

Unsupervised learning

“učenie bez učiteľa”

II. analýza zhlukov

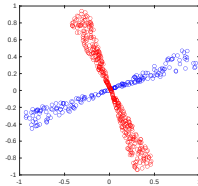
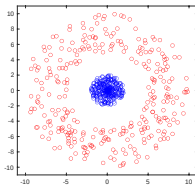
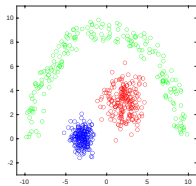
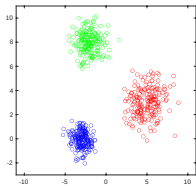
2. časť

Zuzana Rošťáková

21. október 2021

Seminár UM

Analýza zhlukov



- 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 a divízívne**

Nehierarchické metódy zhlukovej analýzy

2. Model-based clustering

Model-based clustering

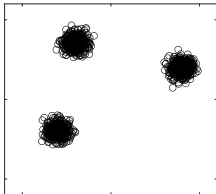
- **predpoklad:** dáta z i -teho zhluku C_i pochádzajú zo známeho rozdelenia

→ najčastejšie: dáta z i -teho zhluku C_i pochádzajú z $\mathcal{N}_p(\mu_i, \Sigma_i)$

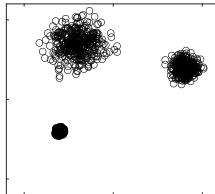
- **cieľ:** zaradenie dát do K zhlukov + odhad charakteristík zhlukov μ, Σ
- predpoklady na kovariančné matice zhlukov:
 - $\Sigma_1 = \Sigma_2 = \dots = \Sigma_K \rightarrow$ zhluky sa líšia len v strednej hodnote
 - $\exists i, j : \Sigma_i \neq \Sigma_j \rightarrow$ niektoré zhluky sa líšia aj v kovariančnej matici
 - $\Sigma_1, \dots, \Sigma_K$ sú diagonálne
 - $\Sigma_1, \dots, \Sigma_K$ môžu byť ľubovoľné $\rightarrow$ tvorba aj eliptických zhlukov

Model-based clustering

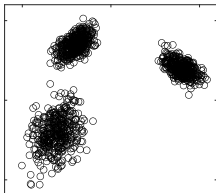
$\Sigma_1 = \Sigma_2 = \Sigma_3$
diagonalna struktura



$\Sigma_1 \neq \Sigma_2 \neq \Sigma_3$
diagonalna struktura



$\Sigma_1 \neq \Sigma_2 \neq \Sigma_3$
lubovolna struktura



Model-based clustering

- **algorithmus:** tzv. EM algoritmus (expectation-maximisation)
 - iteračný algoritmus
 - založený na maximalizácii (komplikovanej) vierohodnostnej funkcie
- **výstup:** tzv. soft clustering
 - matica pravdepodobností $B \in \mathbb{R}^{n \times K}$
 B_{ij} = pravdepodobnosť zaradenia prvku x_i do zhluku C_j
- **hard clustering** - prvok x_i patrí do zhluku C_j , ak $B_{ij} > B_{il}, \quad \forall l \neq j$

Model-based clustering v MATLAB-e

`GM = fitgmdist(X,K, Name, Value)`

`clX = cluster(GM,X)`

`X` - $n \times p$ matica, dáta

`K` - počet zhlukov

- voliteľné parametre

- Start - inicializácia $\mu_i, \Sigma_i, \dots$ (plus, randSample, user-defined)
- Replicates - počet opakovaní algoritmu s inými vstupnými hodnotami
- CovarianceType - diagonal, full
- SharedCovariance - true ($\Sigma_1 = \Sigma_2 = \dots = \Sigma_K$), false
- RegularisationValue - malá konštanta pripočítaná k diagonále Σ na zabezpečenie kladnej definitnosti
- ...

`GM` - štruktúra, napr. `GM.mu` je odhad $\mu_1, \dots, \mu_K, \dots$; pre viac detailov `fields(GM)`

`clX` - zaradenie do zhlukov na základe odhadnutého modelu z `GM`

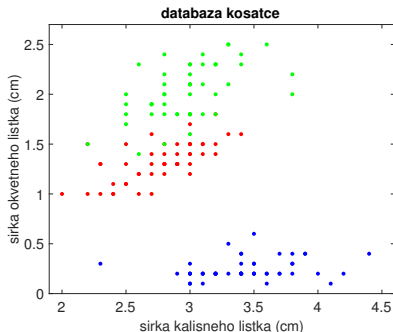
Model-based clustering - Príklad: Kosatce

- databáza *FisherIris* z MATLAB-u

<https://www.mathworks.com/help/stats/clustering-using-gaussian-mixture-models.html>

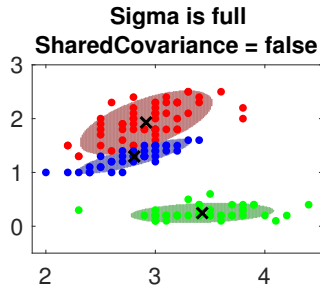
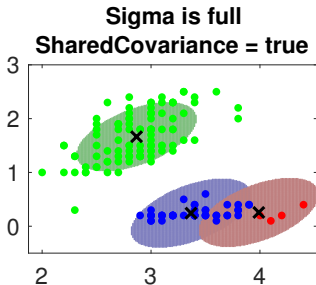
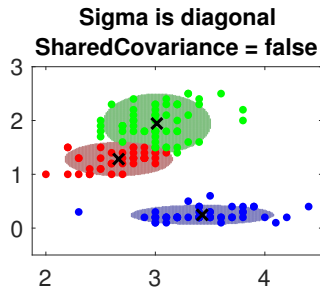
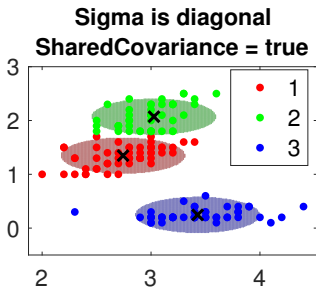
- šírka kališného a okvetného lístka pre 150 kosatcov (3 druhy)

$$X \in \mathbb{R}^{150 \times 2}$$



<https://medium.com/@Nivitus./iris-flower-classification-machine-learning-d4e337140fa4>

Model-based clustering - Príklad: Kosatce



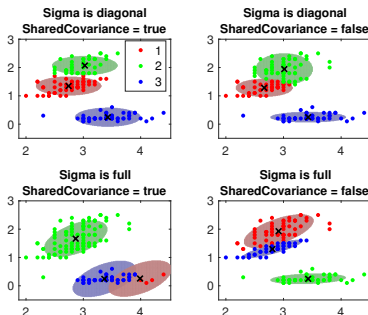
Model-based clustering - parametre, K

- Ako zvoliť vhodnú štruktúru kovariančných matíc?
- Ako stanoviť vhodný počet zhlukov?
- súčasťou GM sú aj Akaikeho a Bayesovo informačné kritérium

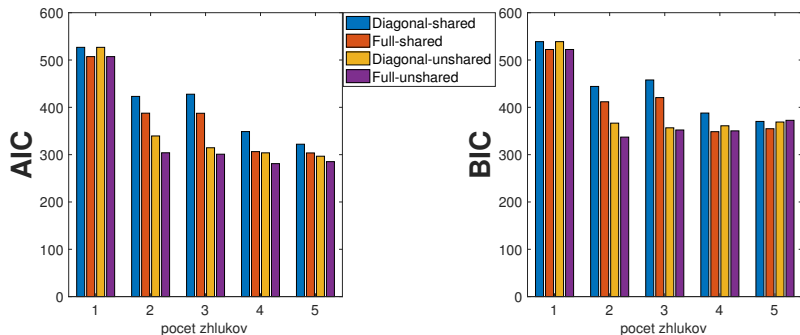
GM.AIC

GM.BIC

- modely s rôznym K , rôznymi predpokladmi na $\Sigma_1, \dots, \Sigma_K$
→ min. AIC a/alebo BIC



Model-based clustering - parametre, K



- kovariančné matice - bez reštrikcií (full), rozdielne medzi zhukmi (unshared)
- počet zhukov
 - AIC - 5, 4, resp. 3 zhuky → AIC zvykne K nadhodnocovať
 - BIC - 2, resp. 3 zhuky → BIC zvykne K podhodnocovať

Model-based clustering - zhrnutie

- **výhody**

- tvorba aj eliptických zhlukov, nie len sférických
- nie je ovplyvnené škálovaním premenných
- vie nájsť aj “skryté” zhluky v rámci “väčších” zhlukov

- **nevýhody**

- ťažšie na pochopenie aj implementáciu

Nehierarchické metódy zhlukovej analýzy

3. Density-based clustering

Density-based clustering

- DBSCAN [Ester et al., 1996]
- neparametrický prístup
- určí automaticky vhodný počet zhlukov K
- parametre
 - minN - minimálny počet susedov
 - ϵ -okolie bodu
 - príliš veľké $\rightarrow$ 1 zhluk
 - príliš malé $\rightarrow$ veľa outlierov

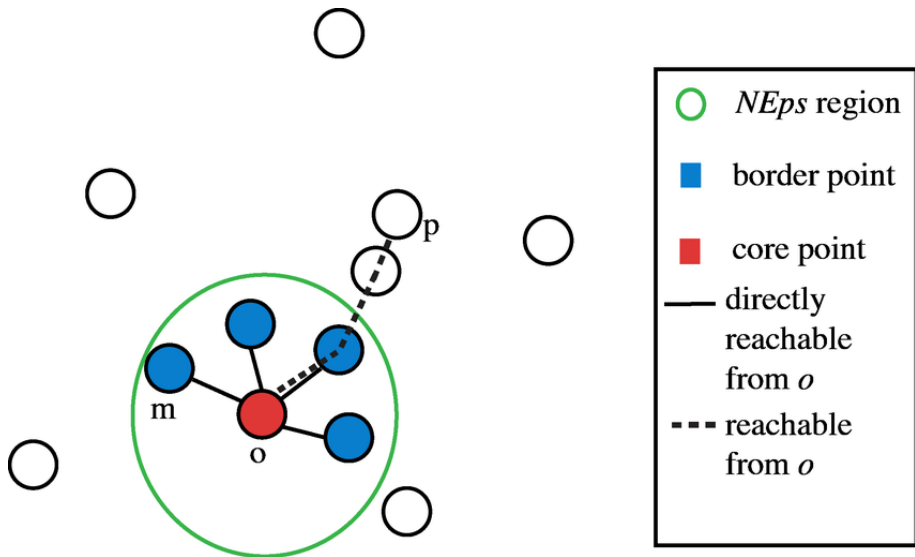
Density-based clustering

- **jadrový bod** (*core point*) - vo svojom ϵ -okolí má aspoň $minN$ susedov
- x_j je **priamo dosiahnuteľný** (directly reachable) z x_i , ak

$$D(x_i, x_j) < \epsilon \text{ a } x_i \text{ je core point}$$

- x_j je **dosiahnuteľný** (reachable) z x_i , ak
 - $\exists$ postupnosť $x_i = x_{l_1}, x_{l_2}, \dots, x_{l_{p-1}}, x_{l_p} = x_j$
 - $x_{l_1}, \dots, x_{l_{p-1}}$ sú *jadrové body*
 - $x_{l_{k+1}}$ je priamo dosiahnuteľný z x_{l_k}
 - všetky body, ktoré sú navzájom dosiahnuteľné, sú v jednom zhluku
- **outlier** - bod, ktorý nie je dosiahnuteľný zo žiadneho iného bodu

Density-based clustering



[Naqshbandi et al., 2016]

DBSCAN v MATLAB-e

`[idx, CorePts] = dbscan(X, eps, minN, Name, Value)`

`X` - $n \times p$ matica, dáta

`eps` > 0 - polomer eps-okolia

`minN` - minimálny počet susedov

- voliteľné parametre

- Distance

- euclidean

- minkowski $\rightarrow$ 'P' - exponent v Minkowskeho vzdialenosti

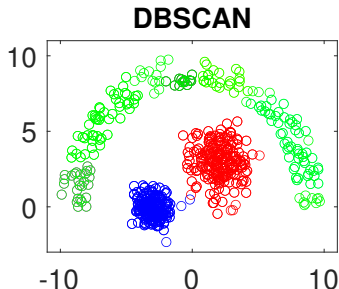
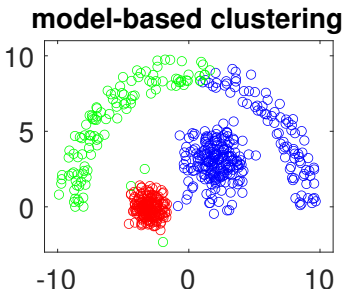
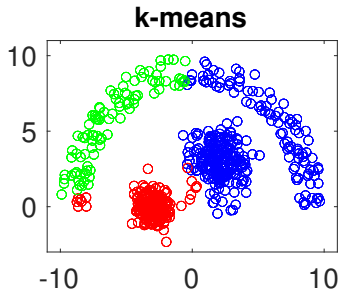
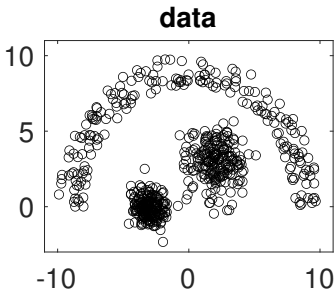
- mahalanobis $\rightarrow$ 'Cov' - kovariančná matica

- ...

`idx` - indexy zhhlukov, outlieri sú označení -1

`CorePts` - indexy *core points*

DBSCAN - Příklad: “banánový” zhluk

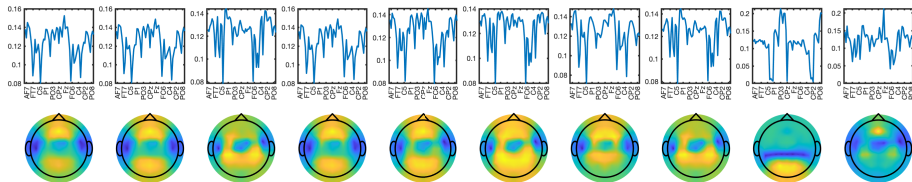


DBSCAN - Príklad: topografické mapy mozgu

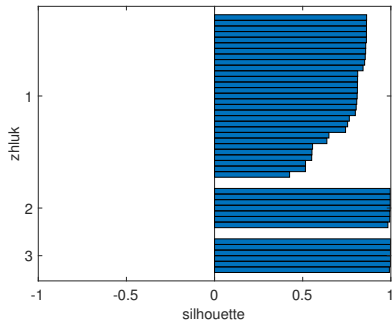
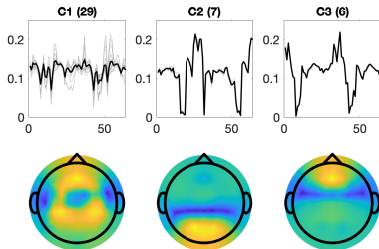
- $n = 53$ topomáp - lokalizácia zdrojov oscilačnej aktivity v mozgu
- $p = 64$ elektród - vektor reprezentujúci 1 mapu

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

- cieľ: nájsť K typických topomáp

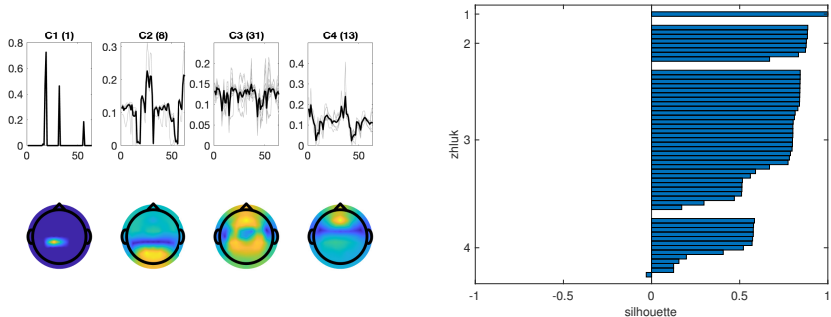


DBSCAN - Príklad: topografické mapy mozgu



- $\min N = 5$
- $\epsilon = 0.15$
- nájdené tri zhluky $\rightarrow$ podobne ako k -means
- 11 outlierov z 53 pozorovaní

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

Density-based clustering - zhrnutie

- **výhody**

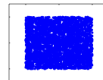
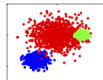
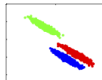
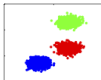
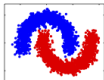
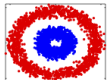
- nie je potrebné stanoviť počet zhlukov $K \rightarrow$ automatická selekcia
- zhluky ľubovoľného tvaru - nie len sférické a eliptické, ale aj “banánové” (nekonvexné)
- robustný na prítomnosť outlierov, ich automatická detekcia

- **nevýhody**

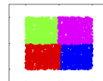
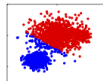
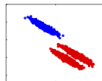
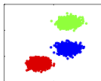
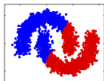
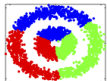
- výsledok závisí od zvolenej miery vzdialenosti
- potrebné dobre poznať dáta, aby sme vedeli stanoviť vhodné $minN$ a ϵ

Porovnanie nehierarchických metód

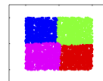
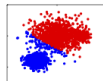
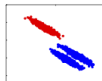
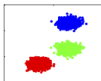
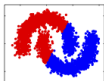
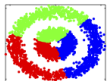
spravne zaradenie do zhlukov



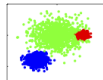
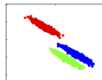
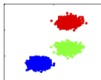
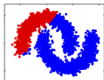
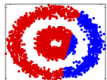
k-means



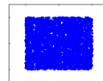
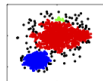
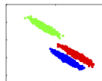
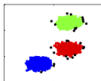
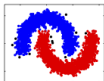
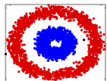
k-medoids



GMM



DBSCAN



Porovnanie nehierarchických metód

- viac porovnaní (aj ďalších) metód zhlukovej analýzy

<https://scikit-learn.org/stable/modules/clustering.html>

- ***k*-means, *k*-medoids**

- metódy “prvej voľby”
- ok pre “klasické” problémy s predpokladom sférických zhlukov
- môžu nám dať predstavu o charaktere dát

- **GMM**

- vie nájsť aj “menšie zhluky vnorené do väčších zhlukov”

- **DBSCAN**

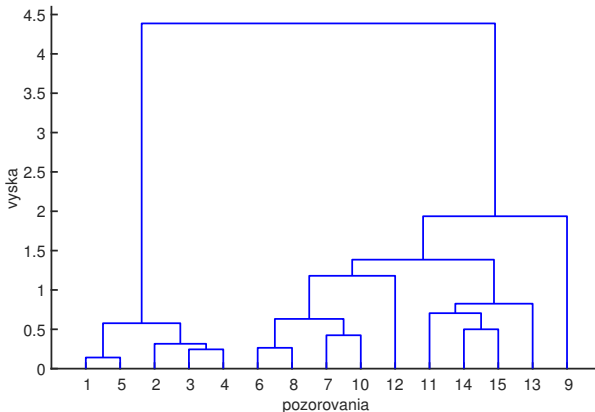
- vhodný pre dáta, v ktorých očakávame špecifické tvary zhlukov alebo keď euklidovská vzdialenosť nie je najvhodnejší typ vzdialenosti

Hierarchické metódy zhlukovej analýzy

Hierarchické metódy zhlukovej analýzy

- výstupom nie je priamo zaradenie do zhlukov, ale tzv. strom závislostí (dendrogram) → vytvorí sa “hierarchia” pozorovaní

1. aglomeratívne zhlukovanie
2. divizívne zhlukovanie



Agglomeratívne zhlukovanie

- tzv. bottom-up

1. každé pozorovanie je vo vlastnom zhluku
 2. dve najpodobnejšie pozorovania sú spojené do jedného zhluku
 - ...
 - k. spojenie dvoch najpodobnejších pozorovaní/zhlukov do jedného zhluku
 - ...
 - L. všetky pozorovania sú v jednom zhluku
- ⇒ okrem vzdialenosti medzi pozorovaniami potrebujeme definovať aj “vzdialenosť medzi zhlukmi”

Aglom. zhukovanie - vzdialenosť medzi zhukmi

- single linkage - minimálna vzdialenosť medzi prvkami zhukov

$$D_{sl}^*(C_i, C_j) = \min_{x \in C_i, y \in C_j} D(x, y)$$

- average linkage - priemerná vzdialenosť medzi prvkami zhukov

$$D_{al}^*(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i} \sum_{y \in C_j} D(x, y)$$

- complete linkage - maximálna vzdialenosť medzi prvkami zhukov

$$D_{cl}^*(C_i, C_j) = \max_{x \in C_i, y \in C_j} D(x, y)$$

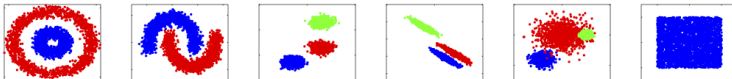
- Wardova vzdialenosť

$$D_{Wd}^*(C_i, C_j) = \sum_{m \in C_i \cup C_j} D(m, c_{ij}) - \left[\sum_{x \in C_i} D(x, c_i) + \sum_{y \in C_j} D(y, c_j) \right]$$

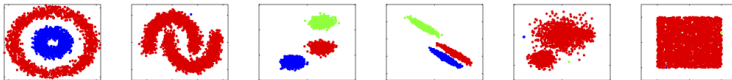
- c_{ij} je centroid $C_i \cup C_j$; c_i je centroid C_i ; c_j je centroid C_j

Vzdialenosť medzi zhlukmi - porovnanie

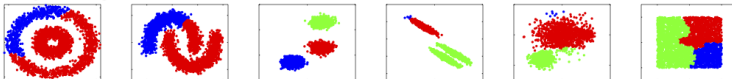
spravne zaradenie do zhlukov



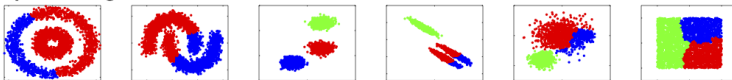
single linkage



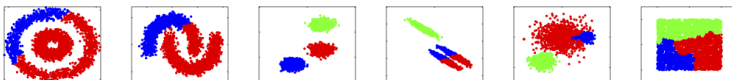
average linkage



complete linkage



Ward distance



Vzdialenosť medzi zhlukmi - porovnanie

- **single linkage**

- nie je robustná v prípade zašumených dát
- efektívna pre veľký počet pozorovaní
- veľké rozdiely vo veľkostiach zhlukov

- **Ward distance**

- zhluky majú podobnú veľkosť
- pomalšie na výpočet

- **average linkage**

- “zlatá stredná cesta” pre neeuklidovské vzdialenosti

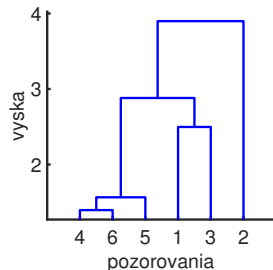
Agglomeratívne zhľukovanie v MATLAB-e

$Z = \text{linkage}(X, \text{'method'}, \text{'metric'})$

X - $n \times p$ matica pozorovaní

- method = single, average, complete, Ward, ...
- metric = vzdialenosť medzi prvkami (euclidean, minkowski, mahalanobis, ...)

Z - matica s 3 stĺpcami, príklad pre $n = 6$



x	y	$D(x, y)$	
4	6	1.40	→ x_4 a x_6 spojené do $C_1 = x_7$
5	7	1.57	→ x_5 a $C_1 = x_7$ spojené do $C_2 = x_8$
1	3	2.50	→ x_1 a x_3 spojené do $C_3 = x_9$
8	9	2.88	→ $C_2 = x_8$ a $C_3 = x_9$ spojené do $C_4 = x_{10}$
2	10	3.90	→ x_2 a $C_4 = x_{10}$ spojené do $C_5 = x_{11}$

Aglomeratívne zhľukovanie v MATLAB-e

`[H,T,O] = dendrogram(Z, P, Name, Value)`

Z - výstup z funkcie *linkage*

P - maximálny počet vykreslených zhľukov; ak $P = 0$ alebo $P = n \Rightarrow$ kompletný dendrogram

- voliteľné parametre

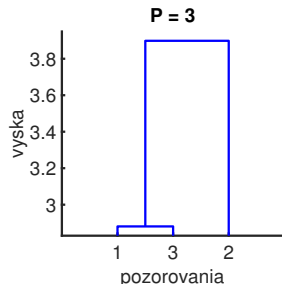
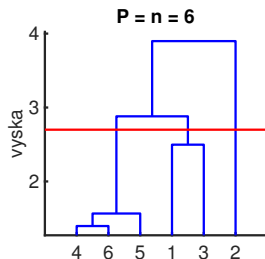
- Orientation - orientácia dendrogramu (top, bottom, left, right)
- Labels - označenia pozorovaní
- Reorder - vektor permutácie zhľukov
- CheckCrossing - kontrola pretínania "vetiev"

- vykreslí aj strom závislostí

H - handles "vetiev" v dendrograme (grafe)

T - indexy prvkov v dendrograme, vhodné ak $P < n$ (príklad: $T = [1, 2, 1, 3, 3, 3]$)

O - označenia zhľukov



`ind = cluster(Z,'maxClust',K)`

- zaradenie do zhukov na základe dendrogramu

Z - výstup z funkcie *linkage*

- maxClust $K \in \mathbb{N}$ - vytvorí sa maximálne K zhukov

Agglomeratívne zhľukovanie - kvalita zhľukovania

$$Y = \text{pdist}(X)$$

X - $n \times p$ matica pozorovaní

Y - $n \times n$ matica vzájomných vzdialeností pozorovaní

$$[C,D] = \text{cophenet}(Z,Y)$$

- korelácia medzi vzdialenosťou prvkov z dendrogramu a ich reálnou vzdialenosťou
- na určenie vhodného typu *linkage*

Z - výstup z $Z = \text{linkage}(X, 'method', 'metric')$

Y - výstup z $Y = \text{pdist}(X)$

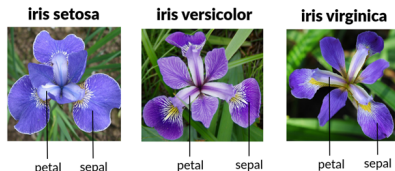
C - kofenetická korelácia, $C \in [-1, 1]$

D - kofenetické vzdialenosti medzi pozorovaniami

Agglomeratívne zhľukovanie - Príklad: Kosatce

- dĺžka a šírka okvetných a kališných lístkov pre 150 kosatcov

$$X \in \mathbb{R}^{150 \times 4}$$

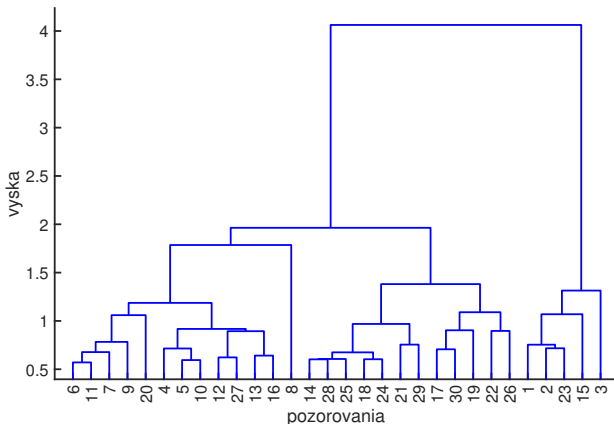


- vzdialenosť = euklidovská $Y = \text{pdist}(X, 'euclidean')$
- určenie vhodnej linkage pomocou *cophenet*
 - $Z1 = \text{linkage}(X, 'single')$
 $[C1, \sim] = \text{cophenet}(Z1, Y) \rightarrow C1 = 0.86$
 - $Z2 = \text{linkage}(X, 'average')$ $\rightarrow C2 = 0.88$
 - $Z3 = \text{linkage}(X, 'complete')$ $\rightarrow C3 = 0.73$
 - $Z4 = \text{linkage}(X, 'ward')$ $\rightarrow C4 = 0.87$

Aglomeratívne zhľukovanie - Príklad: Kosatce

$[H, T] = \text{dendrogram}(Z2)$

- MATLAB vykresľuje max. 30 koncových objektov $\rightarrow$ z T vieme, že
 - objekt "6" = zhľuk, obsahuje pozorovania 54, 60, 70, 81, 82, 90
 - objekt "9" = pozorovanie 63



Aglomeratívne zhlukovanie - Príklad: Kosatce

```
ind1 = cluster(Z2,'maxClust',3)
```

- poznáme zaradenie kosatcov do 3 skupín / druhov
- chceme vidieť, ako naše zhlukovanie korešponduje so skutočnými druhmi

ind1 / druh	setosa	versicolor	virginica
1	0	0	36
2	0	50	14
3	50	0	0

Hierarchické metódy - divizívne zhlukovanie

- tzv. top-down

1. všetky pozorovania sú spolu v jednom zhluku

2. nájdeme “najodlišnejšie” pozorovanie x^*

→ nájdeme pozorovania podobnejšie x^* ako zvyšku pozorovaní

→ rozdelenie na dva zhluky

...

k. rozdelenie zhluku s najväčšou vnútornou nepodobnosťou podľa 2.

...

L. každé pozorovanie je vo vlastnom zhluku

- divizívne zhlukovanie nie je súčasťou MATLAB-u

→ MathWorks, GitHub, ...

- **výhody**

- nepotrebujeme vopred stanoviť počet zhlukov
- grafické znázornenie (dendrogram) pomôže lepšie pochopiť štruktúru v dátach, resp. odhadnúť počet zhlukov pre iné metódy

- **nevýhody**

- závisí na škálovaní dát
- výpočtovo náročné pre veľký počet pozorovaní
- rôzne “linkage” môžu viesť k podstatne odlišným dendrogramom
- pre príliš veľké n môže byť dendrogram neprehľadný

Otázky?



Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996).

A density-based algorithm for discovering clusters in large spatial databases with noise.

In *Second International Conference on Knowledge Discovery & Data Mining*, pages 226–231. American Association for Artificial Intelligence.



Naqshbandi, K., Gedeon, T., and Abdulla, U. (2016).

Automatic clustering of eye gaze data for machine learning.

In *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 001239–001244.