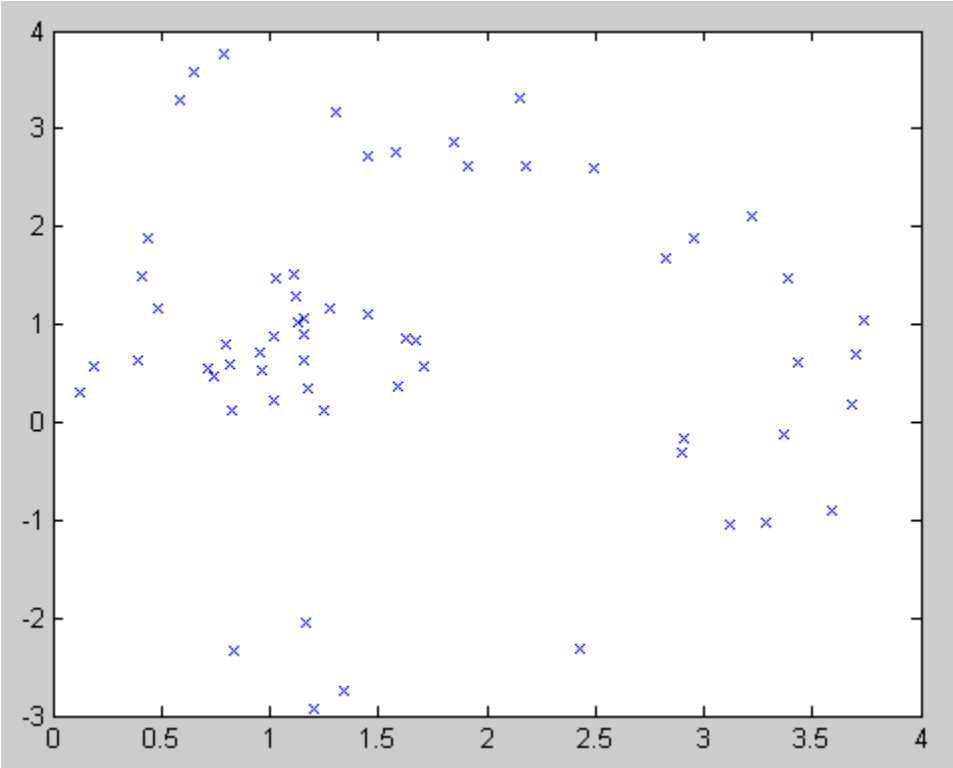
K-means – příklad 1 (náhodné rozdělení)



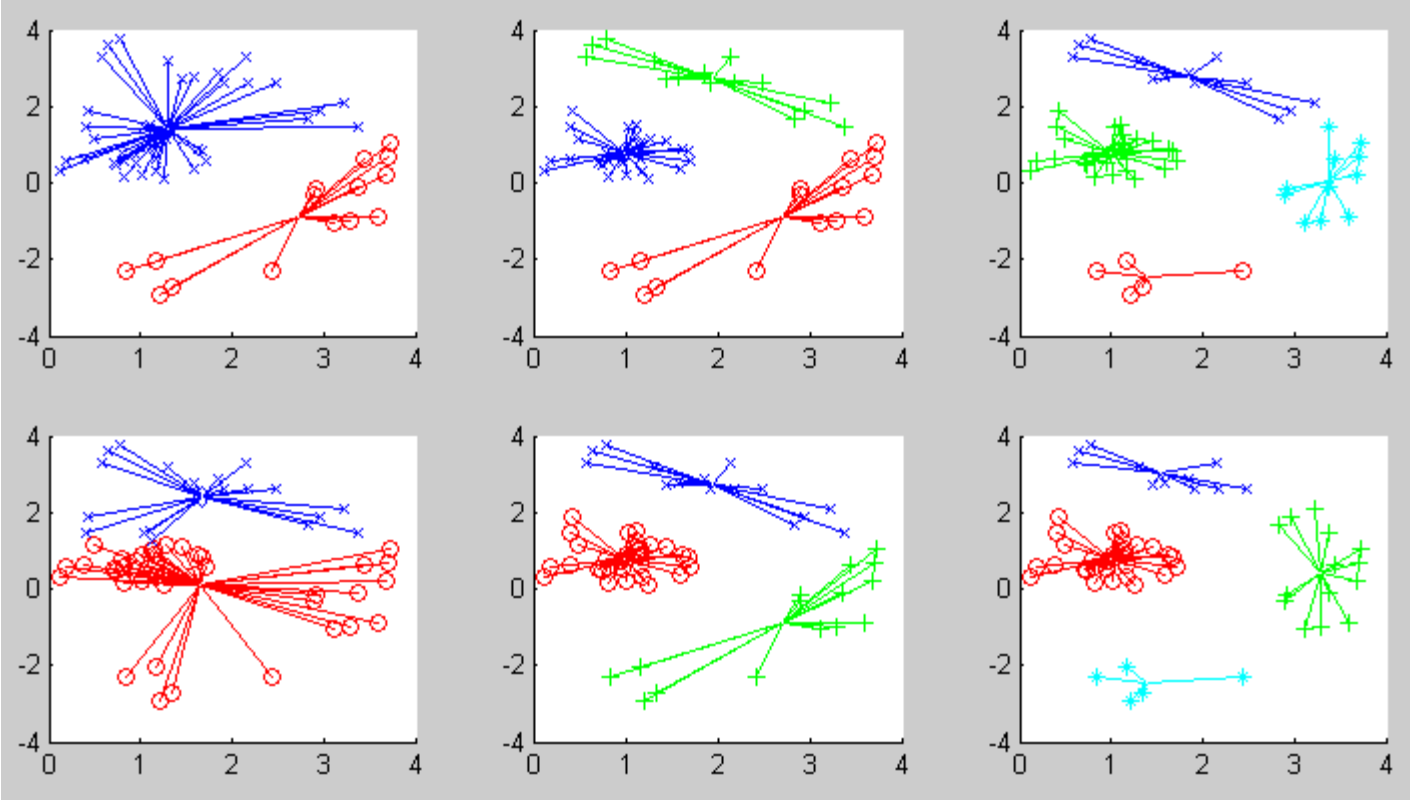
K-means – příklad 1 (náhodné rozdělení)



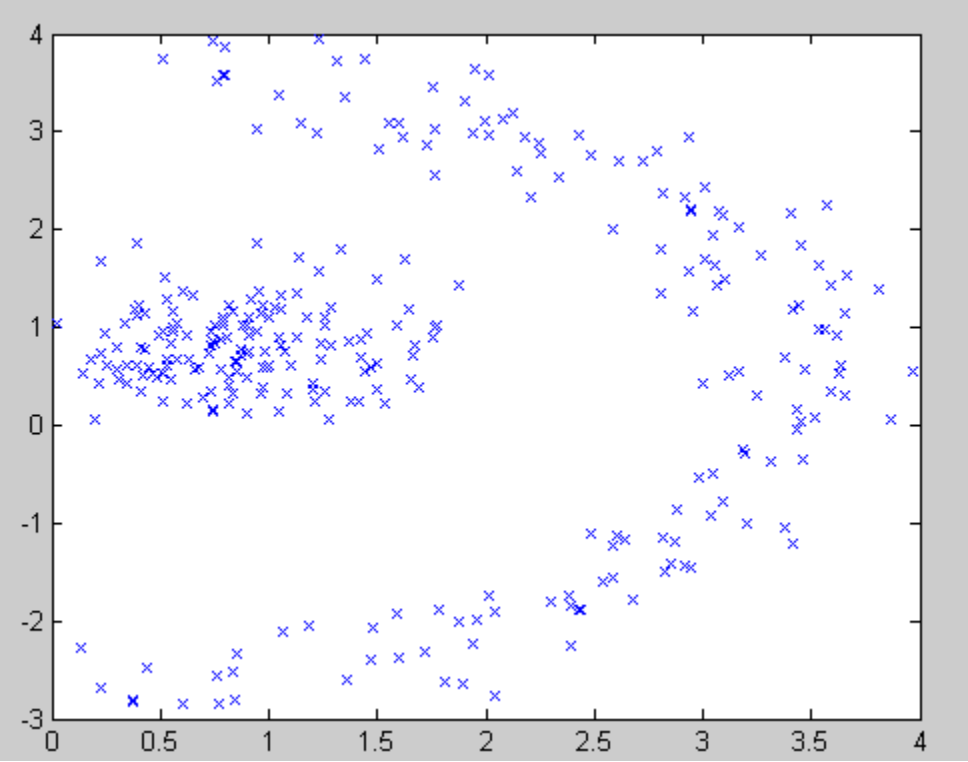
K-means – příklad 2



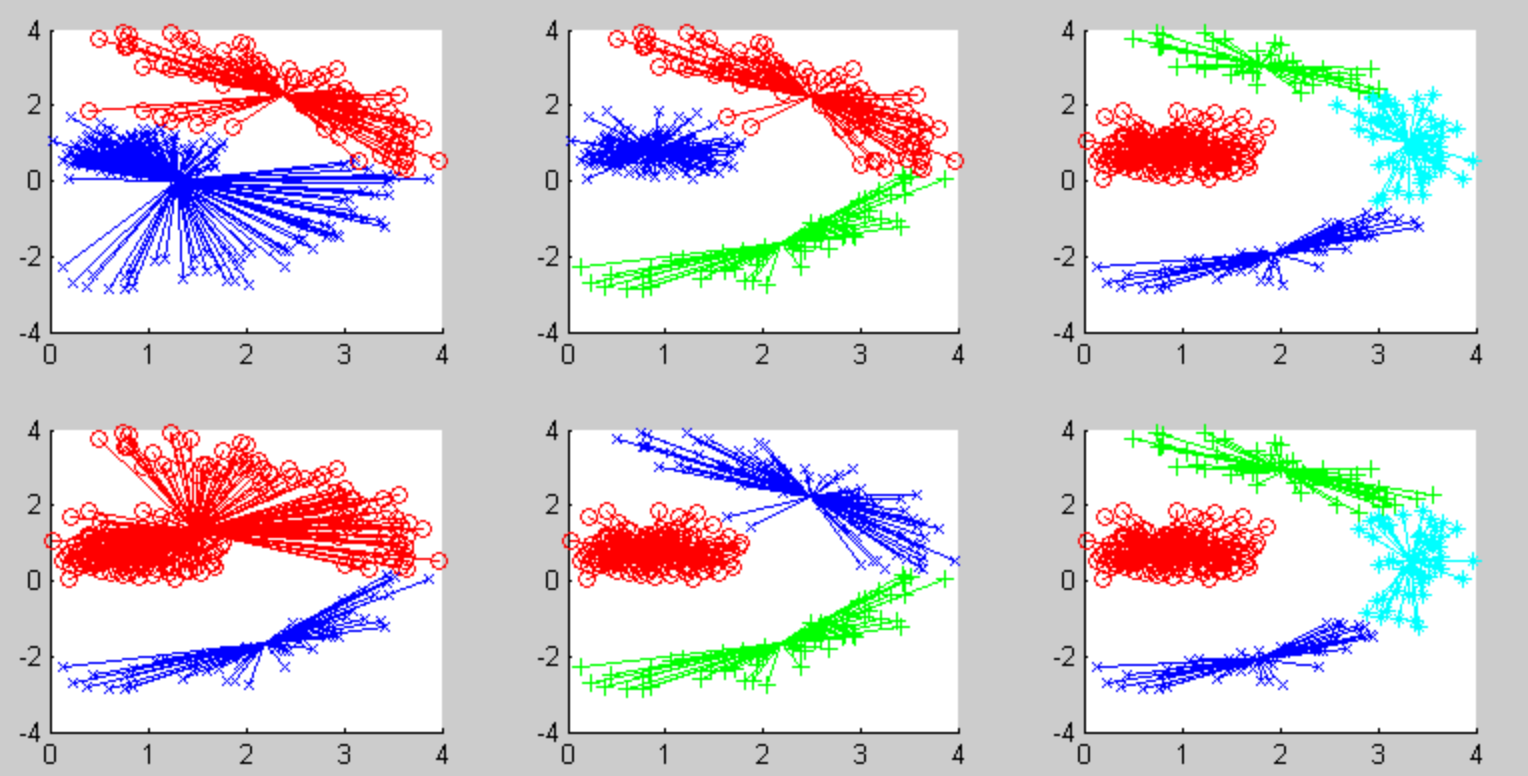
K-means – příklad 2



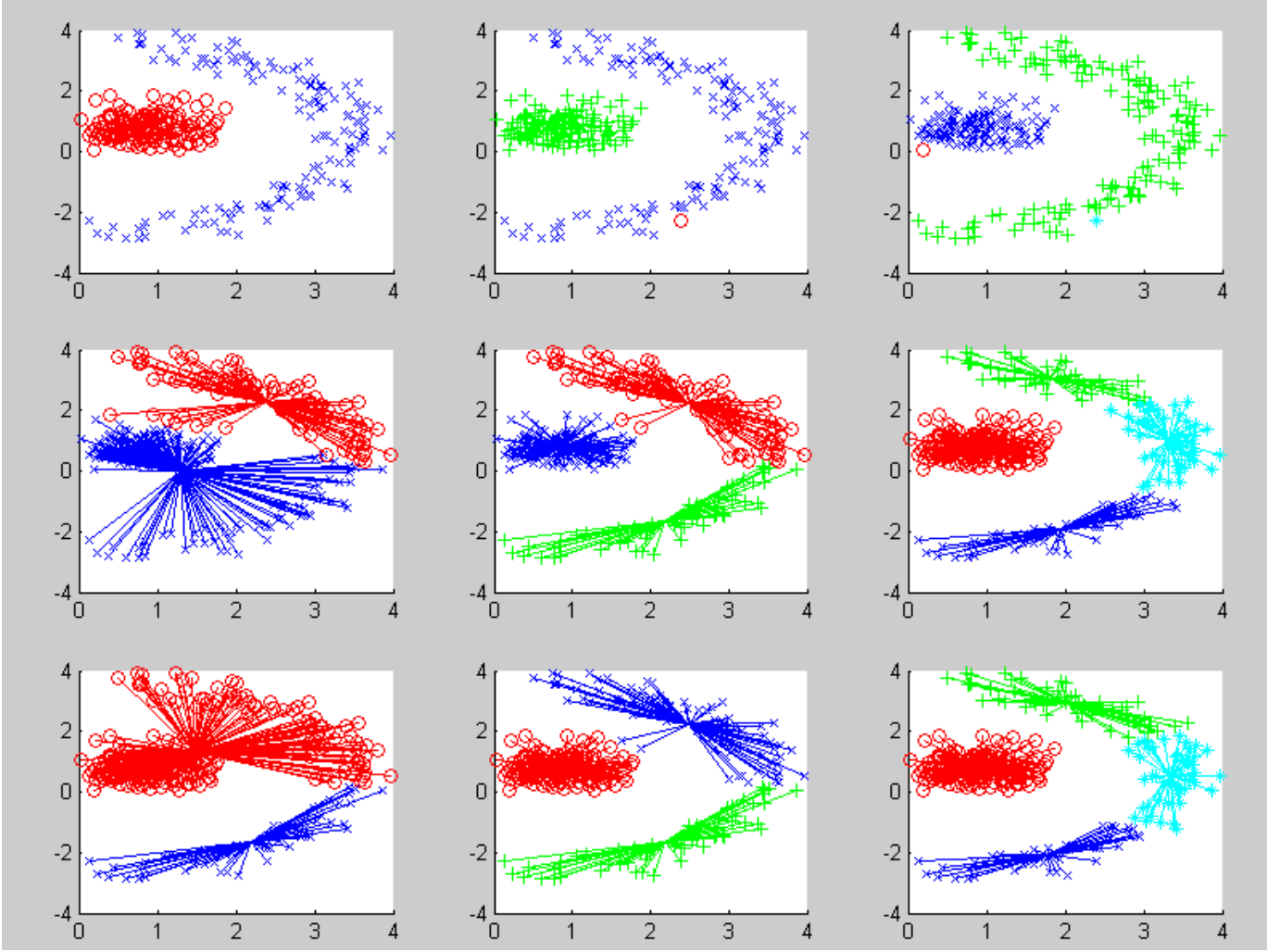
K-means příklad 3



K-means příklad 3



Srovnání hierarchické a nehierarchické metody



Optimální počet shluků

- Jak určit optimální počet shluků?
  - souvisí s mírou separability (preprocessing)
  - mělo by rozlišit „více“ a „méně“ podobné shluky
- Řada otázek
  - má záležet na počtu prvků ve shluku?
  - má být závislé na tvaru?
  - má být závislé na prostoru, který shluk zabírá?
  - má být citlivé na outliers?
  - …

Optimální počet shluků

- Absolutní srovnání
  - srovnávají se přímo **hodnoty indexů** pro konkrétní, předem dané počty shluků (typické pro k-means metody)
  - Calinski and Harabasz
  - procento proměnlivosti
  - Davies-Bouldin Index
- Diferenční srovnání
  - srovnávají se **diference indikátoru** při rostoucím počtu shluků
  - Hartigan
  - Krzanowski and Lai

V čem spočívá princip absolutních metod pro srovnání různých rozkladů?

Calinského a Harabaszova metoda 1/3

- Je dána poměrem meziskupinové a vnitroskupinové čtvercové vzdálenosti
- Vnitroskupinová čtvercová vzdálenost pro *k* shluků (within-clustr sum of squares):

$$W_k = \sum_{r=1}^{k} \sum_{i \in C_r} \left\| x_i - \bar{x}^{C_r} \right\|^2$$

- kde
  - $\bar{x}^{C_r}$ je průměr prvků ve shluku C_r
  - *k* je počet shluků

Calinského a Harabaszova metoda 2/3

- Meziskupinová čtvercová vzdálenost pro *k* shluků (between-clustr sum of squares):

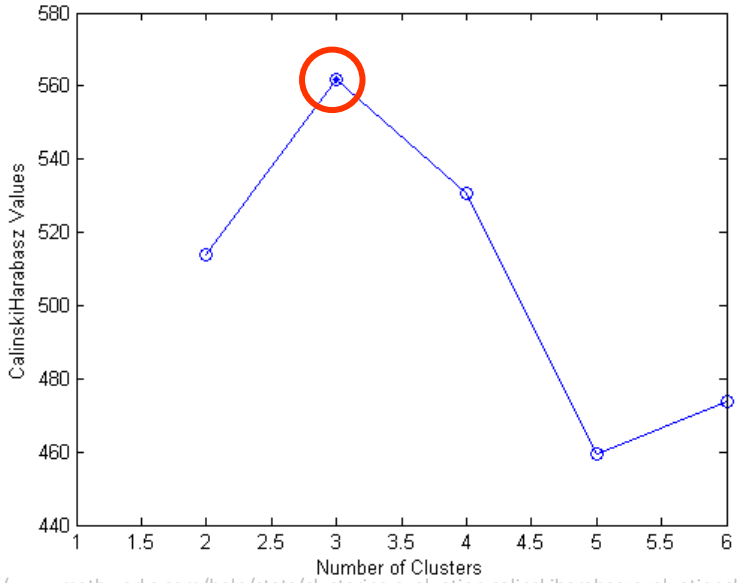
$$B_k = \sum_{r=1}^{k} |C_r| \left\| \bar{x}^{C_r} - \bar{x} \right\|^2$$

  - $\bar{x}^{C_r}$ je průměr prvků ve shluku C_r
  - *k* je počet shluků

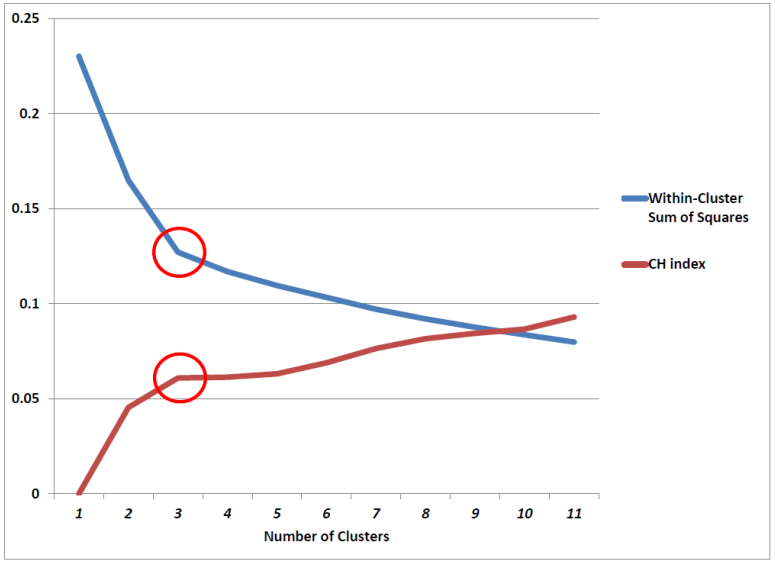
- Míra separability shluků (čím větší, tím lepší) jest:

$$I_{CH} = \frac{B_k / (k-1)}{W_k / (N-k)} = \frac{B_k}{W_k} \cdot \frac{(N-k)}{(k-1)}$$

Calinského a Harabaszova metoda 3/3



http://www.mathworks.com/help/stats/clustering.evaluation.calinskiharabaszevaluationclass.html



http://businessforecastblog.com/wp-content/uploads/2012/10/Ncluster2.png

a) Maximální hodnota pro k=3, v datech se nacházejí 3 klastry

b) Maximální hodnota pro k=N (počet prvků), stanovení počtu shluků není jednoznačné

Procento proměnlivosti PV_k

- Využívá **mezishlukové sumy čtverců** WSS_k vypočtené z vnitroskupinové vzdálenosti W_k

$$WSS_k = \frac{Nm}{Nm - m} \sum_{r=1}^{k} W_r$$

  - N je počet prvků
  - m je počet veličin popisujících prvek (dimenze, počet znaků)
  - k je počet shluků

- Procento proměnlivosti (100% pro k=1, 0% pro k=N)

$$PV_k = 100\% \frac{WSS_k}{WSS_1}$$

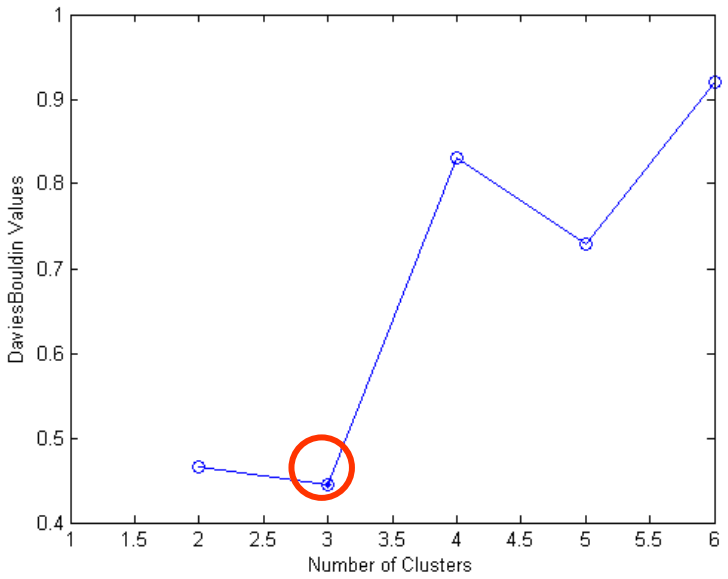
Davies-Bouldinův index

- Průměrná vzdálenost uvnitř shluku $$\Delta_i = \frac{\sum_{j \in C_i} d\left(x_j^{C_i}, \bar{x}^{C_i}\right)}{|C_i|}$$
- Vzdálenost mezi dvěma shluky $$\partial_{ij} = d\left(\bar{x}^{C_i}, \bar{x}^{C_j}\right)$$
- Výsledný index (čím je menší, tím lepší separabilita)

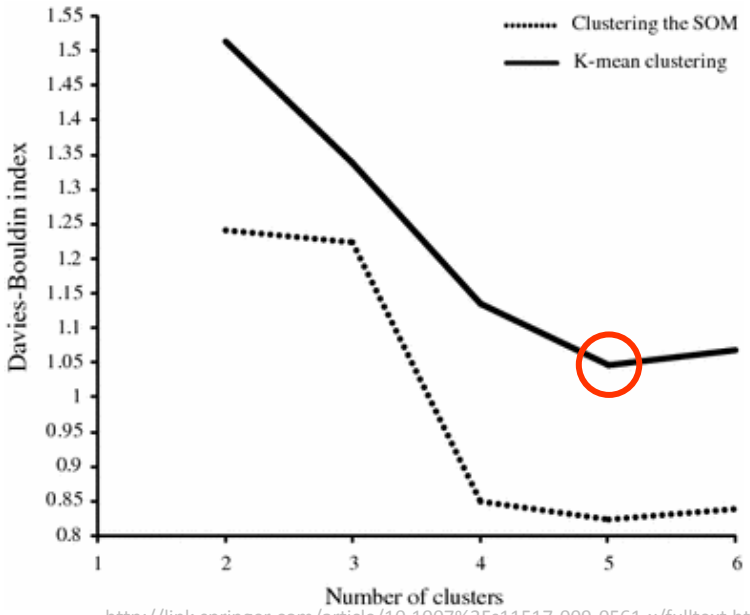
$$I_{DB} = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{\Delta_i + \Delta_j}{\partial_{ij}} \right)$$

  - $\bar{x}^{C_i}$ je průměr prvků ve shluku C_i a $x_j^{C_i}$ je j-tý prvek ve shluku C_i
  - k je počet shluků
  - $d(x,y)$ je vzdálenost mezi prvky x a y

Davies-Bouldinův index



http://www.mathworks.com/help/stats/clustering.evaluation.daviesbouldinevaluationclass.html



http://link.springer.com/article/10.1007%2Fs11517-009-0561-x/fulltext.html

Každý navržený shluk je charakterizován svou „nejhorší“ rozlišitelností vůči ostatním navrženým shlukům („max“ v definici indexu). Průměr těchto „nejhorších“ případů pak definuje index samotný. Minimum indexu určuje odhad nejvhodnějšího počtu shluků.

Diferenční indexy

- Hartiganův index

$$H(k) = \left(\frac{W_k}{W_{k+1}} - 1 \right) \Big/ (n - k - 1)$$

- Krzanowski and Lai

$$KL(k) = \left| \frac{(k-1)^{2/p} W_{k-1} - k^{2/p} W_k}{k^{2/p} W_k - (k+1)^{2/p} W_{k+1}} \right|$$

Dodatek

- Metody měnící počty shluků
  - MacQueen se dvěma parametry
  - Wishartova metoda RELOC (4 parametry)
  - Metoda ISODATA (2 parametry)