Klasteryzacja

Zadania

1. Napisz funkcje do realizacji algorytmów klasteryzacji hierarchicznej, K-means i C-means I i II

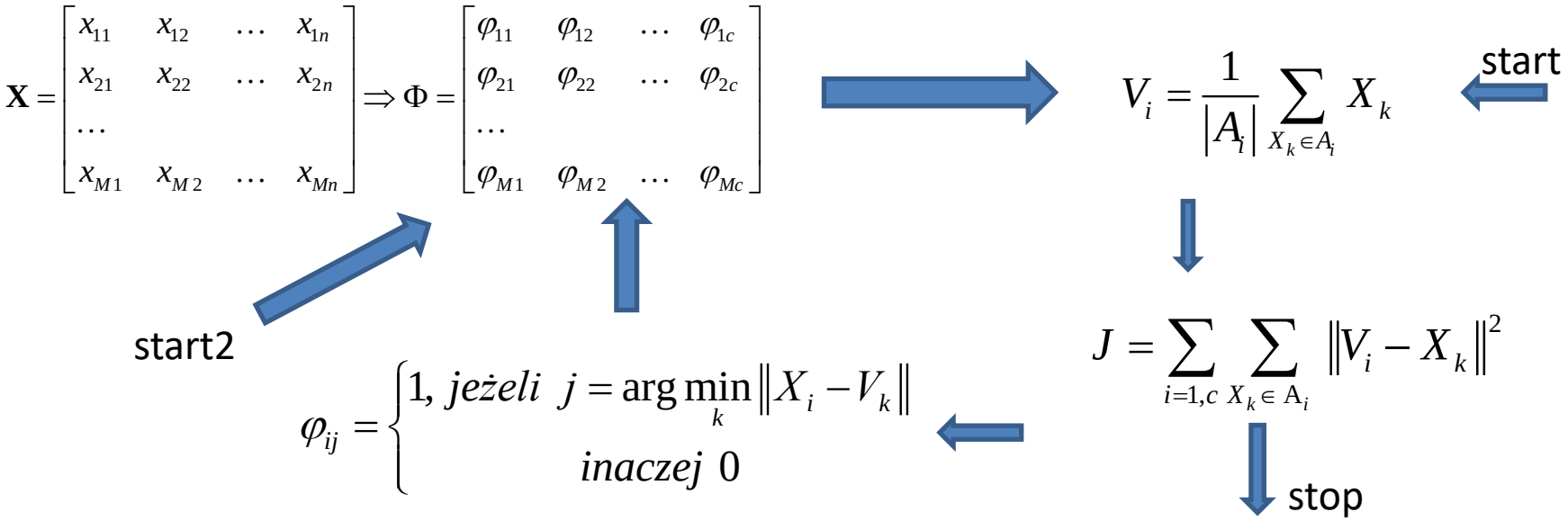
k	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x_1	1	1	1	3	3	3	5	7	9	11	11	11	13	13	13
x_2	1	4	7	2	4	6	4	4	4	2	4	6	1	4	7

2. Dodaj punkty (7,8),(7,16),(7,32) i sprawdź wyniki klasteryzacji

3. Zbadaj wyniki klasyfikacji dla różnych miar odległości
4. Oceń jakość klasteryzacji
5. Sprawdź działanie dla danych z repozytorium UCI, na przykład https://archive.ics.uci.edu/ml/datasets/iris

Algorytm K-means

Ogólna idea algorytmu K-means (MacQueen,1967) polega na tym, że rozpoczyna się od losowo wybranego grupowania punktów, następnie **ponownie przypisuje się punkty tak, aby *otrzymać największy spadek w funkcji oceny***, po czym przelicza się zaktualizowane skupienia, **po raz kolejny przypisuje się punkty** i tak dalej aż do momentu, w którym nie ma już żadnych zmian w funkcji oceny lub w składzie skupień. To zachłanne podejście ma tę zaletę, że jest proste i gwarantuje otrzymanie co najmniej lokalnego maksimum (minimum) funkcji oceny.



Algorytm C-means I

1. Ustalamy C – ilość klastrów, m - waga; ε – kryterium zakończenia

2. Losowa generacja macierzy F $\quad \mathbf{F} = [\mu_{ik}], \quad \mu_{ik} \in [0,1], \quad k = \overline{1,M}, \quad i = \overline{1,c}$

3. Obliczenia centrów klastrów $\quad V_i = \dfrac{\sum_{k=1,M} (\mu_{ik})^m X_k}{\sum_{k=1,M} (\mu_{ik})^m}, \quad i = 1,c$

4. Obliczenia odległości punktów do centrów klastrów

$$D_{ik} = \sqrt{\|X_k - V_i\|^2}, \quad k = \overline{1,M}, \quad i = \overline{1,c}$$

$$\chi = \sum_{i=1,c} \sum_{k=1,M} (\mu_{ik})^m \|X_k - V_i\|^2$$

Algorytm C-means I (cd)

5. Obliczenie nowej macierzy F

Jeżeli $D_{ik} > 0$ to $$\mu_{ik} = \frac{1}{D_{ik}^2 \sum_{j=1,c} \left(\frac{1}{D_{jk}^2} \right)^{2/(m-1)}}$$

Jeżeli $D_{ik} = 0$ to $\mu_{ik} = 1$

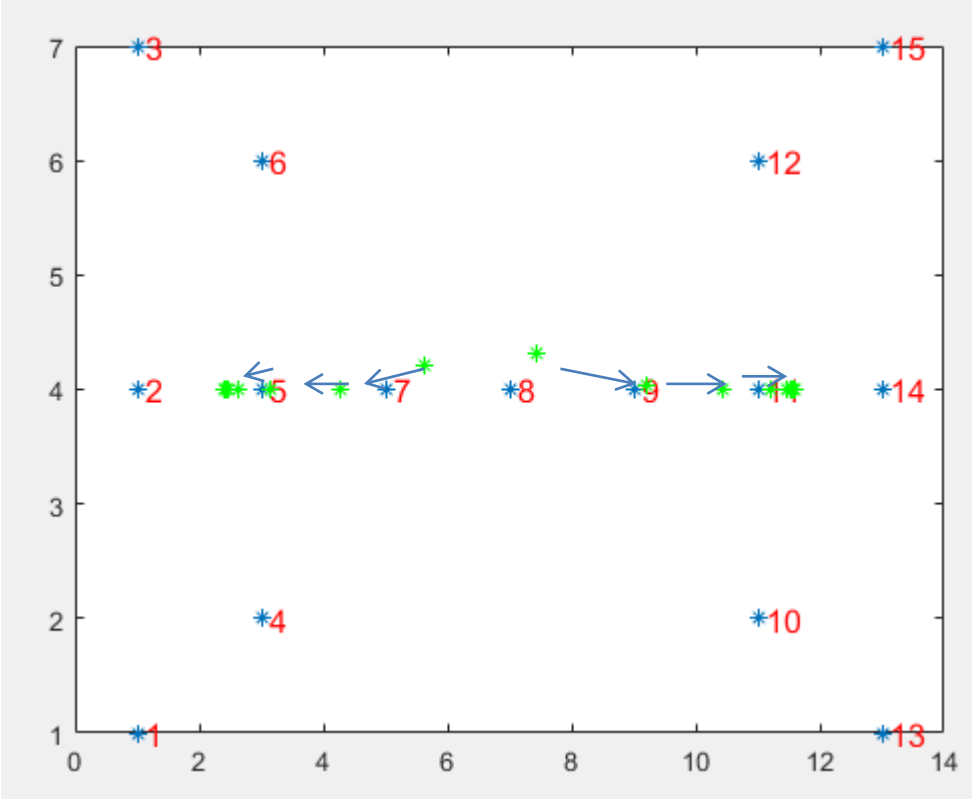
6. Sprawdzanie warunku zakończenia $\left\| F - F^* \right\|^2 < \varepsilon$

$$D_{ik} = \sqrt{\left\| X_k - V_i \right\|^2}, \quad k = \overline{1, M}, \quad i = \overline{1, c}$$

cmeansOS.m

$$\chi = \sum_{i=1,c} \sum_{k=1,M} (\mu_{ik})^m \left\| X_k - V_i \right\|^2$$

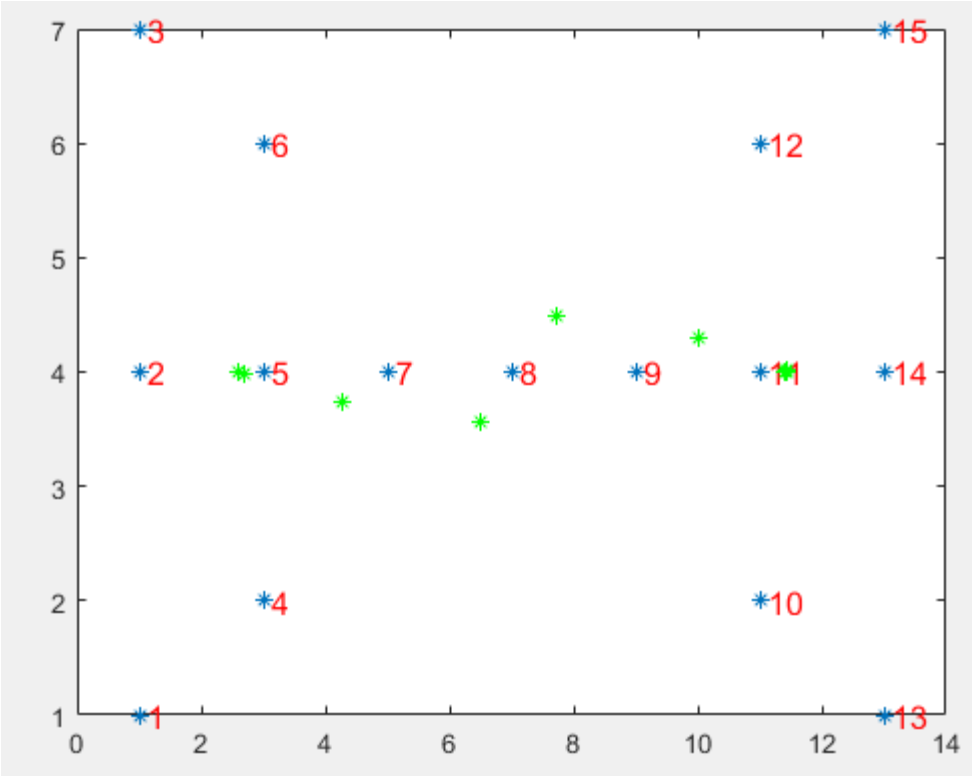
Przykładowe badania



m=5 NumIter=10 J = 45.3951607251449

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0.2312	0.1170	0.2312	0.1912	0.0649	0.1912	0.2824	0.5000	0.7176	0.8088	0.9351	0.8088	0.7688	0.8830	0.7688
0.7688	0.8830	0.7688	0.8088	0.9351	0.8088	0.7176	0.5000	0.2824	0.1912	0.0649	0.1912	0.2312	0.1170	0.2312

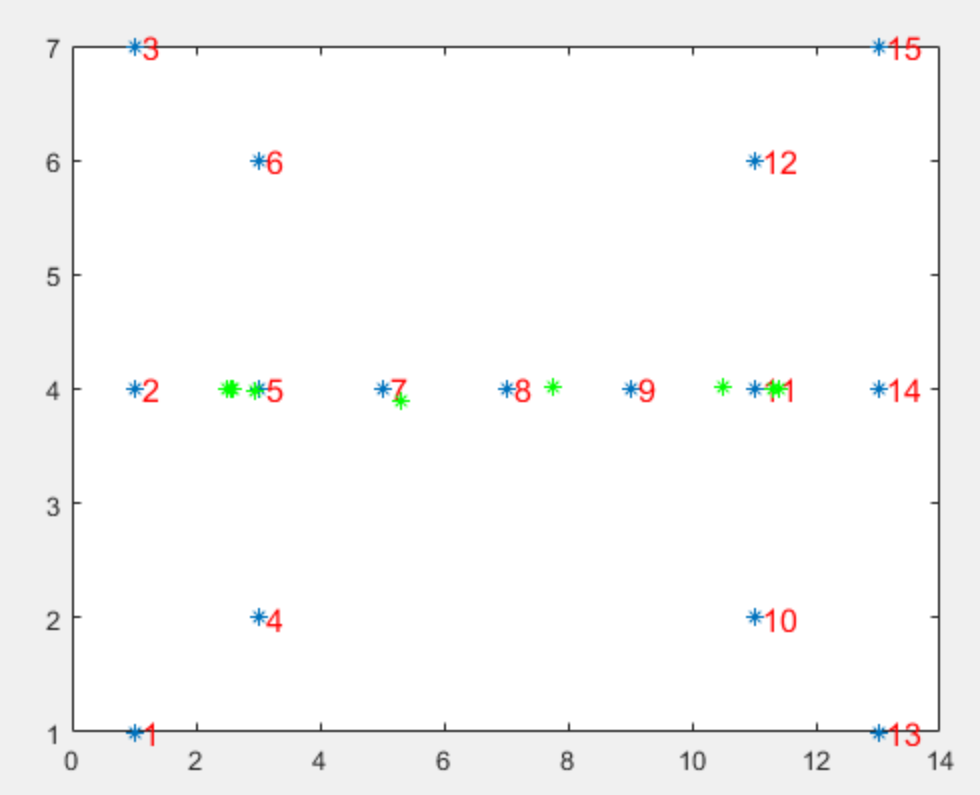
m=2



NumIter=15 J = 169.419907506014

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0.0095	5.3074e-04	0.0095	0.0031	6.1042e-06	0.0031	0.0198	0.5000	0.9802	0.9969	1.0000	0.9969	0.9905	0.9995	0.9905
0.9905	0.9995	0.9905	0.9969	1.0000	0.9969	0.9802	0.5000	0.0198	0.0031	6.1065e-06	0.0031	0.0095	5.3070e-04	0.0095

m=1 (K-means!)



J = 169.42076563828

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0.9905	0.9995	0.9905	0.9969	1.0000	0.9969	0.9802	0.5000	0.0198	0.0031	6.1053e-06	0.0031	0.0095	5.3072e-04	0.0095
0.0095	5.3072e-04	0.0095	0.0031	6.1053e-06	0.0031	0.0198	0.5000	0.9802	0.9969	1.0000	0.9969	0.9905	0.9995	0.9905

Cmeans II

$$J(X;U,V)=\sum_{i=1}^{c}\sum_{k=1}^{M}(\mu_{ik})^m\left\|x_k-v_i\right\|_A^2+\sum_{i=1}^{c}\eta_i\sum_{k=1}^{M}(1-\mu_{ik})^m$$

$$\eta_i=K\frac{\sum_{k=1}^{M}\mu_{ik}^m D_{ik}^2}{\sum_{k=1}^{M}\mu_{ik}^m},1\le i\le c$$

$$t_{ik}=\frac{1}{1+\left(\frac{D_{ik}^2}{\eta_i}\right)^{1/(m-1)}}1\le i\le c;1\le k\le M$$

$$\mathbf{v}_i=\frac{\sum_{k=1}^{M}t_{ik}^m\mathbf{x}_k}{\sum_{k=1}^{M}t_{ik}^m},\quad 1\le i\le c;1\le k\le M$$