

Nehierarchické shlukování – přednášky USK kapitola 4.3

Jde o metody, které se snaží provést optimální rozklad trénovací množiny T do R shluků. Je obvykle definována kritériální funkce a rozkladem by měl být zajištěno dosažení tohoto kritéria. Kritériální funkce se většinou definuje:

$$J = \sum_{i=1}^R J_i = \sum_{i=1}^R \sum_{x \in T_i} d^2(x, \mu_i)$$

hledám minimum tohoto kritéria

kde μ je střední hodnota shluku T_i

Pozn.:

Na první pohled se zdá, že je možné využít vyčerpávající prohledávání trénovací množiny s cílem určit optimální rozklad, který zajistí globální minimum.

V praxi je tato metoda neproveditelná, a to i pro jednoduché úlohy.

Např. pro m prvků a c tříd existuje přesně:

$$\frac{1}{c!} \sum_{i=1}^c \binom{c}{i} (-1)^{c-i} i^m \approx \frac{c^m}{c!} \quad \text{pro } m \gg c$$

Příklad: $c = 5, m = 100: \frac{5^{100}}{5!} = 10^{67}$

K-průměrová metoda (K-means algoritmus, MacQueenův algoritmus 1967)

Metoda je založena na minimalizaci kritériální funkce

$$J = \sum_{i=1}^R \sum_{x \in T_i} d^2(x, \mu_i)$$

Předpokládejme, že je dána trénovací množina $T = \langle x_1, \dots, x_N \rangle$

Definujeme: $J_i(k)$... příspěvek kritériální funkce i -tého shluku v k -tém kroku

$T_i(k)$... i -tý shluk v k -tém kroku

$\mu_i(k)$... střední hodnota i -tého shluku v k -tém kroku

$s_i(k)$... počet obrazů v i -tém shluku v k -tém kroku

Postup:

1. Zvolte R počátečních středů shluků $\mu_1(1), \dots, \mu_R(1)$. Středů shluků lze vybrat libovolně, obvykle jsou však vybrány jako prvních R obrazů daného souboru obrazů.
2. V k -tém iterativním kroku se rozdělují obrazy trénovací množiny T do R shluků $T_1(k), \dots, T_R(k)$ podle vztahu: $x \in T_j(k)$, jestliže $d(x, \mu_j(k)) < d(x, \mu_i(k))$ pro všechny $i, j = 1 \dots R, i \neq j$. Tento vztah je postupně aplikován na všechny obrazy x trénovací množiny T .
3. Z výsledků kroku 2 vypočti pro každý shluk nový střed, tj. $\mu_j(k+1)$, $j = 1, \dots, R$ tak, aby sumě kvadrátů vzdáleností všech obrazů v $T_j(k)$ do nového středu shluku byla minimální. Střed shluku $\mu_j(k+1)$, který minimalizuje kritérium

$$J_j(k+1) = \sum_{x \in T_j(k)} d^2(x, \mu_j(k+1)) \quad j = 1, \dots, R$$

lze určit ze vztahu

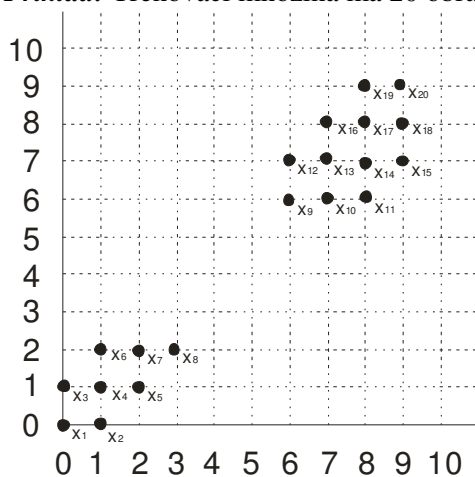
$$\mu_j(k+1) = \frac{1}{s_j(k)} \sum_{x \in T_j(k)} x \quad j = 1, \dots, R$$

4. Jestliže $\mu_j(k+1) = \mu_j(k)$ pro všechny $j = 1, \dots, R$ algoritmus dokonvergoval a procedura je ukončena. Jinak jdi na krok 2.

Pozn.: Proceduru lze alternativně ukončit i v případě, že pokles hodnoty kritériální funkce je již nevýznamný.

Pozn.: k-means zajišťuje pouze lokální minimum. Záleží např. na počátečním rozkladu, tj. jaké μ vyberu nejprve.

Příklad: Trénovací množina má 20 obrazů, rozdělte ji na dva shluky



Vstupní data

Krok 1: Volíme středy

$$\mu_1(1) = x_1 = [0,0]^T$$

$$\mu_2(1) = x_2 = [1,0]^T$$

Krok 2: S novými středy získáme rozklad na T1, T2

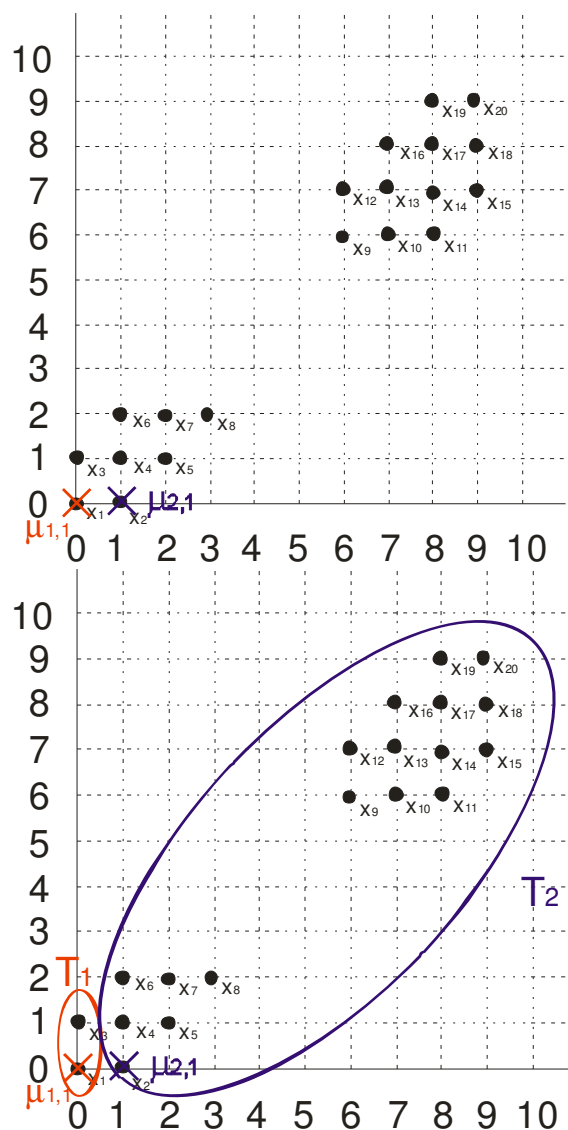
Zařad' do T1 : $d(\mu_1(1), x_i) < d(\mu_2(1), x_i)$

Zařad' do T2 : $d(\mu_1(1), x_i) > d(\mu_2(1), x_i)$

$$T_1(1) = \{x_1, x_3\}; s_1(1) = 2$$

Ostatní obrazy mají menší vzdálenost k $\mu_2(1) \Rightarrow$

$$T_2(1) = \{x_2, x_4, x_5, \dots, x_{20}\}; s_2(1) = 18$$



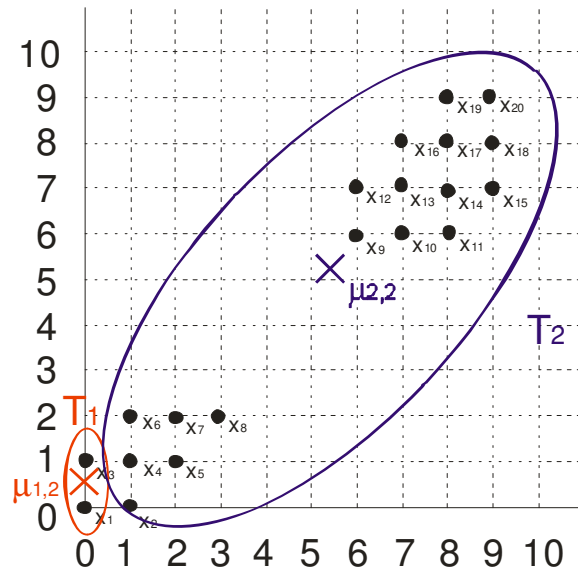
Krok 3: Aktualizace středů a výpočet kritéria

$$\mu_1(2) = \frac{1}{s_1(1)} \sum_{x \in T_1(1)} x = \frac{1}{2} [x_1 + x_3] = [0; 0,5]^T$$

$$\mu_2(2) = \frac{1}{s_2(1)} \sum_{x \in T_2(1)} x = \frac{1}{18} [...] = [5,67; 5,33]^T$$

$$J_1(2) = \sum_{x \in T_j(k)} d^2(x, \mu_j(2)) = 1$$

$$J_2(2) = \sum_{x \in T_j(k)} d^2(x, \mu_j(2)) = 69.2$$



Krok 4: Došlo-li ke změně středů pak přejdi na krok 2

$$\mu_1(2) \neq \mu_1(1) \text{ a } \mu_2(2) \neq \mu_2(1)$$

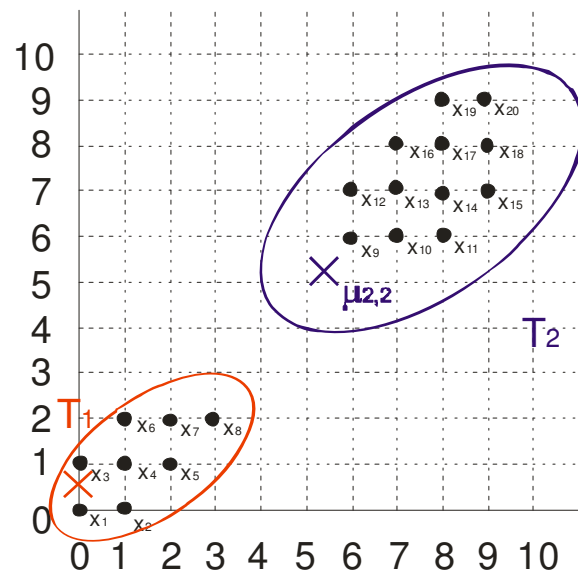
Krok 2:

$$d(\mu_1(2), x_i) < d(\mu_2(2), x_i)$$

$$d(\mu_2(2), x_i) < d(\mu_1(2), x_i)$$

$$T_1(2) = \{x_1, x_2, \dots, x_8\}$$

$$T_2(2) = \{x_9, x_{10}, \dots, x_{20}\}$$



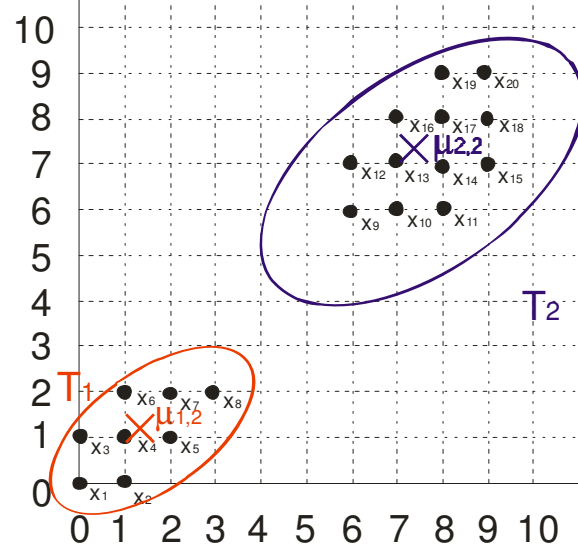
Krok 3:

$$\mu_1(3) = \frac{1}{s_1(2)} \sum_{x \in T_1(2)} x = [1,25; 1,13]^T$$

$$\mu_2(3) = \frac{1}{s_2(2)} \sum_{x \in T_2(2)} x = [7,67; 7,33]^T$$

$$J_1(3) = \sum_{x \in T_j(k)} d^2(x, \mu_j(3)) = 9.1$$

$$J_2(3) = \sum_{x \in T_j(k)} d^2(x, \mu_j(3)) = 16.3$$



Krok 4: Protože $\mu_1(3) \neq \mu_1(2)$ a $\mu_2(3) \neq \mu_2(2) \Rightarrow$ návrat na krok 2

Krok 2: Tento krok již nepřinese žádnou změnu $\Rightarrow T_1(3) = T_1(2)$ a $T_2(3) = T_2(2)$

Krok 3: stejné výsledky, výpočet středů i kriteria je stejný

Krok 4: $\mu_1(4) = \mu_1(3)$ a $\mu_2(4) \neq \mu_2(3) \Rightarrow$ algoritmus končí (už nic nezměním)

Algoritmus K-means konvergoval:

$$\begin{aligned} T_1(3) = T_1(2) &= \{x_1, x_2, \dots, x_8\} & \mu_1 &= [1,25; 1,13]^T \\ T_2(3) = T_2(2) &= \{x_9, x_{10}, \dots, x_{20}\} & \mu_2 &= [7,67; 7,33]^T \end{aligned}$$

