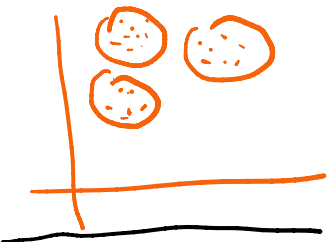


Príklady: zhlukovanie - k -means } nehierarchické
 - k -medoids
 - hierarchické: aglomeratívne



výška spájania: určená vzdialenosťou (linkage)
 zhlukov

⇒ "vzdialenosť": vertikálna, nie horizontálna

↑
 výška bodu, kde boli prvýkrát spojené

• Hlavné zhluky: useknutie vodorovnou čiarou →

3 zhluky: 1, 2, 3
 : 4, 5
 : 6, 7

Výhody:

- hierarchická štruktúra je zaujímavá sama osebe
- nemusíme stanoviť # zhlukov vopred
- ľahká identifikácia outlierov

Nevýhody:

- neprehľadné pre veľké dáta

Všeobecné k zhlukovaniu:

- voľba metriky (vzdialenosti):

(i) standardizácia vopred - môže mať vplyv

(ii) napr. "korelačná vzd."

- združenie podľa # nálekov $x = \begin{pmatrix} \# \in 1. \text{ kategória} \\ \# \in 2. \text{ kat.} \\ \# \in 3. \text{ kat.} \end{pmatrix}$

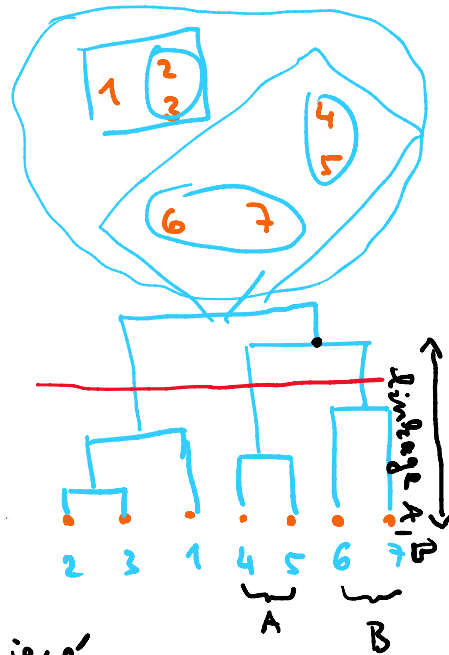
$$x_1 = \begin{pmatrix} 3 \\ 0 \\ 3 \end{pmatrix}, \quad x_2 = \begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix}, \quad x_3 = \begin{pmatrix} 30 \\ 2 \\ 20 \end{pmatrix}$$

- klasická eur. vzd.: 1. a 2. sú podobní

: silný dôraz na celkovú veľkosť nálekov

- "korelačná" vzd.: podľa $\text{cor}(x_i, x_j)$

: silný dôraz na štruktúru



: 1. a 3. sú podobní

- opätne s výstupom - nemusí byť "správny" → ideálne overiť na novom dátade

Princípy pravdepodobnostného modelovania

$$X = \begin{bmatrix} \text{---} x_1^T \text{---} \\ \vdots \\ \text{---} x_n^T \text{---} \end{bmatrix}_{n \times p}$$

$$x_1 = \begin{pmatrix} x_{11} \\ \vdots \\ x_{1p} \end{pmatrix} \in \mathbb{R}^p : \text{realizácia náh. vektora } X$$

premenne (veľ, výška): náhodne premenne $X_1, \dots, X_p$

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} - \text{náh. vektor}$$

→ $\begin{pmatrix} \text{náh. výška} \\ \text{náh. veľ} \\ \vdots \end{pmatrix}$

1. stĺpec: n realizácií x_1 (n vektor)

• pp. modely: typický: také, kde práca s pravdepodobnostným rozdelením (vektora X) je hlavným nástrojom modelu

$X \rightarrow$ spojité: hustota $f(x) \rightarrow$ rozdelenie váh, výšok, ... v populácii
 → diskrétny: rozdelenie pp.

Pravdepodobnosť

(i) diskrétna náh. prem.: spočítateľne veľa hodnôt

$$\rightarrow X = \begin{cases} 1, & \text{pp. } p & \dots & \text{firma zbankrotuje} \\ 0, & \text{pp. } 1-p & \dots & \text{nezbankrotuje} \end{cases}$$

$$X = \# \text{ ízieb v náh. dome} = \begin{cases} 1 & \dots & P(X=1) = \square \\ 2 & \dots & P(X=2) = \square \\ 3 & \dots & \\ \vdots & & \\ 50 & \dots & \\ \vdots & & \\ (\infty) & & \end{cases}$$

(ii) spojité: nespočítateľne veľa hodnôt

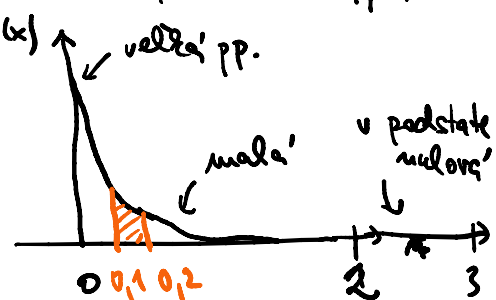
$X =$ dĺžka čakania na autobus

$X \in [0, 24]$ - ak by sme každému číslu $x \in [0, 24]$ priradili pp $\neq 0$

→ sečítali by sa na ∞ $f(x) \propto$...

$X \in [0, 24]$ - ak by sme každému číslu z $[0, 24]$ priradili pp $\neq 0$

$\rightarrow$ sečítali by sa na ∞ $f(x)$
 $\Rightarrow$ hustota $f(x)$



$$P(\text{číslo medzi } 0,1 \text{ a } 0,2) = \int_{0,1}^{0,2} f(x) dx$$

vo všeob.: $\int_a^b f(x) dx = P(X \in (a, b))$

• hustota čísel = ?

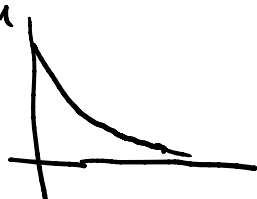
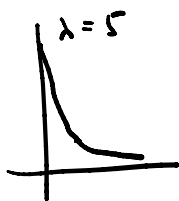
- aby sme nemuseli modelovať celú hustotu (každé pp.)

$\rightarrow$ vytvorili sme parametrizované rozdelenia pp.

$\Rightarrow$ stačí odhadnúť pár parametrov

$\rightarrow$ čísel : $X \sim \text{Exp}(\lambda) \dots f(x) = \lambda e^{-\lambda x}, x \geq 0$

$\rightarrow$ stačí mi odhadnúť λ ($E(X) = \frac{1}{\lambda}$)



voľba rozdelenia : zodpovedá procesu, ako vznikli (sedí rozprávača)

: sedí na dáta $\rightarrow$ správne hodnoty (1, 2, 3, ... $\rightarrow$ diskrétne)

$\rightarrow$ správne charakteristiky (rozsah, šírka, výška, (ne)konечná disperzia)...

$\rightarrow$ dobre fituje - súčet štvorcov chýb } lepšie
 - likelihood }
 - informačné } lepšie
 - kritéria

Rozdelenia

Alternatívne : $y \sim \text{Alt}(\tau) : y = \begin{cases} 0 & \text{pp. } 1-p \\ 1 & \text{pp. } p \end{cases}$

$\Rightarrow$ pre 0/1 (binárne)

Binomické : $y \sim \text{Bin}(N, p) \rightarrow y \in \{0, 1, \dots, N\}$
 $\uparrow \quad \uparrow$
 ... 0 ... 1 ...

Binomial: $y \sim \text{Bin}(N, p) \rightarrow y \in \{0, 1, \dots, N\}$
 $\uparrow$ # pokusov $\uparrow$ pp. úspechu $\rightarrow P(y=k) = \binom{N}{k} p^k (1-p)^{N-k}$ $\uparrow$ # k

rozpráva: # úspechov, keď N -krát nezávisle opakujeme pokus

$\rightarrow$ napr.: # šiestok pri 100 hodoch: $\sim \text{Bin}(100, \frac{1}{6})$

: # vyplnených éiaroniek za rok: $\sim \text{Bin}(N, p)$

éiaroniek $\rightarrow$ $\uparrow$ pp. vyplnenia

: # študentov, kt. dostanú A z predmetu

$\sim \text{Bin}(N, p)$

študentov $\uparrow$ $\uparrow$ pp. A-čky (z dlhodobých záznamov)

aby sedelo na dáta: - celčíselné dáta, vieme # pokusov