

Laboratorium 6 – algorytm k-środków

Streszczenie

Dla zbioru danych $\mathbf{X}_{n \times m}$ algorytm k-środków realizuje się za pomocą dwóch macierzy: $\mathbf{P}_{n \times K}$ – macierzy przynależności wektorów danych $\mathbf{x}_i = \mathbf{X}(i, :)$, $i = 1, 2, \dots, n$ do grupy C_k (macierzy stanów), przy czym $p_{ik} = \{0, 1\}$ oraz macierzy środków $\mathbf{C}_{K \times m}$, $\mathbf{c}_k = \mathbf{C}(k, :)$, $k = 1, 2, \dots, K$.

Krok 1. Wektory macierzy \mathbf{C} są inicjowane losowo.

Krok2. Dla każdego i oraz k :

- $p_{ik} = 1$, jeśli dla każdego $l \neq k$ zachodzi $d(\mathbf{x}_i, \mathbf{c}_k) < d(\mathbf{x}_i, \mathbf{c}_l)$; jeśli dla pewnego wektora danych minimalna odległość jest realizowana przez więcej, niż jeden środek grupy, to należy wybrać jeden z tych środków grup losowo;
- $p_{ik} = 0$, w przeciwnym przypadku.

Krok 3. Dla każdego k obliczyć $\mathbf{c}_k = \frac{\sum_{i=1}^n p_{ik} \mathbf{x}_i}{\sum_{i=1}^n p_{ik}}$

Krok 4. Powtarzać kroki 2 i 3 dopóki grupowanie nie ustabilizuje się (macierze \mathbf{P} i \mathbf{C} przestaną się zmieniać).

Krok 5. Każdy obiekt \mathbf{x}_i należy do klasy k w przypadku, gdy $p_{ik} = 1$.

1 Cel

Zapoznanie się z algorytmem k-środków oraz jego implementacja.

2 Zadania

1. Napisać funkcje:

- $\mathbf{d}=\text{distp}(\mathbf{X}, \mathbf{C}, \mathbf{e})$, która wyliczy odległość euklidesową między dwoma zbiorami punktów \mathbf{X} i \mathbf{C} :

$$d_e(\mathbf{x}_i, \mathbf{c}_k) = \sqrt{(\mathbf{x}_i - \mathbf{c}_k)(\mathbf{x}_i - \mathbf{c}_k)^T}.$$

- $\mathbf{d}=\text{distm}(\mathbf{X}, \mathbf{C}, \mathbf{V})$, która wyliczy odległość Mahalanobis'a między dwoma zbiorami punktów \mathbf{X} i \mathbf{C} ; \mathbf{V} jest macierzą kowariancji zbioru \mathbf{X} :

$$d_m(\mathbf{x}_i, \mathbf{c}_k) = \sqrt{(\mathbf{x}_i - \mathbf{c}_k) \mathbf{V}^{-1} (\mathbf{x}_i - \mathbf{c}_k)^T}.$$

- $[C, CX] = \text{ksrodki}(X, k)$, która dla zadanej macierzy wzorców X oraz liczby grup k , wyznaczy centra C i sąsiedztwa CX .
2. Zaimplementować algorytm k-środków. W postaci zbioru X wybrać zbiór `autos`.
 3. Zilustrować graficznie wyniki działania algorytmu.
 4. Obliczyć jakość grupowania:

$$F(C) = \frac{\sum_{1 \leq k < l \leq K} \sum_{x \in C_k} d(\mathbf{c}_k, \mathbf{c}_l)}{\sum_{k=1}^K \sum_{x \in C_k} d^2(\mathbf{x}, \mathbf{c}_k)}.$$