Neuronové sítě

Shluková analýza

Metrický prostor

Metrický prostor $\langle \mathbb{M}, \mathrm{d} \rangle$

$\mathrm{d}(\overline{x}, \overline{y})$ – metrika

$\forall x, y, z \in \mathbb{M}$:

1. $\mathrm{d}(\overline{x}, \overline{y}) \geq 0$
2. $\mathrm{d}(\overline{x}, \overline{y}) = \mathrm{d}(\overline{y}, \overline{x})$
3. $\mathrm{d}(\overline{x}, \overline{y}) = 0 \Leftrightarrow \overline{x} = \overline{y}$
4. $\mathrm{d}(\overline{x}, \overline{y}) \leq \mathrm{d}(\overline{x}, \overline{z}) + \mathrm{d}(\overline{z}, \overline{y})$

Semimetrický prostor

platí 1) – 3)
$\mathrm{d}(\overline{x}, \overline{y})$ – semimetrika

Vzdálenosti

Minkowského vzdálenost

$\langle \mathbb{R}^n, \mathrm{d}_1 \rangle \quad \mathrm{d}_1(\overline{x}, \overline{y}) = \sum_{i=1}^{n} |x_i - y_i|$ Manhattanská vzdálenost

$\langle \mathbb{R}^n, \mathrm{d}_2 \rangle \quad \mathrm{d}_2(\overline{x}, \overline{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$ Euklidovská vzdálenost

$\langle \mathbb{R}^n, \mathrm{d}_\infty \rangle \quad \mathrm{d}_\infty(\overline{x}, \overline{y}) = \max_{i=1,\ldots,n} |x_i - y_i|$ Čebyševova vzdálenost

$\langle \mathbb{R}^n, \mathrm{d}_p \rangle \quad \mathrm{d}_p(\overline{x}, \overline{y}) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}$ Minkowského vzdálenost

$p \geq 1 \quad (0 < p < 1 - \text{semimetrika})$

Vzdálenosti

Hammingova vzdálenost

$\langle \mathbb{M}^n, \mathrm{d}_H \rangle$

$\mathrm{d}_H(\alpha, \beta) = |\{i \in \hat{n} | a_i \neq b_i\}|, \quad \mathrm{d}_H(\alpha, \beta) \in \langle 0, n \rangle$

Levenshteinova vzdálenost

$\langle \mathbb{M}^*, \mathrm{d}_L \rangle$

$\mathrm{d}_L(\alpha, \beta) = n_I + n_D + l\, n_R, \quad l \in \langle 1, 2 \rangle$

$|n_1 - n_2| \leq \mathrm{d}_L(\alpha, \beta) \leq \max(n_1, n_2) \qquad |\alpha| = n_1, |\beta| = n_2$
$n_1 = n_2 \Rightarrow \mathrm{d}_L(\alpha, \beta) \leq \mathrm{d}_H(\alpha, \beta)$

Střed shluku

Shluk (cluster) $\begin{cases} \text{jak vybrat reprezentanta } x \in \mathbb{C} \text{ shluku } \mathbb{C} \\ \text{jak vybrat obecného reprezentanta shluku } \mathbb{C} \end{cases}$

Střed $\mathbb{C}$ v $\langle \mathbb{M}, \mathrm{d} \rangle$

$$\mathrm{f}(x) = \sum_{k=1}^{m} \mathrm{d}^2(x, x_k), \qquad x \in \mathbb{M}, \; x_k \in \mathbb{C} \subset \mathbb{M}$$

Každý prvek $x \in \mathbb{M}$, pro který platí $\mathrm{f}(x) = \min$, nazýváme středem shluku, $x \in s(\mathbb{C})$.

Souvislosti v $\mathbb{R}^n$

- $\langle \mathbb{R}, \mathrm{d}_1 \rangle$: $s(\mathbb{C}) = \{\mu\}$
  $$\mu = \frac{1}{m} \sum_{k=1}^{m} x_k$$
- $\langle \mathbb{R}^n, \mathrm{d}_2 \rangle$: $\overline{s}(\mathbb{C}) = \{\overline{\mu}\}$
  $$\overline{\mu} = \frac{1}{m} \sum_{k=1}^{m} \overline{x}_k$$

Shluková analýza

Shlukování

shluk – skupina s malým rozptylem a vekou vzdáleností od ostatních shluků

shlukování – disjunktní rozdělení dat na shluky

- rozdělení m objektů do H shluků
- metrický prostor $\langle \mathbb{M}, \mathrm{d} \rangle$
- $\mathbb{C} = \{x_1, x_2, \ldots, x_m\} \subset \mathbb{M}$
- $\mathbb{C} = \mathbb{C}_1 \cup \mathbb{C}_2 \cup \cdots \cup \mathbb{C}_H, \qquad i \neq j : \mathbb{C}_i \cap \mathbb{C}_j = \emptyset$

rozptyl shluku $\mathbb{C}$: $\sum_{x \in \mathbb{C}} \mathrm{d}^2(x, s(\mathbb{C}))$

součet přes všechny shluky : $J_e = \sum_{i=1}^{H} \sum_{x \in \mathbb{C}_i} (\mathrm{d}^2(x, s(\mathbb{C}_i)))$

- kritérium při shlukování, chybí zde ale podmínka velkých vzdáleností od ostatních shluků
- používá se při znalosti počtu shluků H

Shluková analýza

Algoritmy shlukové analýzy

k–means

1. Zvolíme k náhodných bodů (objektů) z $\mathbb{M}$ (nebo z $\mathbb{C}$) za středy shluků.
2. Přiřadíme každý objekt k nejbližšímu středu shluku.
3. Přepočítáme středy shluků.
4. Pokud se příslušnost objektů ke shlukům změnila, opakujeme druhý a třetí krok.

Poznámka: V našem případě $k = H$.

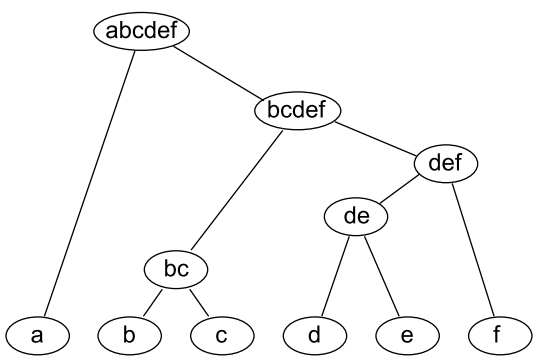
Cíl: Minimalizace kritéria J_e, které určuje kvalitu shlukování.

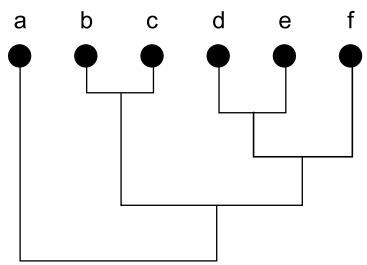
Algoritmy shlukové analýzy

Hierarchické shlukování

- Není nutné znát předem předpokládaný počet shluků, které data vytvářejí.
- Vycházíme z matice vzdáleností všech objektů.
- Na počátku každý objekt vytváří jeden shluk.
- V každém kroku dochází ke sloučení dvou nejbližších shluků.

dendrogram





Algoritmy shlukové analýzy

Vzdálenost mezi shluky $\mathbb{C}_i, \mathbb{C}_j \subset \mathbb{R}^n$ můžeme určit:

- metodou nejbližšího souseda
$$\mathrm{d}(\mathbb{C}_i, \mathbb{C}_j) = \min\{\mathrm{d}_2(\overline{x}, \overline{x}') \mid \overline{x} \in \mathbb{C}_i,\ \overline{x}' \in \mathbb{C}_j,\ i \neq j\},$$
- metodou nejvzdálenějšího souseda
$$\mathrm{d}(\mathbb{C}_i, \mathbb{C}_j) = \max\{\mathrm{d}_2(\overline{x}, \overline{x}') \mid \overline{x} \in \mathbb{C}_i,\ \overline{x}' \in \mathbb{C}_j,\ i \neq j\},$$
- průměrnou vzdáleností shluků $\mathbb{C}_i$ a $\mathbb{C}_j$
$$\mathrm{d}(\mathbb{C}_i, \mathbb{C}_j) = \frac{1}{m_i\, m_j} \sum_{\overline{x} \in \mathbb{C}_i} \sum_{\overline{x}' \in \mathbb{C}_j} \mathrm{d}_2(\overline{x}, \overline{x}'),$$
- **vzdáleností středů shluků $\mathbb{C}_i$ a $\mathbb{C}_j$**
$$\mathrm{d}(\mathbb{C}_i, \mathbb{C}_j) = \mathrm{d}_2(\overline{\mu}_i, \overline{\mu}_j),$$
- Wardovou metodou minimálního rozptylu
$$\mathrm{d}(\mathbb{C}_i, \mathbb{C}_j) = \frac{m_i\, m_j}{m_i + m_j}\, \mathrm{d}_2^2(\overline{\mu}_i, \overline{\mu}_j).$$

Poznámka: obecně lze vzít místo d_2 jinou metriku, na volbě metriky při shlukování záleží!

Shluková analýza

Algoritmy shlukové analýzy

Vzdálenost mezi shluky $\mathbb{C}_i, \mathbb{C}_j \subset \mathbb{M}^*$ můžeme určit:

- metodou nejbližšího souseda
$$\mathrm{d}(\mathbb{C}_i, \mathbb{C}_j) = \min\{\mathrm{d}_L(x, x') \mid x \in \mathbb{C}_i,\ x' \in \mathbb{C}_j,\ i \neq j\},$$
- **metodou nejvzdálenějšího souseda**
$$\mathrm{d}(\mathbb{C}_i, \mathbb{C}_j) = \max\{\mathrm{d}_L(x, x') \mid x \in \mathbb{C}_i,\ x' \in \mathbb{C}_j,\ i \neq j\},$$
- průměrnou vzdáleností shluků $\mathbb{C}_i$ a $\mathbb{C}_j$
$$\mathrm{d}(\mathbb{C}_i, \mathbb{C}_j) = \frac{1}{m_i\, m_j} \sum_{x \in \mathbb{C}_i} \sum_{x' \in \mathbb{C}_j} \mathrm{d}_L(x, x'),$$

Poznámka: V případě $\mathbb{C}_i, \mathbb{C}_j \subset \mathbb{M}^n$ lze použít také Hammingovu vzdálenost.