

Unsupervised learning

“učenie bez učiteľa”

II. analýza zhlukov

1. časť

Zuzana Rošťáková

14. október 2021

Seminár UM

Unsupervised learning - označenie

- vstupné údaje: na skupine **n** objektov pozorujeme **p** premenných

$$X = \begin{bmatrix} - & \mathbf{x}_1^T & - \\ - & \mathbf{x}_2^T & - \\ & \vdots & \\ - & \mathbf{x}_n^T & - \end{bmatrix} = \begin{bmatrix} | & | & & | \\ \mathbf{x}_1^* & \mathbf{x}_2^* & \dots & \mathbf{x}_p^* \\ | & | & & | \end{bmatrix} \in \mathbb{R}^{n \times p}$$

- $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$

→ vektor hodnôt **p** premenných pre *i*-ty objekt, $i = 1, \dots, n$

- $\mathbf{x}_l^* = (x_{1l}, x_{2l}, \dots, x_{nl})^T$

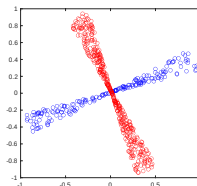
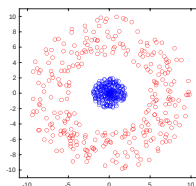
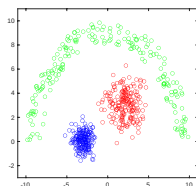
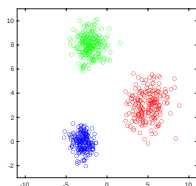
→ vektor hodnôt *l*-tej premennej, $l = 1, \dots, p$ pre **n** objektov

Analýza zhlukov

- **cieľ**: nájsť rozdelenie n pozorovaní $x_1, \dots, x_n \in \mathbb{R}^p$ do K zhlukov $C_1, \dots, C_K$

$$\bigcup_{k=1}^K C_k = \{1, \dots, n\} \quad \text{a} \quad C_i \cap C_j = \emptyset, i \neq j$$

- pozorovania v **jednom** zhluke sú si maximálne **podobné**
- pozorovania z **rôznych** zhlukov sú si maximálne **nepodobné**



⇒ potrebujeme definovať vzdialenosť medzi pozorovaniami

Typy vzdialeností medzi pozorovaniami

- **Minkowskeho**
$$D_{Min}(x, y) = (\sum_{i=1}^p |x_i - y_i|^r)^{\frac{1}{r}}$$
 - $r = 1 \rightarrow$ Manhattan / cityblock distance $D_C(x, y) = \sum_{i=1}^p |x_i - y_i|$
 - $r = 2 \rightarrow$ euklidovská vzdialenosť $D_E(x, y) = \sqrt{\sum_{i=1}^p (x_i - y_i)^2}$
- **Mahalanobisova**
$$D_{Mah}(x, y) = \sqrt{(x - y)^T S^{-1} (x - y)}$$
 - $S = I \rightarrow$ euklidovská vzdialenosť $D_E(x, y) = \sqrt{(x - y)^T (x - y)}$
- Čebyševova $D_{Che}(x, y) = \max_{i=1, \dots, p} |x_i - y_i|$
- **korelácia** $D_{corr}(x, y) = 1 - corr(x, y)$
- kosínusová $D_{cos}(x, y) = 1 - \frac{x^T y}{\sqrt{(x^T x)(y^T y)}}$
- Hammingova - skôr v teórii kódov $D_{Ham}(x, y) = \#\{i : x_i \neq y_i\}$
- ...

- nehierarchické

- **centroid-based methods**

- k -means/ k -means++, k -medoids, k -medians, ...

- **model-based clustering**

- gaussian mixture model (GMM), ...

- **density-based clustering**

- DBSCAN, OPTICS, ...

- ...

- hierarchické

- aglomeratívne

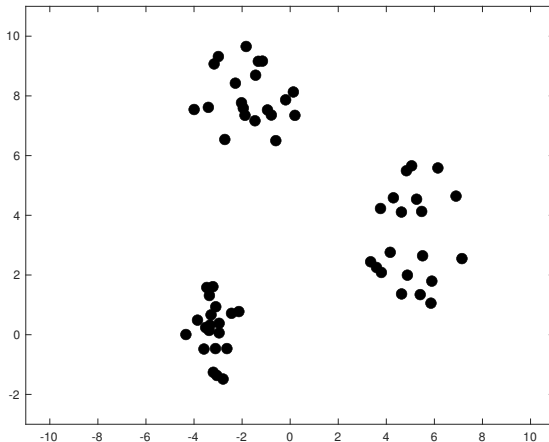
- divizívne

Nehierarchické metódy zhlukovej analýzy

1. Centroid-based clustering

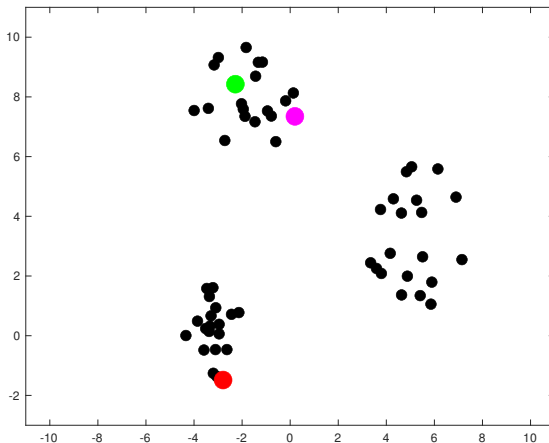
Centroid-based clustering - vstúp

- dáta $X = (x_1, x_2, \dots, x_n)^T \in \mathbb{R}^{n \times p}$
- počet zhlukov K



Centroid-based clustering - algoritmus

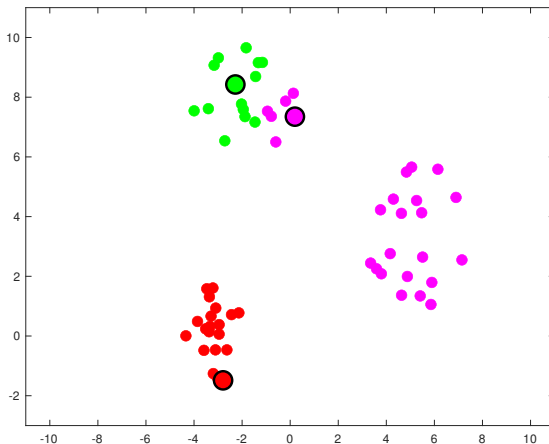
- 1.) z n pozorování $x_1, \dots, x_n$ náhodně zvolíme K
→ charakteristické prvky zhlukov: $y_{C_1}, \dots, y_{C_K}$



Centroid-based clustering - algoritmus

2.) zaradenie do zhlukov:

$$x_i \in C_j, \text{ ak } D(x_i, y_{C_j}) < D(x_i, y_{C_k}), \quad k \neq j$$



Centroid-based clustering - algoritmus

3.) vypočítame charakteristický prvok y_{C_i} i -teho zhluku, $i = 1, \dots, K$

- k -means (**centroid**)

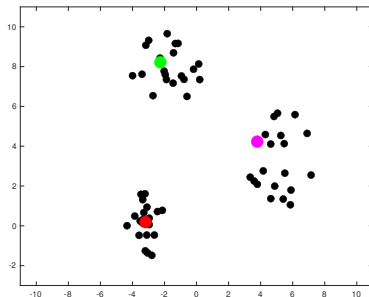
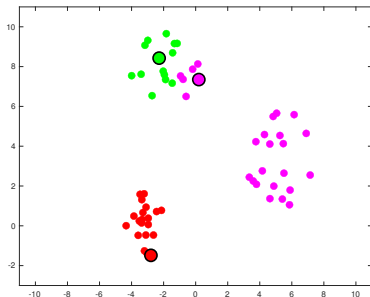
$$y_{C_i} = \frac{1}{|C_i|} \sum_{j: x_j \in C_i} x_j$$

- k -medoids (**medoid**)

$y_{C_i} \in C_i : \sum_{j: x_j \in C_i} D(x_j, y_{C_i})$ je minimálna

- k -medians (**medián**)

$y_{C_i} = p$ -rozmerný medián zhluku C_i



Centroid-based clustering - algoritmus

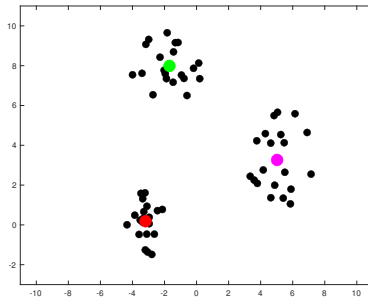
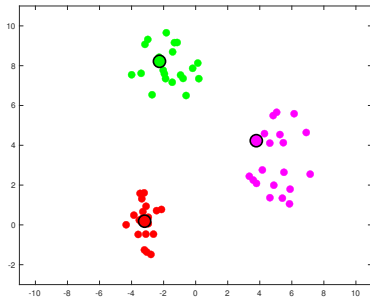
4.) opakujeme kroky 2.) a 3.), kým nie sú splnené zastavovacie kritéria

- počet iterácií

- zmena v zhľukovaní

- $$\frac{1}{n} \sum_{k=1}^K \sum_{i: x_i \in C_k} D(x_i, y_{C_k}) < \varepsilon$$

- ...



Centroid-based clustering - voľba počtu zhlukov

- apriórna informácia

- “rule of thumb” $K \approx \sqrt{\frac{n}{2}}$

- priemerná nepodobnosť v zhluchoch

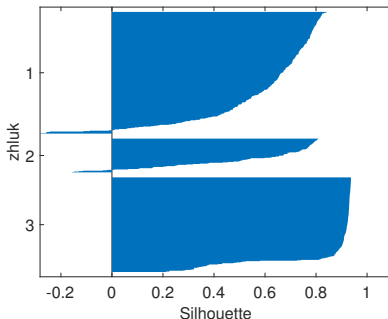
- zhluková analýza pre $k = 2, \dots, K_{max}$ zhlukov

- $$b_k = \frac{1}{n} \sum_{l=1}^k \sum_{i: x_i \in C_l} D(x_i, y_{C_l})$$

- elbow diagram pre $b_k, k = 2, \dots, K_{max} \Rightarrow$ zmena v profile grafu

Centroid-based clustering - voľba počtu zhlukov

- **silhouette** $sil(x_i) \in [-1, 1]$
 - $sil(x_i) \approx 1 \rightarrow$ pozorovanie x_i je zaradené v správnom zhluku
 - $sil(x_i) \approx 0 \rightarrow$ pozorovanie x_i leží na hranici medzi dvoma zhlukmi
 - $sil(x_i) \approx -1 \rightarrow$ pozorovanie x_i je zaradené v nesprávnom zhluku
 - v MATLAB-e **silhouette(X, ind)**



- priemerná **silhouette** pre $k = 2, \dots, K_{max}$ zhlukov
- hľadáme maximálnu priemernú **silhouette**

k -means v MATLAB-e

[IDX, C, sumD, D] = kmeans(X, K, Name, Value)

X - $n \times p$ matica pozorovaní

K - požadovaný počet zhlukov

- voliteľné parametre

- Distance - sseuclidean, cityblock, cosine, correlation, hamming
- MaxIter - maximálny počet iterácií (≥ 100)
- Replicates - počet opakovaní algoritmu, iné inicializačné centroidy
- Start - voľba inicializačných centroidov (plus, sample, uniform, cluster, ...)
- ...

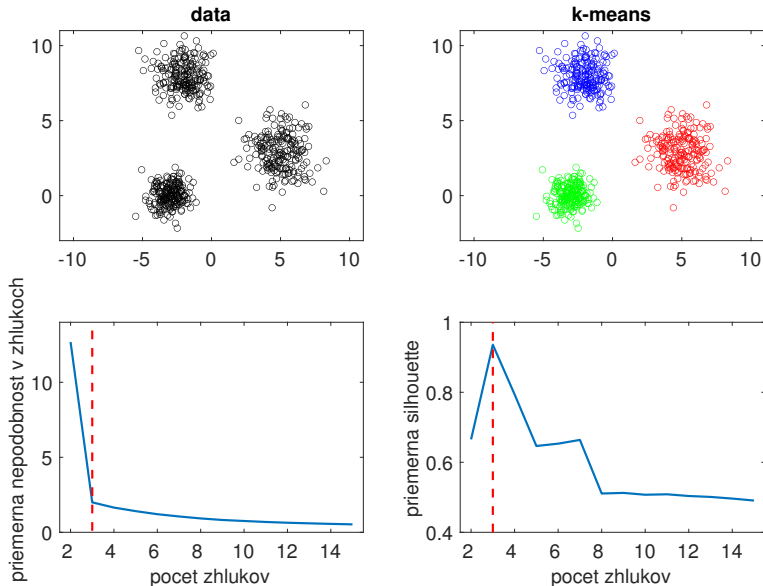
IDX $\in \mathbb{R}^n$ = indexy zhlukov pre n prvkov

C $\in \mathbb{R}^{K \times p}$ = centroidy zhlukov

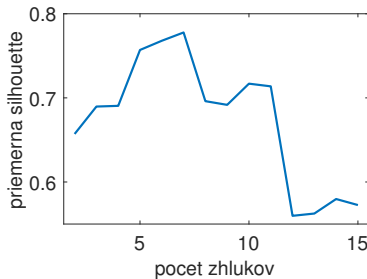
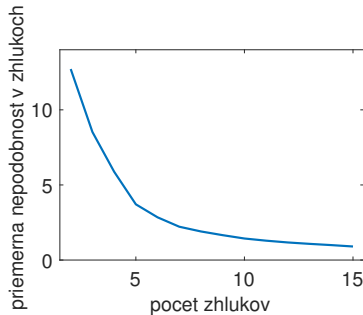
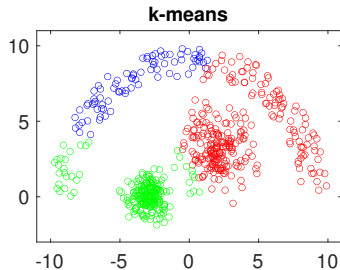
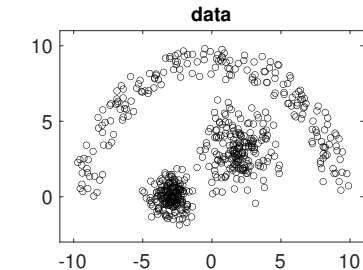
sumD $\in \mathbb{R}^K$ = súčet vzdialeností medzi prvkami zhliku a centroidom

D $\in \mathbb{R}^{n \times K}$ = vzdialenosť každého bodu od každého centroidu

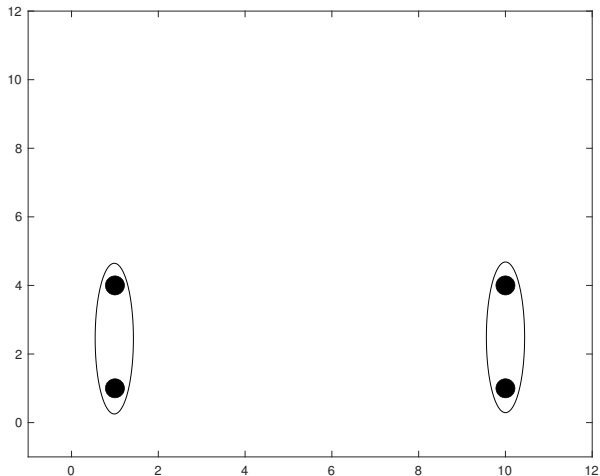
k-means - Príklad: 3 sférické zhľuky



k-means - Príklad: “banánový” zhluk

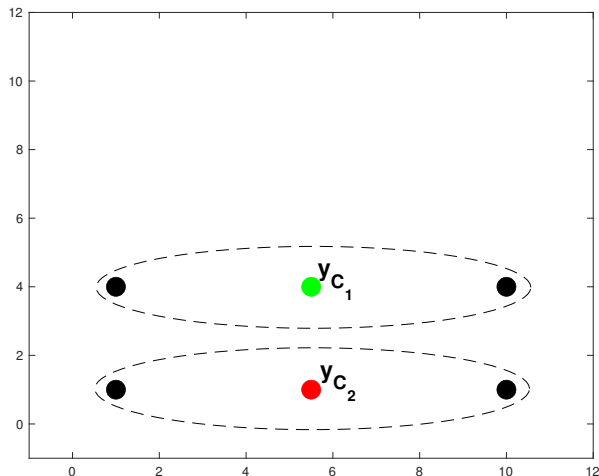


k -means - Príklad: Zlé inicializačné zhľukovanie



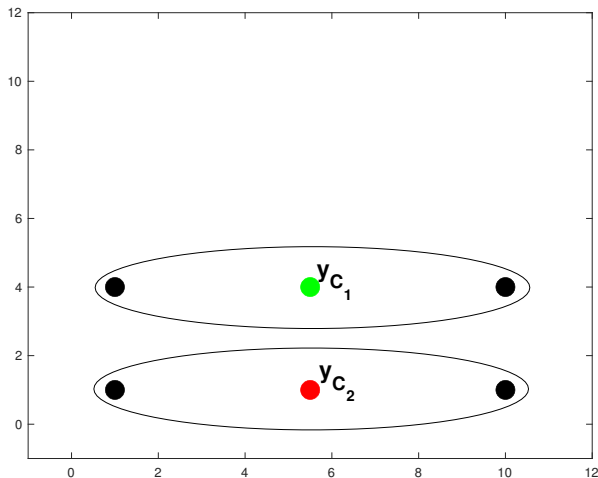
- 4 pozorovania tvoriace 2 “zvislé” zhľuky

k-means - Príklad: Zlé inicializačné zhľukovanie



- zvolíme inicializačné zhľukovanie - “vodorovné” (nesprávne) zhľuky
- vypočítame centroidy

k -means - Príklad: Zlé inicializačné zhľukovanie



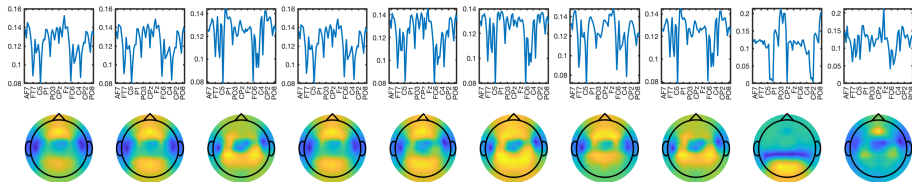
- zhľukovanie je stále rovnaké
- dôležité voliť rôzne inicializačné zhľukovania

k -means - Príklad: topografické mapy mozgu

- $n = 53$ topomáp - lokalizácia zdrojov oscilačnej aktivity v mozgu
- $p = 64$ elektród - vektor reprezentujúci 1 mapu

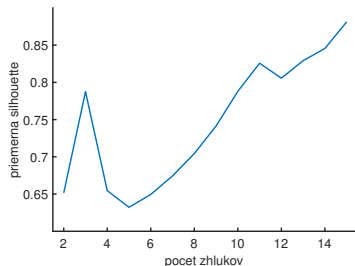
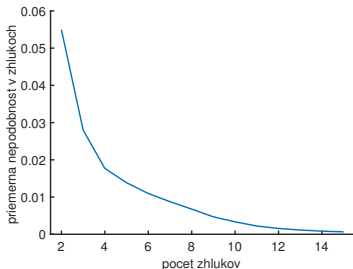
$$\rightarrow X \in \mathbb{R}^{53 \times 64}$$

- cieľ: nájsť K typických topomáp



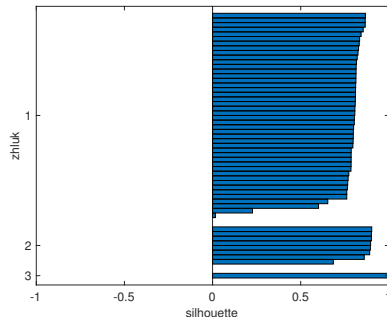
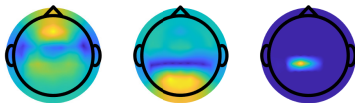
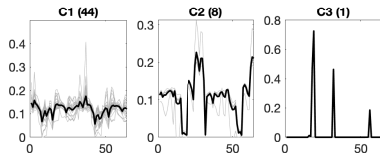
k-means - Príklad: topografické mapy mozgu

- určenie vhodného počtu zhlukov



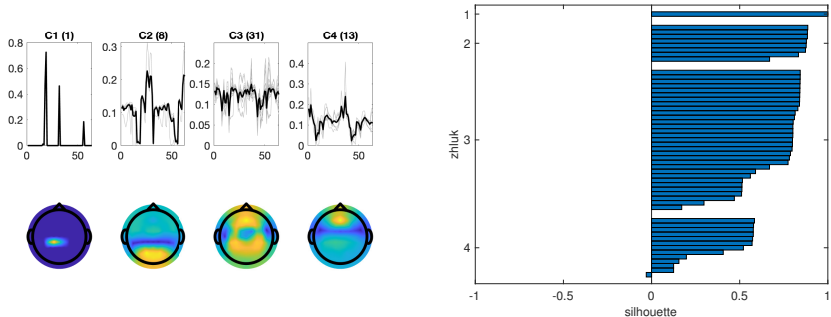
- rule of thumb $K \approx \sqrt{\frac{53}{2}} \approx 5$
- priemerná nepodobnosť v zhlukoch - slabé "zlomy" pre $K \in \{3, 4\}$
- silhouette - $K \in \{3, 11, 15\}$

k-means - Príklad: topografické mapy mozgu



- $K = 3$
- C1 - aktivita vo frontálnej oblasti, najväčší zhluk
- C2 - aktivita v posteriórnej oblasti
- C3 - outlierske pozorovanie, bez interpretácie

k -means - Príklad: topografické mapy mozgu



- $K = 4$
- C1 - outlierske pozorovanie, bez interpretácie
- C2 - aktivita v posteriórnej oblasti
- C3 - aktivita v centrálnej oblasti, najväčší zhluk
- C4 - aktivita vo frontálnej oblasti
- zvyšovanie K viedlo len k deleniu zhľuku C3

- **výhody**

- jednoduchý algoritmus (na pochopenie, implementovanie)
- pomerne rýchla konvergencia pre euklidovskú vzdialenosť

- **nevýhody**

- odlišné inicializačné centroidy → iné finálne zhľukovanie
- “zlé” inicializačné centroidy → “nesprávne” zhľukovanie
→ viac spustení algoritmu
- závislosť na jednotkách merania
→ hodnoty premenných rádovo odlišné → štandardizácia
- tvorí len sférické zhľuky, nevie nájsť nekonvexné zhľuky
- pozorovania musia byť reálne vektory
- pomalšia konvergencia pre iné typy vzdialeností

k -medoids v MATLAB-e

[IDX, C, sumD, D, midX, info] = kmedoids(X, K, Name, Value)

X - $n \times p$ matica dát

K - počet zhlukov

• voliteľné parametre

- Algorithm - pam, small, clara, large
- Distance - sqeuclidean, cityblock, cosine, correlation, hamming
- Replicates - počet opakovaní algoritmu, iné inicializačné medoidy
- ...

IDX $\in \mathbb{R}^n$ = indexy zhlukov pre n prvkov

C $\in \mathbb{R}^{K \times p}$ = medoidy zhlukov

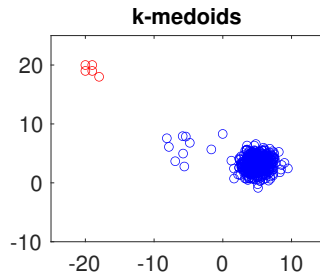
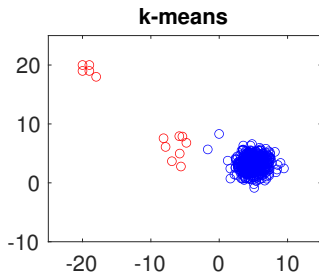
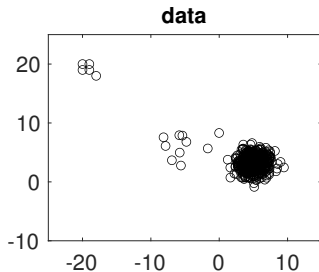
sumD $\in \mathbb{R}^K$ = súčet vzdialeností medzi prvkami zhuku a medoidom

D $\in \mathbb{R}^{n \times K}$ = vzdialenosť každého bodu od každého medoidu

midX - indexy pozorovaní, ktoré sú medoidmi; $y_C = X(\text{midX}, :)$

info - informácie o algoritme, kovergencií, ...

k -medoids a k -means - Porovnanie



- **výhody**

- jednoduchý algoritmus (na pochopenie, implementovanie)
- pomerne rýchla konvergencia
- **menej citlivý na prítomnosť outlierov** ako k -means

- **nevýhody**

- odlišné inicializačné medoidy $\rightarrow$ iné finálne zhukovanie
- závislosť na jednotkách merania
 $\rightarrow$ hodnoty premenných rádovo odlišné $\rightarrow$ štandardizácia
- tvorí len sférické zhluky, nevie nájsť nekonvexné zhluky

Otázky?