Přednáška 1 – Úvod. Multimodální data. Základní modely

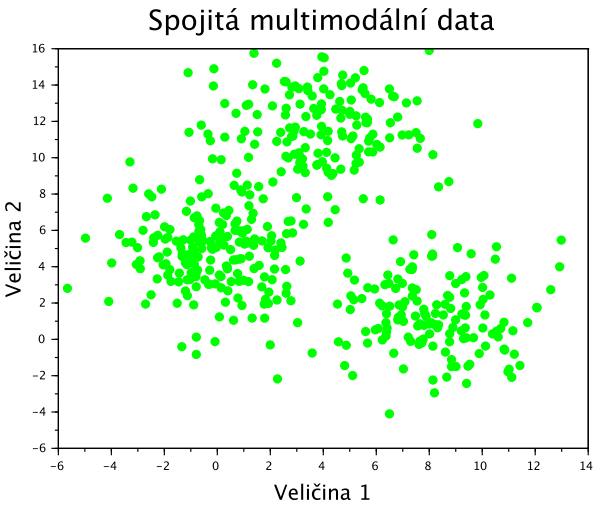
- Evženie Uglickich, http://staff.utia.cas.cz/uglickich
  - skripta, přednášky, materiály ke cvičení
  - doporučená literatura:
    - I. Nagy, E. Suzdaleva. Algorithms and Programs of Dynamic Mixture Estimation. Unified Approach to Different Types of Components, Springer, 2017.
    - D. T. Larose. Discovering Knowledge in Data. An Introduction to Data Mining. Willey, 2005.
    - J. Han, M. Kamber, J. Pei. Data Mining: Concepts and Techniques, 3rd Edition. Morgan Kaufmann, 2011.
    - M. J. Zaki, W. Meira Jr. Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press, 2014.
  - konzultační hodiny - osobně/MS Teams
  - moodle
- Zápočet: body za práci v hodinách
- Zkouška:
  - ústní forma, 5 otázek
  - pouze v zimním zkouškovém období, otázky jsou na webu
  - předtermín, doktorandi – témata v oblasti analýzy dat

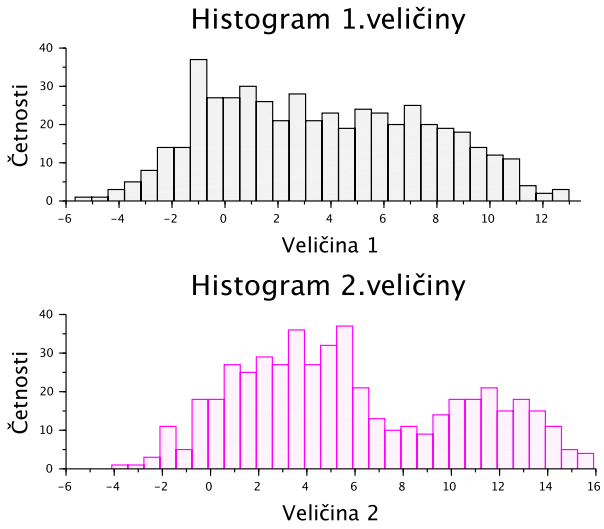
O čem je předmět Matematické metody analýzy dat?

- Analýza multimodálních dat – spojitých a diskrétních
- Shluková analýza a klasifikace
- Metody na základě modelu, s učitelem a bez učitele

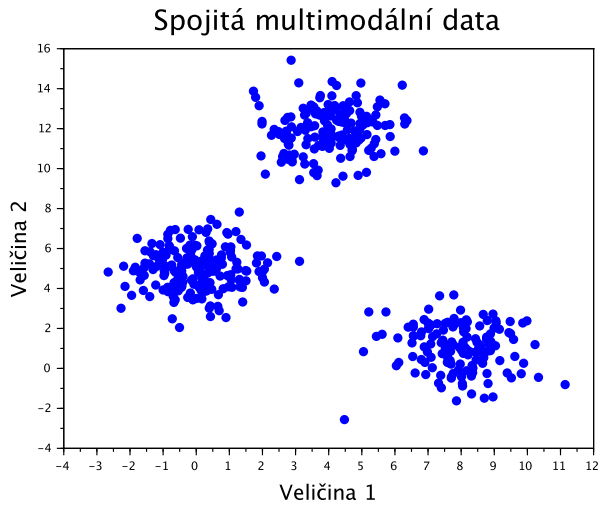
Multimodální data – spojitá

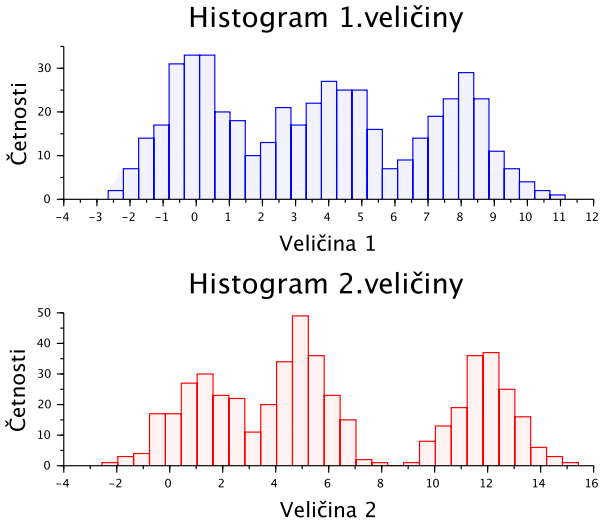
- systém – část reality, kterou modelujeme
- multimodální systém – přepíná mezi jednotlivými režimy
  Příklad: styl jízdy, dopravní špičky, atd
- naměřená data – multimodální
- tvoří shluky (klastry)
- vzdálenost mezi shluky – rozptyl





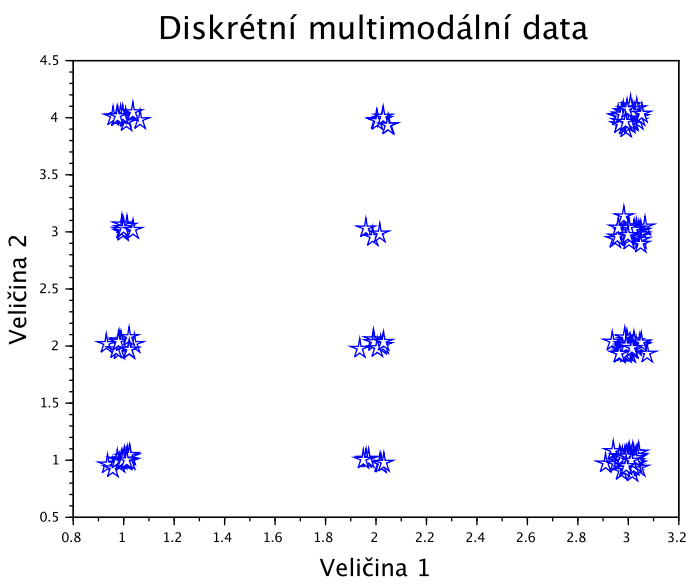
- spojitá data

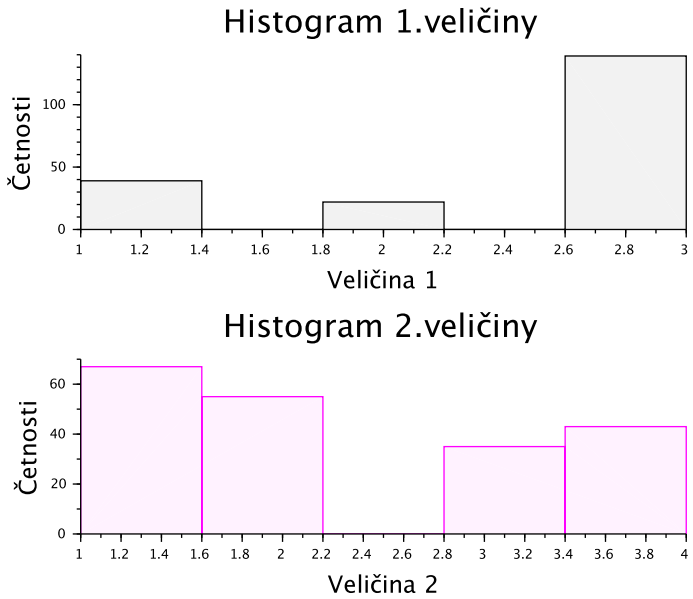




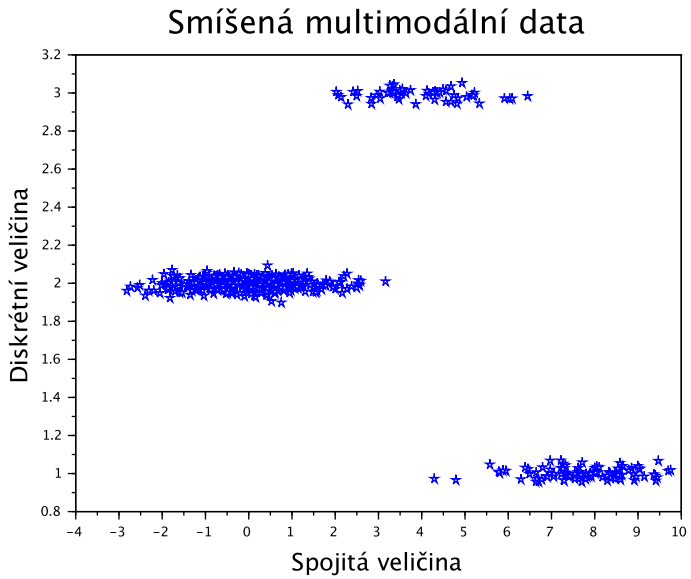
Multimodální data – diskrétní data, smíšená data

- diskrétní data



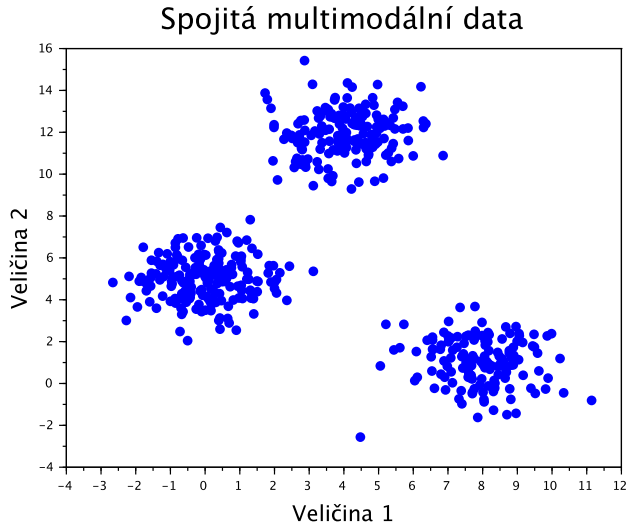


- smíšená data

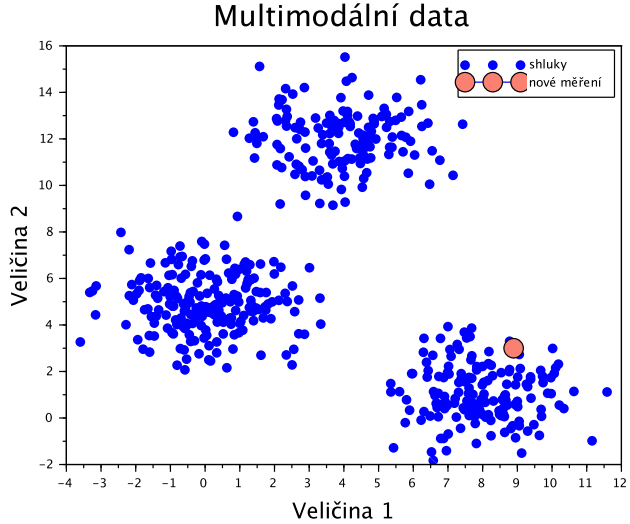


Úlohy shlukování vs. klasifikace

- Shlukování (bez učitele)
  - nalezení shluků dat s podobnými vlastnostmi (třídění dat do skupin)
- Algoritmy na základě:
  - modelu (model směsi distribucí)
  - centroidů (k-means, k-medoids)
  - fuzzy logiky (c-means)
  - hustoty (DBSCAN)
  - hierarchie shluků



- Klasifikace (s učitelem)
  - třídění dat do předem známých skupin
- Algoritmy:
  - model (směsi, log. regrese, naivní Bayes)
  - k-nejbližších sousedů
  - rozhodovací stromy, náhodné lesy
  - podpůrné vektorové stroje
  - neuronové sítě



Základní modely – statický normální model pro spojitá data

$y_t = \theta + e_t$

y_t – modelovaná náhodná veličina (vektor)
θ – vektor regresních koeficientů (neznámé parametry)
e_t – bílý šum s normálním rozdělením

Příklad:

$$\begin{bmatrix} y_{1;t} \\ y_{2;t} \end{bmatrix} = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} + \begin{bmatrix} e_{1;t} \\ e_{2;t} \end{bmatrix}$$

- Statický model – vhodný pro shlukování a klasifikaci

Mnohorozměrný lineární normální model ve tvaru hp – obecně

$$f(y_t|\underbrace{\theta, R}_{\Theta}) = (2\pi)^{-\frac{n}{2}} \det(R)^{-\frac{1}{2}} \exp\{-\frac{1}{2}(y_t - \theta)' R^{-\frac{1}{2}}(y_t - \theta)\}$$

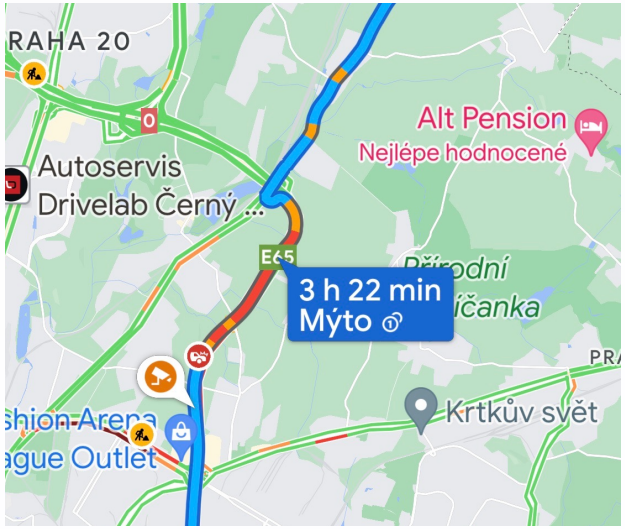
- Střední hodnota, kovarianční matice, Gaussova křivka

Základní modely – kategorický model pro diskrétní data

$y_t \in \{1, 2, 3, \ldots, N\}$ – modelovaná veličina

Příklad:

stupeň dopravy (volný provoz až dopravní kolaps)



y_t	1	2	3	4	5
Θ_i	$\Theta_1 = 0.05$	$\Theta_2 = 0.05$	$\Theta_3 = 0.1$	$\Theta_4 = 0.6$	$\Theta_5 = 0.2$

$\Rightarrow 1$

- Model ve tvaru pravděpodobnostní funkce $f(y_t|\Theta)$
- $\Theta \equiv \{\Theta_i\}_{i=1}^{N}$ (neznámé parametry)
- Statický model, neurčitost v modelu

Základní modely – Poissonův model pro sčítací data

- $y_t \in \{0, \ldots\}$ – modelovaná veličina
- počty náhodných nezávislých událostí za jednotku času

Příklad: počet vozidel, cyklistů, cestujících (za min, h, ...)

- pozor na počet možných realizací

Model ve tvaru hp – Poissonovo rozdělení

$$f(y_t|\underbrace{\lambda}_{\Theta}) = \exp\{-\lambda\}\frac{\lambda^{y_t}}{y_t!}, \qquad \lambda - \text{neznámé parametry}$$

- předpoklad λ = střední hodnota = rozptyl