ANALIZA SKUPIEŃ

Zbiór metod służących do wyodrębniania jednorodnych podzbiorów obiektów populacji nosi nazwę ***analiza skupień***.

Podstawową ideą analizy skupień jest rozdzielenie obiektów na pewną (ustaloną lub nieustaloną z góry) liczbę grup „podobnych” do siebie obiektów, które jednocześnie nie są „podobne” do obiektów z pozostałych grup.

- Jak określić prawdopodobieństwo obiektów?
- Jakimi metodami zidentyfikować skupienia?
- Jakie są założenia i ograniczenia analizy skupień?

Miary odległości (1)

Jest n obiektów: $O_1, O_2, .., O_n$. Należy rozdzielić te obiekty na podzbiory $S_1, S_2, ..., S_K$ populacji generalnej Ω, spełniającej następujące warunki:

- rozłączność: $S_i \cap S_k = \varnothing, i \neq k$;
- zupełność: $\bigcup S_i = \Omega$.

Niech każdy obiekt opisany jest przez parametry, np. obiekt O_x przez parametry $x_1, x_2, ..., x_p$, a O_y przez parametry $y_1, y_2, ..., y_p$, gdzie $x = 1,2,...,n$; $y = 1,2,...,n$. Każdy obiekt jest zatem opisany przez pewien punkt w przestrzeni p wymiarowej.

Miary odległości (2)

- odległość Czebyszewa:

$$d\left(O_x, O_y\right) = \max_{i=1,2,\ldots,p} \left\{ \left| x_i - y_i \right| \right\}.$$

- odległość Euklidesowa:

$$d\left(O_x, O_y\right) = \sqrt{\sum_{i=1}^{p} \left(x_i - y_i \right)^2}.$$

- odległość miejska (Manhattan, City Block):

$$d\left(O_x, O_y\right) = \sum_{i=1}^{p} \left| x_i - y_i \right|.$$

Miary odległości (3)

- odległość Euklidesowa do kwadratu:

$$d\left(O_x, O_y\right) = \sum_{i=1}^{p} \left(x_i - y_i\right)^2 .$$

Wybieramy ją, gdy chcemy przypisać większą wagę obiektom, które są bardziej oddalone.

- Jeżeli dwie zmienne opisujące obiekt wysoko ze sobą korelują, to odległość euklidesowa może dawać wyniki mylące i wtedy polecane jest stosowanie odległości Mahalanobisa:

$$d\left(O_x, O_y\right) = \sqrt{\sum_{i=1}^{p} \sum_{j=1}^{p} \left(x_i - y_i\right) \cdot \left(x_j - y_j\right) \cdot s_{ij}} ,$$

gdzie s_{ij} odpowiedni element macierzy, odwrotnej do macierzy kowariancji.

Miary odległości (4)

Wybranie konkretnej metryki umożliwia utworzenie macierzy odległości:

$$D = \begin{pmatrix} 0 & d_{12} & d_{13} & \dots & d_{1n} \\ d_{21} & 0 & d_{23} & \dots & d_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ d_{n1} & d_{n2} & d_{3} & \dots & 0 \end{pmatrix}$$

, gdzie d_{ij} odległość obiektu i od obiektu j.

Macierz ta stanowi punkt wyjścia dla wielu procedur analizy skupień. Zwykle zmienne muszą być standaryzowane.

Miary odległości (5)

Metody analizy skupień dzielą się na *hierarchiczne* i *niehierarchiczne*.

Hierarchiczne metody zawierają dwie grupy technik:

- aglomeracyjne – początkowo każdy obiekt jest odrębnym skupieniem. Następnie stopniowo łączymy najbliższe siebie obiekty w nowe skupienia, aż do uzyskania jednego skupienia;
- podziałowe – początkowo wszystkie obiekty tworzą jedno skupienie, które kolejno dzielimy (rozszczepiamy) na mniejsze, aż do momentu uzyskania jednoelementowych skupień.

I. Grupowanie metodą k-średnich

Algorytm

1. Wybiera się losowo *k* obiektów jako początkowe środki *k* klas.

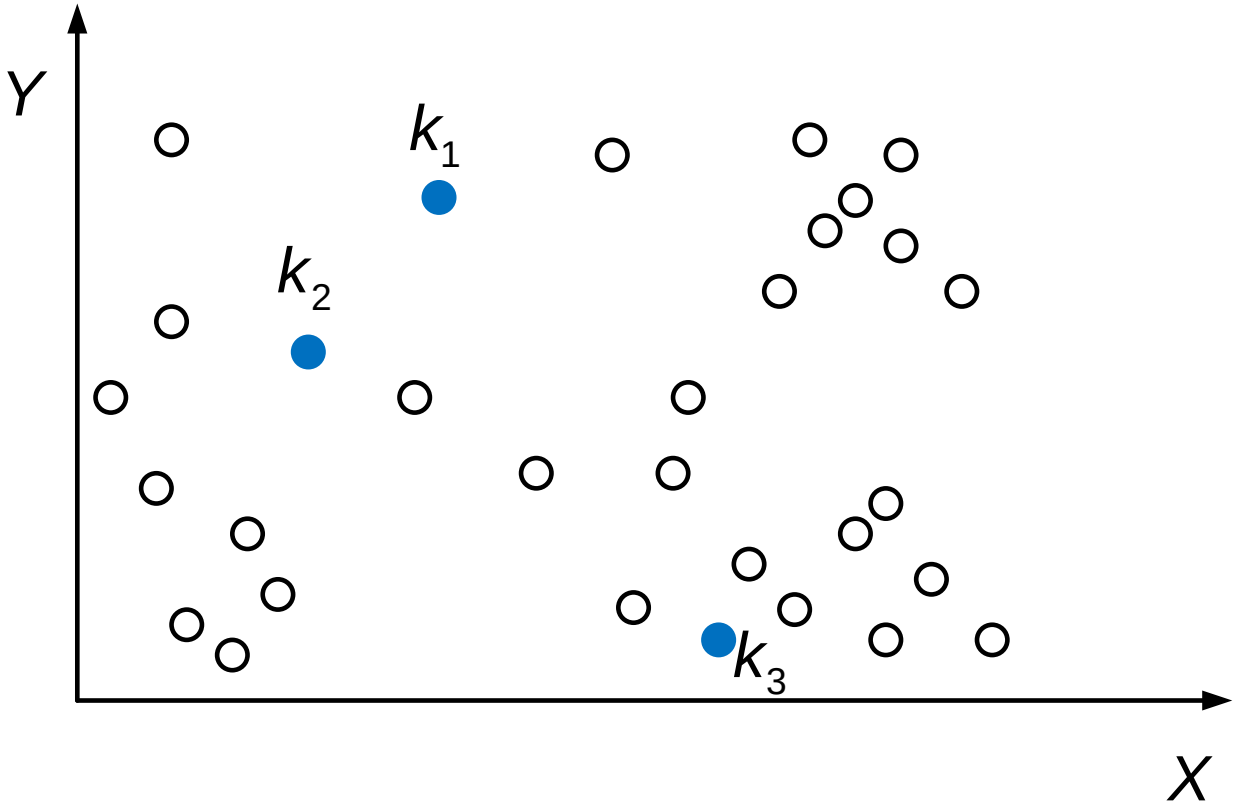
2. Obiekty przypisują się do klas: każdy obiekt jest przydzielany do tej klasy, dla której odległość obiektu od środka klasy jest najmniejsza.

3. Po przypisaniu (alokacji) obiektów do klas, uaktualniane są wartości średnie klas (środki klas) i powrót do kroku 2. Może się okazać, że na skutek aktualizacji średnich klas zachodzi konieczność przepisania obiektów. Proces przepisania obiektów i uaktualniania średnich klas jest powtarzany tak długo, jak długo występują zmiany przydziału obiektów do klas.

Warunek stopu może być zdefiniowany również w inny sposób, np. warunek time-out'u, zadana liczba iteracji, itp.

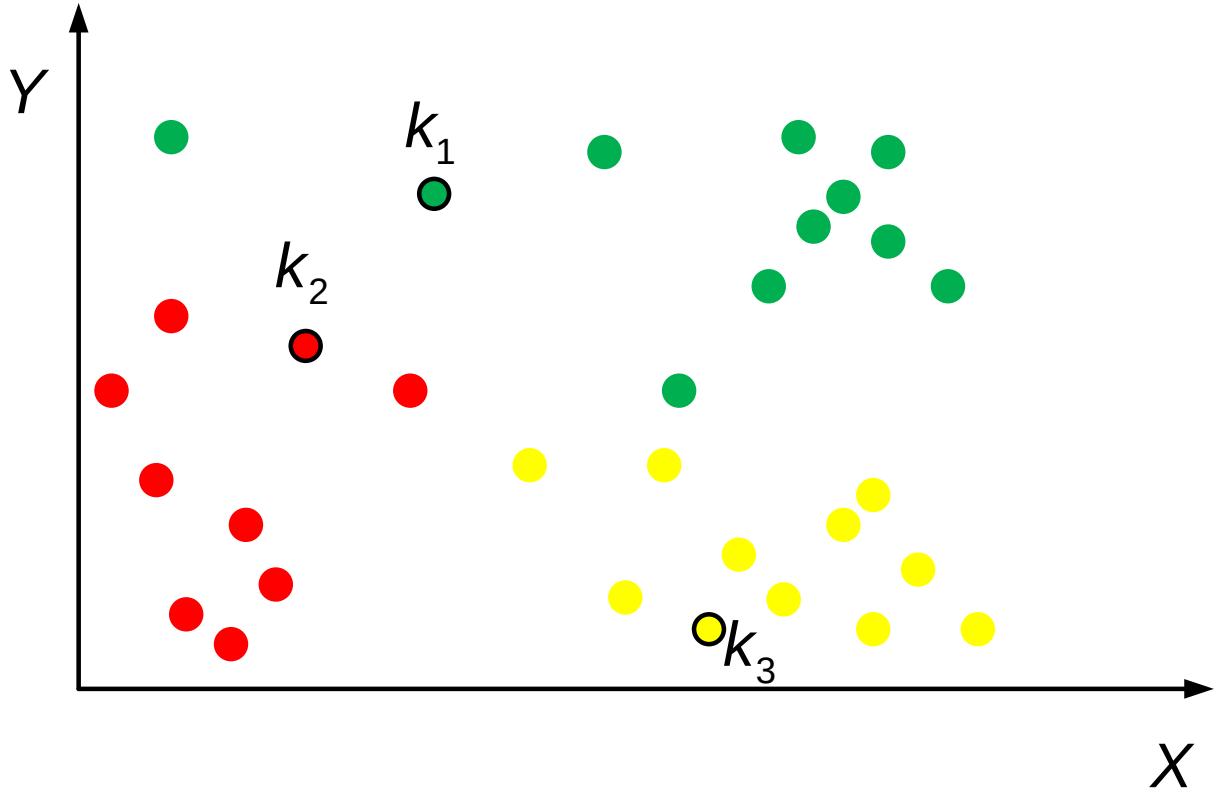
Przykład 1 (1)

W pierwszym kroku, algorytm wybiera losowo 3 punkty k_1, k_2, k_3, spośród zbioru obiektów, które początkowo stanowią środki trzech klas c_1, c_2, c_3.



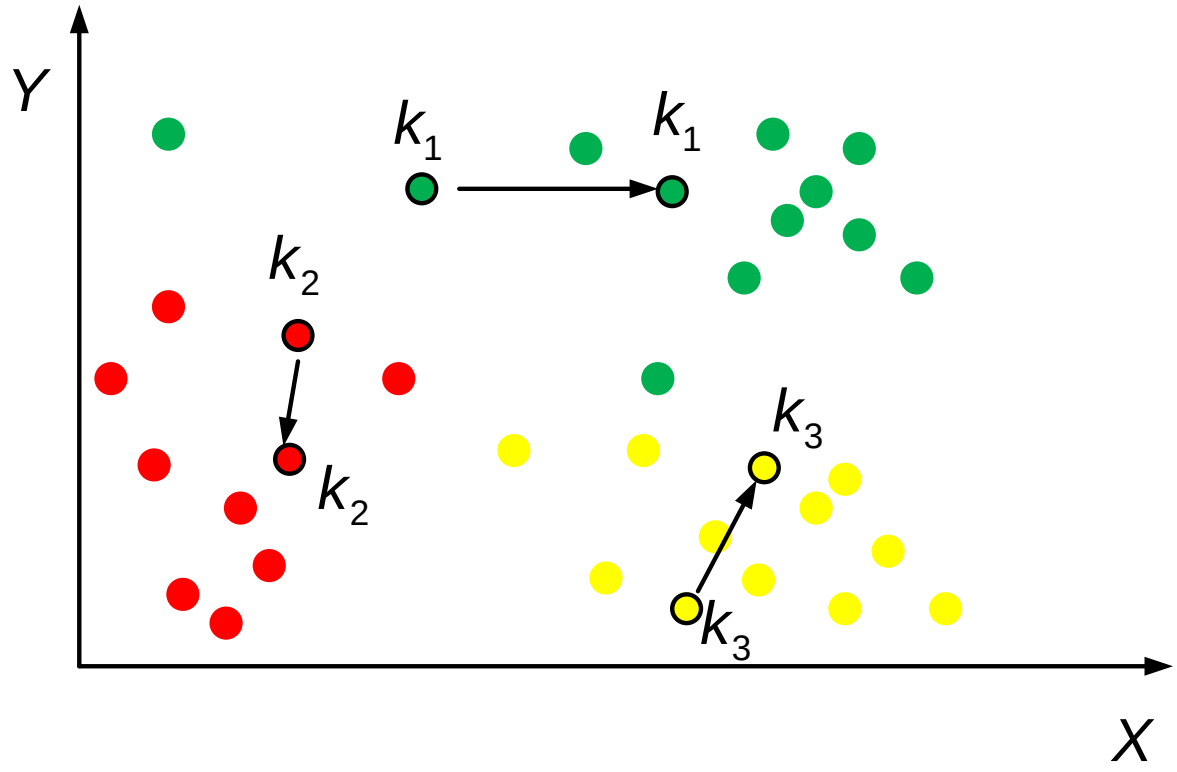
Przykład 1 (2)

W kroku drugim, algorytm przydziela obiekty do klas. Przydział obiektów do klas jest zaznaczony na rysunku kolorami.



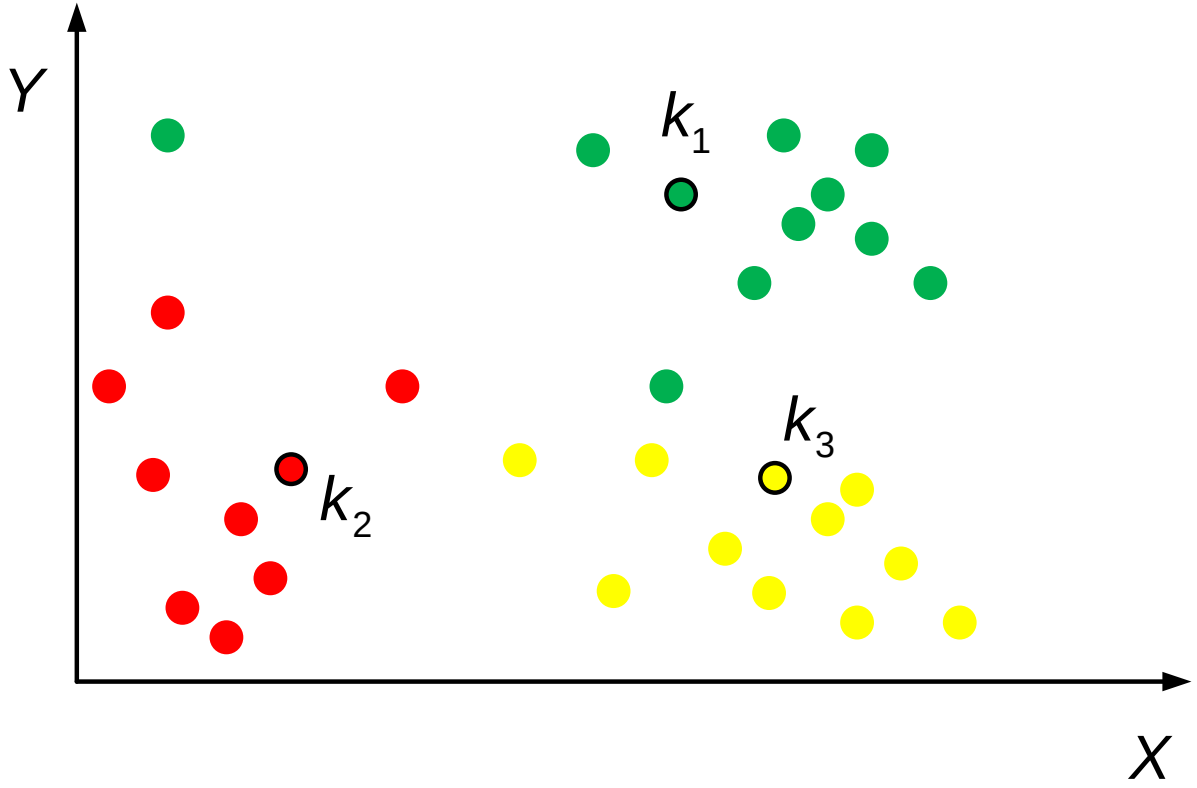
Przykład 1 (3)

W kolejnym kroku algorytmu następuje uaktualnienie średnich wszystkich klas.



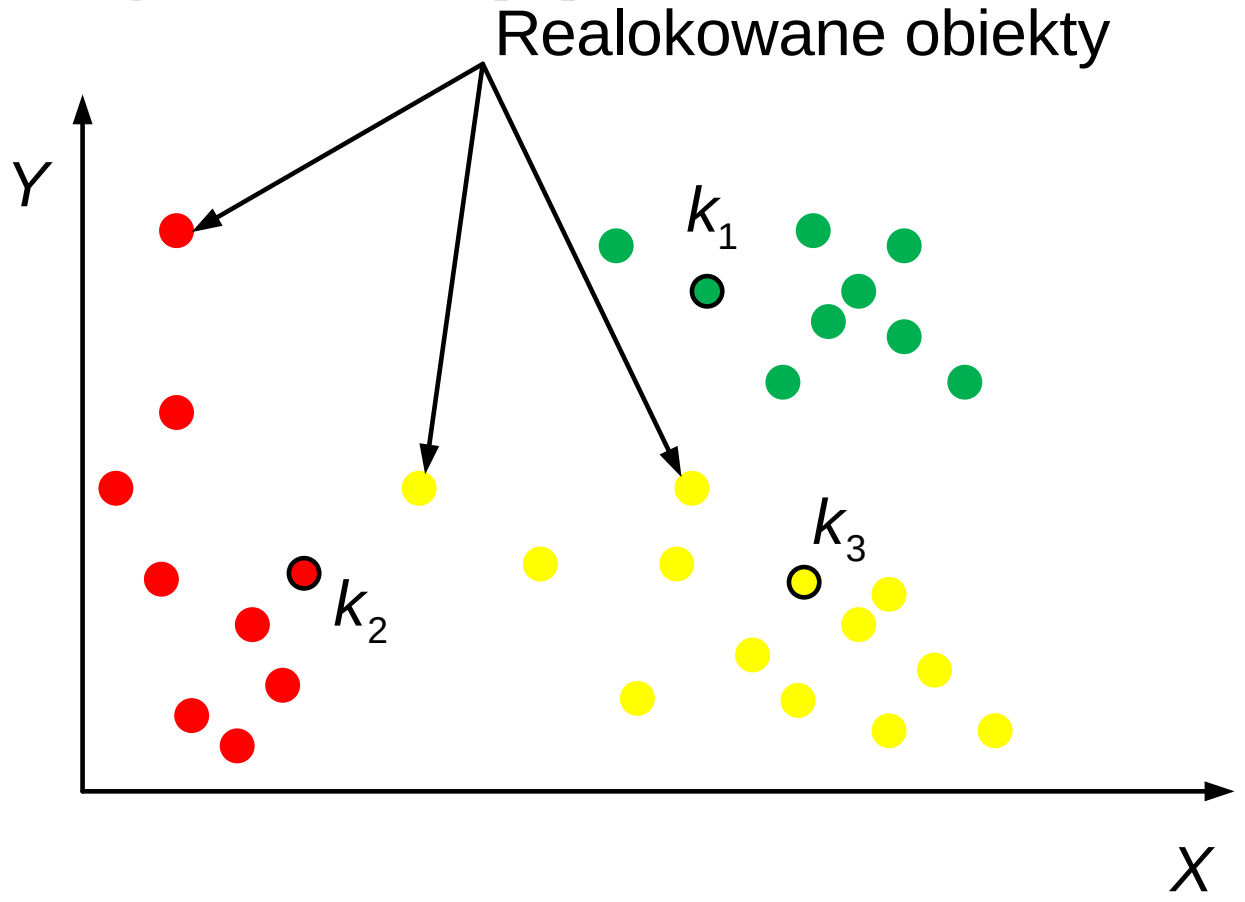
Przykład 1 (4)

Po uaktualnieniu średnich wszystkich klas następuje powrót do kroku drugiego – przepisaniu obiektów.



Przykład 1 (5)

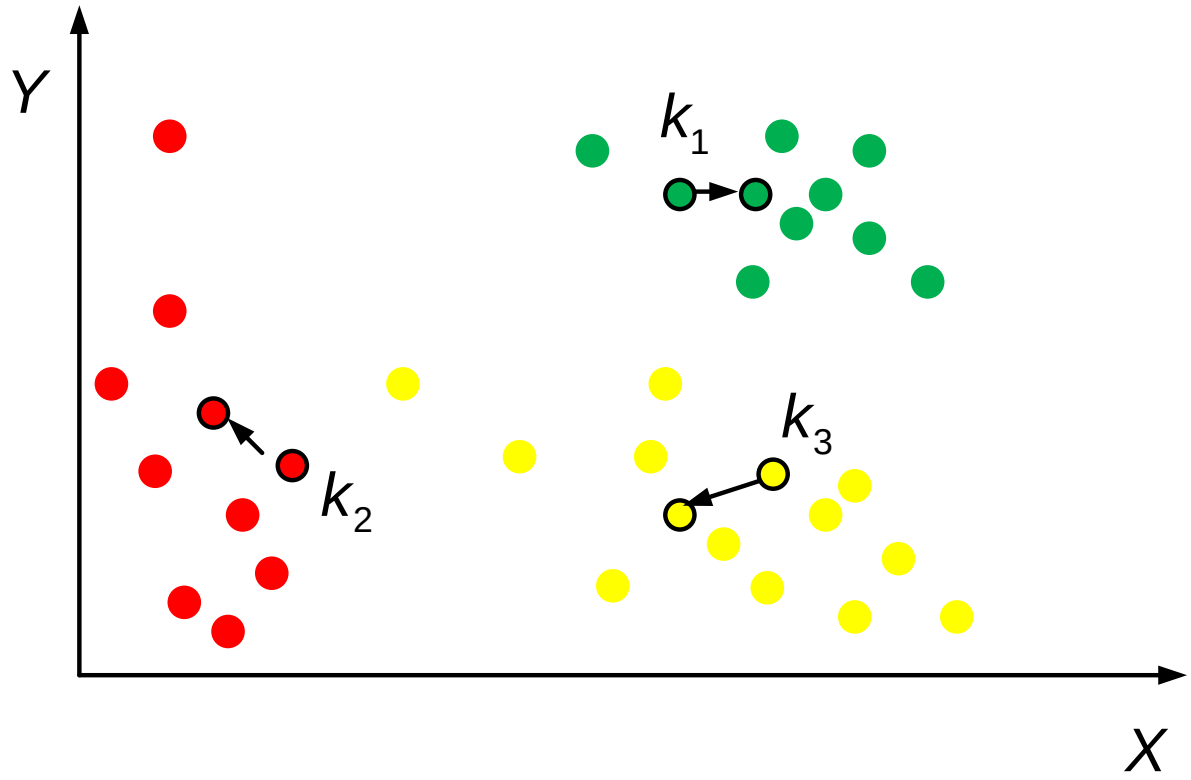
Konieczność przepisania (realokacji) dotyczy trzech obiektów.



Realokujemy obiekty do najbliższych klas i ponownie, przechodzimy do kolejnego kroku aktualizacji środków klas.

Przykład 1 (6)

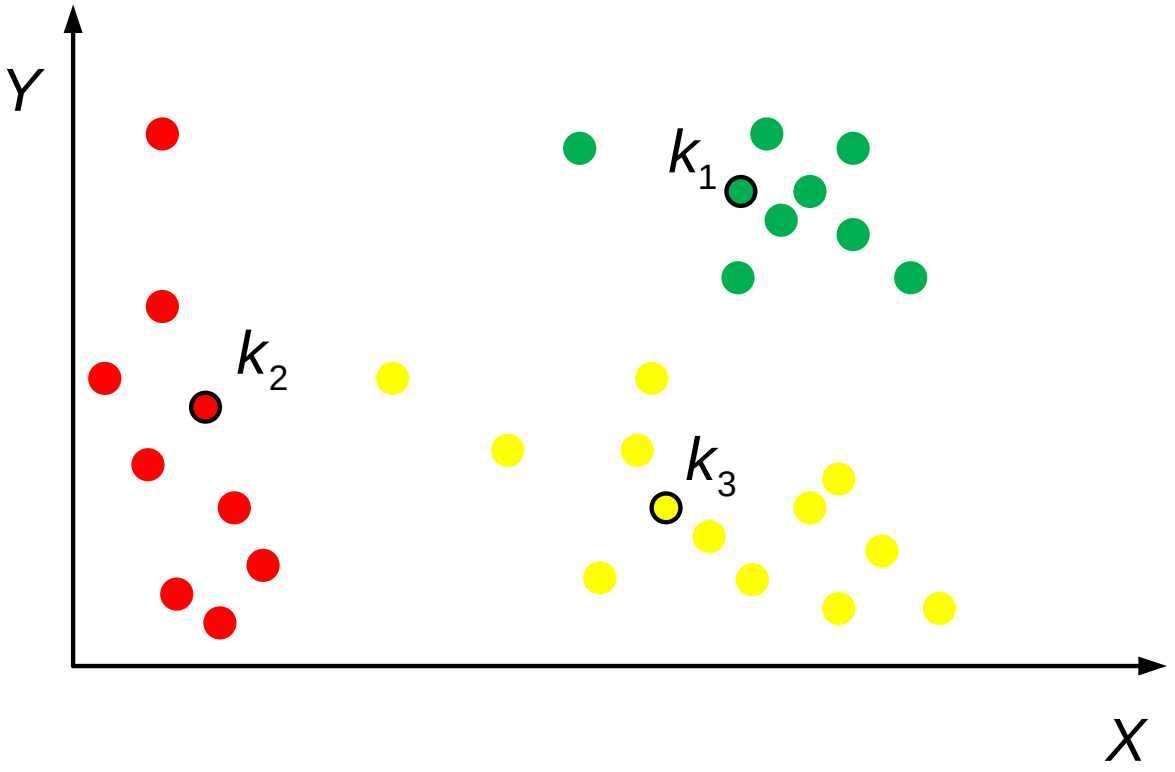
Ponownie aktualizujemy średnie wszystkich klas i wracamy do kroku przepisania obiektów.



Dla każdego obiektu następuje weryfikacja, czy obiekt ten podlega przepisaniu. Jeżeli żaden z obiektów nie wymaga przepisania, następuje zakończenie działania algorytmu.

Przykład 1 (7)

W wyniku działania algorytmu uzyskujemy trzy klasy c_1, c_2, c_3 przedstawione na rysunku.



Złożoność algorytmu k-średnich

Złożoność algorytmu k-średnich jest rzędu $O(k \cdot n \cdot I)$, gdzie I oznacza liczbę iteracji algorytmu, n liczbę grupowanych obiektów, k oznacza zadaną liczbę klas.

Zaletą algorytmu jest wysoka efektywność.

Wady:

- algorytm ten jest bardzo czuły na dane zaszumione lub dane zawierające punkty osobliwe;
- wynik działania algorytmu (tj. ostateczny podział obiektów pomiędzy klasami) mocno zależy od początkowego podziału obiektów.
- algorytm może „wpaść” w optimum lokalne, które może odbiegać od optimum globalnego.

II. Algorytm grupowania metodą EM (1)

Podstawową ideą algorytmu EM (*expectation-maximization*) jest założenie, że badany zbiór danych może być zmodelowany za pomocą liniowej kombinacji wielomianowych rozkładów normalnych, a celem jest ocena parametrów rozkładów, które maksymalizują logarytmiczną funkcję prawdopodobieństwa, która z kolej wykorzystywana jest jako miara jakości modelu.

Algorytm grupowania metodą EM (2)

Zalety:

- mocna podstawa statystyczna;
- liniowy wzrost złożoności przy zwiększeniu ilości danych;
- odporność na szum i braki danych;
- szybka zbieżność przy udanej inicjalizacji.

Wady:

- nie zawsze zmienne mają rozkład normalny;
- przy nieudanej inicjalizacji zbieżność może być wolna;
- algorytm może zatrzymać się w lokalnym minimum i dać quasi-optymalne rozwiązanie.

Algorytm grupowania metodą EM (3)

Zadaniem jest podział n obiektów na k klas w zależności od wartości zmiennych $X_1, X_2, ..., X_q$, charakteryzujących każdy obiekt.

Zapiszmy macierz wartości danych na dwa sposoby:

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1q} \\ x_{21} & x_{22} & \dots & x_{2q} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nq} \end{pmatrix} = \begin{pmatrix} X_1 & X_2 & \dots & X_q \end{pmatrix} = \begin{pmatrix} X^{(1)} \\ X^{(2)} \\ \dots \\ X^{(n)} \end{pmatrix},$$

gdzie $l = \overline{1,q}$ - liczba zmiennych, $i = \overline{1,n}$ - liczba obiektów.

Algorytm grupowania metodą EM (4)

Parametry modelu:

$W = \{w_1, w_2, ..., w_k\}$ - zbiór wag; dla wag musi być spełniony warunek $\sum_{j=1}^{k} w_j = 1$.

$$M = \begin{pmatrix} m_{11} & m_{12} & \dots & m_{1q} \\ m_{21} & m_{22} & \dots & m_{2q} \\ \dots & \dots & \dots & \dots \\ m_{k1} & m_{k2} & \dots & m_{kq} \end{pmatrix} = \begin{pmatrix} M_1 \\ M_2 \\ \\ M_k \end{pmatrix}$$ - zbiór wartości średnich,

gdzie m_{jl} - średnia wartość zmiennej X_l w klasie j.

C_j - macierz kowariancji w klasie j, $\dim(C_j) = q \times q$.

Algorytm grupowania metodą EM (5)

1. Początkowe parametry inicjalizacji algorytmu:

- Początkowe wagi są jednakowe: $w_j = \frac{1}{k}$.
- Początkowe wartości średnich są wybierane w sposób losowy.
- Początkowe macierze kowariancji są jednakowe w każdej klasie i są równe macierzy kowariancji zmiennych X

2. Dla każdego obiektu $X^{(i)}$ rozpatruje się mieszanka rozkładów normalnych:

$$p_j\left(X^{(i)}\right) = \frac{1}{(2\pi)^{\frac{q}{2}}\sqrt{\det\left(C_j\right)}} e^{-\frac{1}{2}\left(X^{(i)}-M_j\right)C_j^{-1}\left(X^{(i)}-M_j\right)^T}.$$

Algorytm grupowania metodą EM (6)

3. Krok *E* - *oczekiwanie*: obliczenie oczekiwanych wartości ukrytych zmiennych $g_{ij}, i = \overline{1,n}, j = \overline{1,k}$ wg parametrów (W,M,C):

$$g_{ij} = \frac{w_j p_j\left(x^{(i)}\right)}{\sum_{j=1}^{k} w_j p_j\left(x^{(i)}\right)},$$

czyli prawdopodobieństw przynależności i-go obiektu do klasy j.

$\dim(G) = n \times k$.

Algorytm grupowania metodą EM (7)

4. Obliczenie funkcji wiarygodności:

$$LL = \ln \prod_{i=1}^{n} p\left(X^{(i)}\right) = \sum_{i=1}^{n} \ln \sum_{j=1}^{k} w_j p_j\left(X^{(i)}\right)$$

5. Krok *M* - maksymalizacja logarytmu funkcji wiarygodności wynikiem której są nowe wartości parametrów modelu:

$$w_j = \frac{\sum_{i=1}^{n} g_{ij}}{n}; \quad M_j = \frac{\sum_{i=1}^{n} g_{ij} X^{(i)}}{n w_j};$$

$$C_j = \frac{\left(X^{(i)} - M_j\right)^T g_{ij} \left(X^{(i)} - M_j\right)}{n w_j}.$$

Algorytm grupowania metodą EM (8)

6. Porównujemy wartość funkcji wiarygodności z wartością funkcji z poprzedniej iteracji (jeśli nr iteracji $t = 1$, to przyjmujemy, że na poprzednim kroku $LL = 0$).

Jeśli wartość bezwzględna różnicy funkcji wiarygodności jest mniejsza, niż dopuszczalny błąd obliczeń δ (który zadaje wykonawca), to algorytm się kończy:

$$if \quad \Delta LL = LL\left(iteracja\ t\right) - LL\left(iteracja\left(t-1\right)\right) \le \delta,$$

Inaczej należy wrócić do kroku 2.

Jeszcze jednym dodatkowym ograniczeniem jest podanie maksymalnej liczby iteracji.

Przykład 2 (1)

Podzielić 6 obiektów wg 2 zmiennych na dwie grupy:

X_1	X_2
1	1
6	7
3	2
4	6
5	7
2	1

Błąd dopuszczalny δ wybrać 0,001.

Przykład 2 (2)

Dzielimy $n = 6$ obiektów na $k = 2$ klasy w zależności od wartości zmiennych X_1, X_2, $q = 2$, charakteryzujących każdy obiekt.

Macierz danych:

$$X = \begin{pmatrix} 1 & 1 \\ 6 & 7 \\ 3 & 2 \\ 4 & 6 \\ 6 & 7 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} X_1 & X_2 \end{pmatrix} = \begin{pmatrix} x^{(1)} \\ x^{(2)} \\ \dots \\ x^{(6)} \end{pmatrix},$$

Przykład 2 (3)

1. Początkowe parametry inicjalizacji algorytmu:

Początkowe wagi są jednakowe: $w_j = \frac{1}{k}$, musi być spełniony warunek $\sum_{j=1}^{k} w_j = 1$.

$$W = \{w_1, w_2\} = \left\{\frac{1}{2}, \frac{1}{2}\right\}.$$

Przykład 2 (4)

Początkowe wartości średnich są wybierane w sposób losowy:

$$M = \begin{pmatrix} 3,5 & 3,8 \\ 3,5 & 4 \end{pmatrix}_{k \times q} = \begin{pmatrix} M_1 \\ M_2 \end{pmatrix}$$

- zbiór wartości średnich, gdzie przykładowo 3,8 wartość średniej dla zmiennej X_2 w klasie 1.

Przykład 2 (5)

Początkowe macierze kowariancji są jednakowe w każdej klasie i są równe macierzy kowariancji zmiennych X:

$$\overline{X} = \begin{pmatrix} 3{,}5 & 4 \end{pmatrix}$$

$$C_1 = C_2 = \frac{1}{6} \begin{pmatrix} 1-3{,}5 & 1-4 \\ 6-3{,}5 & 7-4 \\ 3-3{,}5 & 2-4 \\ 4-3{,}5 & 6-4 \\ 5-3{,}5 & 7-4 \\ 2-3{,}5 & 1-4 \end{pmatrix}^T \begin{pmatrix} 1-3{,}5 & 1-4 \\ 6-3{,}5 & 7-4 \\ 3-3{,}5 & 2-4 \\ 4-3{,}5 & 6-4 \\ 5-3{,}5 & 7-4 \\ 2-3{,}5 & 1-4 \end{pmatrix} = \begin{pmatrix} 2{,}9 & 4{,}3 \\ 4{,}3 & 7{,}3 \end{pmatrix}$$

$$\dim\left(C_j\right) = q \times q = 2 \times 2$$

Przykład 2 (6)

2. Dla każdego obiektu $X^{(i)}$ rozpatruje się mieszanka rozkładów normalnych ($j = 1,2$ - nr klasy, $i = 1,2,..6$ - nr obiektu, $q = 2$ liczba zmiennych):

$$p_j\left(X^{(i)}\right) = \frac{1}{(2\pi)^{\frac{q}{2}}\sqrt{\det\left(C_j\right)}} e^{-\frac{1}{2}\left(X^{(i)}-M_j\right)C_j^{-1}\left(X^{(i)}-M_j\right)^T}$$

$d^{(i)} = \left(X^{(i)} - M_j\right)C_j^{-1}\left(X^{(i)} - M_j\right)^T$ - odległość Mahalanobisa.

$$p_j\left(X^{(i)}\right) = \frac{1}{2\pi\sqrt{\det\left(C_j\right)}} e^{-\frac{1}{2}d^{(i)}}$$

Przykład 2 (7)

Klasa 1 ($j = 1$):

$x^{(i)}$		M_1	$\left(x^{(i)} - M_1\right)$	$d^{(i)}$	$p_1\left(x^{(i)}\right)$
$x^{(1)}$	$(1 \quad 1)$	$(3{,}5 \quad 3{,}8)$	$(-2{,}5 \quad -2{,}8)$	3,1	0,02
$x^{(2)}$	$(6 \quad 7)$	$(3{,}5 \quad 3{,}8)$	$(2{,}5 \quad 3{,}2)$	2,4	0,03
$x^{(3)}$	$(3 \quad 2)$	$(3{,}5 \quad 3{,}8)$	$(-0{,}5 \quad -1{,}8)$	1,3	0,05
$x^{(4)}$	$(4 \quad 6)$	$(3{,}5 \quad 3{,}8)$	$(0{,}5 \quad 2{,}2)$	2,5	0,03
$x^{(5)}$	$(5 \quad 7)$	$(3{,}5 \quad 3{,}8)$	$(1{,}5 \quad 3{,}2)$	1,8	0,04
$x^{(6)}$	$(2 \quad 1)$	$(3{,}5 \quad 3{,}8)$	$(-1{,}5 \quad -2{,}8)$	1,1	0,06

Przykład 2 (8)

Przykładowo dla $X^{(4)}$:

$$d^{(4)} = \left(X^{(4)} - M_1\right) C_1^{-1} \left(X^{(4)} - M_1\right)^T =$$

$$= \begin{pmatrix} 0,5 & 2,2 \end{pmatrix} \begin{pmatrix} 2,9 & 4,3 \\ 4,3 & 7,3 \end{pmatrix}^{-1} \begin{pmatrix} 0,5 \\ 2,2 \end{pmatrix} = 2,5$$

$$p_1\left(X^{(4)}\right) = \frac{1}{2\pi\sqrt{\det\left(C_1\right)}} e^{-\frac{1}{2}d^{(4)}} =$$

$$= \frac{1}{2\pi \cdot 1,62} e^{-0,5 \cdot 2,5} = 0,03$$

Przykład 2 (9)

Klasa 2 ($j = 2$):

$X^{(i)}$		M_2	$\left(X^{(i)} - M_2\right)$	$d^{(i)}$	$p_2\left(X^{(i)}\right)$
$X^{(1)}$	$(1 \quad 1)$	$(3{,}5 \quad 4)$	$(-2{,}5 \quad -3)$	2,7	0,03
$X^{(2)}$	$(6 \quad 7)$	$(3{,}5 \quad 4)$	$(2{,}5 \quad 3)$	2,7	0,03
$X^{(3)}$	$(3 \quad 2)$	$(3{,}5 \quad 4)$	$(-0{,}5 \quad -2)$	1,9	0,04
$X^{(4)}$	$(4 \quad 6)$	$(3{,}5 \quad 4)$	$(0{,}5 \quad 2)$	1,9	0,04
$X^{(5)}$	$(5 \quad 7)$	$(3{,}5 \quad 4)$	$(1{,}5 \quad 3)$	1,4	0,05
$X^{(6)}$	$(2 \quad 1)$	$(3{,}5 \quad 4)$	$(-1{,}5 \quad -3)$	1,4	0,05

Przykład 2 (10)

3. Krok *E* - *oczekiwanie*: $g_{ij} = \dfrac{w_j p_j\left(x^{(i)}\right)}{\sum_{j=1}^{k} w_j p_j\left(x^{(i)}\right)}$

$p_1\left(x^{(i)}\right)$	$p_2\left(x^{(i)}\right)$	$w_1 p_1\left(x^{(i)}\right)$	$w_2 p_2\left(x^{(i)}\right)$	$\sum_{j=1}^{2} w_j p_j\left(x^{(i)}\right)$
0,02	0,03	$0{,}5 \cdot 0{,}02 = 0{,}01$	0,015	0,01+0,015= =0,025
0,03	0,03	0,015	0,015	0,03
0,05	0,04	0,025	0,02	0,045
0,03	0,04	0,015	0,02	0,035
0,04	0,05	0,02	0,025	0,045
0,06	0,05	0,03	0,025	0,055

Przykład 2 (11)

$w_1 p_1\left(x^{(i)}\right)$	$w_2 p_2\left(x^{(i)}\right)$	$\sum_{j=1}^{2} w_j p_j\left(x^{(i)}\right)$	g_{i1}	g_{i2}
0,01	0,015	0,025	$\frac{0,01}{0,025} = 0,4$	$\frac{0,015}{0,025} = 0,6$
0,015	0,015	0,03	$\frac{0,015}{0,03} = 0,5$	$\frac{0,015}{0,03} = 0,5$
0,025	0,02	0,045	0,55	0,44
0,015	0,02	0,035	0,43	0,55
0,02	0,025	0,045	0,44	0,55
0,03	0,025	0,055	0,55	0,45

Przykład 2 (12)

4. Obliczenie funkcji wiarygodności:

$$LL = \sum_{i=1}^{n} \ln \sum_{j=1}^{k} w_j p_j\left(X^i\right) = \sum_{i=1}^{6} \ln \sum_{j=1}^{2} w_j p_j\left(X^{(i)}\right)$$

i	$\sum_{j=1}^{2} w_j p_j\left(X^{(i)}\right)$	$\ln \sum_{j=1}^{2} w_j p_j\left(X^{(i)}\right)$
1	0,025	–3,7
2	0,03	–3,5
3	0,045	–3,1
4	0,035	–3,35
5	0,045	–3,1
6	0,055	–2,9
	LL	–19,65

$$LL = -19{,}65.$$

Przykład 2 (13)

5. Krok *M* - maksymalizacja logarytmu funkcji wiarygodności.

Nowe wartości parametrów modelu: $w_j = \left(\sum_{i=1}^{n} g_{ij} \right) / n$

i	g_{i1}	g_{i2}
1	0,4	0,6
2	0,5	0,5
3	0,55	0,44
4	0,43	0,55
5	0,44	0,55
6	0,55	0,45
$\sum_{i=1}^{n} g_{ij}$	2,87	3,1

$$w_1 = \frac{\sum_{i=1}^{n} g_{i1}}{6} = \frac{2,87}{6} = 0,48$$

$$w_2 = \frac{\sum_{i=1}^{n} g_{i2}}{6} = \frac{3,1}{6} = 0,52$$

Czyli nowe wagi:

$W = \{w_1, w_2\} = \{0,48;\ 0,52\}.$

Przykład 2 (14)

$$M_j = \left(\sum_{i=1}^{n} g_{ij} x^{(i)} \right) / \left(n w_j \right)$$

i	$x^{(i)}$	g_{i1}	g_{i2}	$g_{i1} x^{(i)}$	$g_{i2} x^{(i)}$
1	$(1 \quad 1)$	0,4	0,6	$(0,4 \quad 0,4)$	$(0,6 \quad 0,6)$
2	$(6 \quad 7)$	0,5	0,5	$(3 \quad 3,5)$	$(3 \quad 3,5)$
3	$(3 \quad 2)$	0,55	0,44	$(1,65 \quad 1,1)$	$(1,32 \quad 0,88)$
4	$(4 \quad 6)$	0,43	0,55	$(1,72 \quad 2,58)$	$(2,2 \quad 3,3)$
5	$(5 \quad 7)$	0,44	0,55	$(2,2 \quad 3,08)$	$(2,75 \quad 3,85)$
6	$(2 \quad 1)$	0,55	0,45	$(1,1 \quad 0,55)$	$(0,9 \quad 0,45)$
$\sum_{i=1}^{n} g_{ij} x^{(i)}$				$(10,07 \quad 11,21)$	$(10,77 \quad 12,58)$

Przykład 2 (15)

$$M_1 = \frac{\sum_{i=1}^{n} g_{i1} x^{(i)}}{n w_1} = \frac{\begin{pmatrix} 10,07 & 11,21 \end{pmatrix}}{6 \cdot 0,48} = \begin{pmatrix} 3,5 & 3,9 \end{pmatrix}$$

$$M_2 = \frac{\sum_{i=1}^{n} g_{i2} x^{(i)}}{n w_2} = \frac{\begin{pmatrix} 10,77 & 12,58 \end{pmatrix}}{6 \cdot 0,52} = \begin{pmatrix} 3,45 & 4,03 \end{pmatrix}$$

Nowa macierz średnich:

$$M = \begin{pmatrix} M_1 \\ M_2 \end{pmatrix} = \begin{pmatrix} 3,5 & 3,9 \\ 3,45 & 4,03 \end{pmatrix}$$

Przykład 2 (16)

Macierz kowariancji C_1, klasa 1: $C_1 = \dfrac{\left(X^{(i)} - M_1\right)^T g_{i1} \left(X^{(i)} - M_1\right)}{6 \cdot w_1}$

g_{i1}	$X^{(i)}$	M_1	$\left(X^{(i)} - M_1\right)$	$g_{i1}\left(X^{(i)} - M_1\right)$
0,4	$\begin{pmatrix}1 & 1\end{pmatrix}$	$\begin{pmatrix}3,5 & 3,9\end{pmatrix}$	$\begin{pmatrix}-2,5 & -2,9\end{pmatrix}$	$\begin{pmatrix}-1 & -1,16\end{pmatrix}$
0,5	$\begin{pmatrix}6 & 7\end{pmatrix}$	$\begin{pmatrix}3,5 & 3,9\end{pmatrix}$	$\begin{pmatrix}2,5 & 3,1\end{pmatrix}$	$\begin{pmatrix}1,25 & 1,55\end{pmatrix}$
0,55	$\begin{pmatrix}3 & 2\end{pmatrix}$	$\begin{pmatrix}3,5 & 3,9\end{pmatrix}$	$\begin{pmatrix}-0,5 & -1,9\end{pmatrix}$	$\begin{pmatrix}-0,28 & -1,04\end{pmatrix}$
0,43	$\begin{pmatrix}4 & 6\end{pmatrix}$	$\begin{pmatrix}3,5 & 3,9\end{pmatrix}$	$\begin{pmatrix}0,5 & 2,1\end{pmatrix}$	$\begin{pmatrix}0,21 & 0,9\end{pmatrix}$
0,44	$\begin{pmatrix}5 & 7\end{pmatrix}$	$\begin{pmatrix}3,5 & 3,9\end{pmatrix}$	$\begin{pmatrix}1,5 & 3,1\end{pmatrix}$	$\begin{pmatrix}0,66 & 1,36\end{pmatrix}$
0,55	$\begin{pmatrix}2 & 1\end{pmatrix}$	$\begin{pmatrix}3,5 & 3,9\end{pmatrix}$	$\begin{pmatrix}-1,5 & -2,9\end{pmatrix}$	$\begin{pmatrix}-0,83 & 1,59\end{pmatrix}$

Przykład 2 (17)

$$C_1 = \frac{1}{6 \cdot 0,48} \begin{pmatrix} -2,5 & -2,9 \\ 2,5 & 3,1 \\ -0,5 & -1,9 \\ 0,5 & 2,1 \\ 1,5 & 3,1 \\ -1,5 & 2,9 \end{pmatrix}^T \begin{pmatrix} -1 & -1,16 \\ 1,25 & 1,55 \\ -0,28 & -1,04 \\ 0,21 & 0,9 \\ 0,66 & 1,36 \\ -0,83 & 1,59 \end{pmatrix} = \begin{pmatrix} 2,81 & 4,23 \\ 4,23 & 7,26 \end{pmatrix}$$

Analogicznie szukamy nową macierz kowariancji C_2 w klasie 2:

$$C_2 = \begin{pmatrix} 3 & 4,4 \\ 4,4 & 7,32 \end{pmatrix}$$

Przykład 2 (18)

6. Porównujemy wartość funkcji wiarygodności z wartością funkcji z poprzedniej iteracji (jeśli nr iteracji $t = 1$, to przyjmujemy, że na poprzednim kroku $LL = 0$).

$$\Delta LL = \left| LL(t) - LL(t-1) \right| = \left| -19{,}65 - 0 \right| = 19{,}65$$

$$\delta = 0{,}001$$

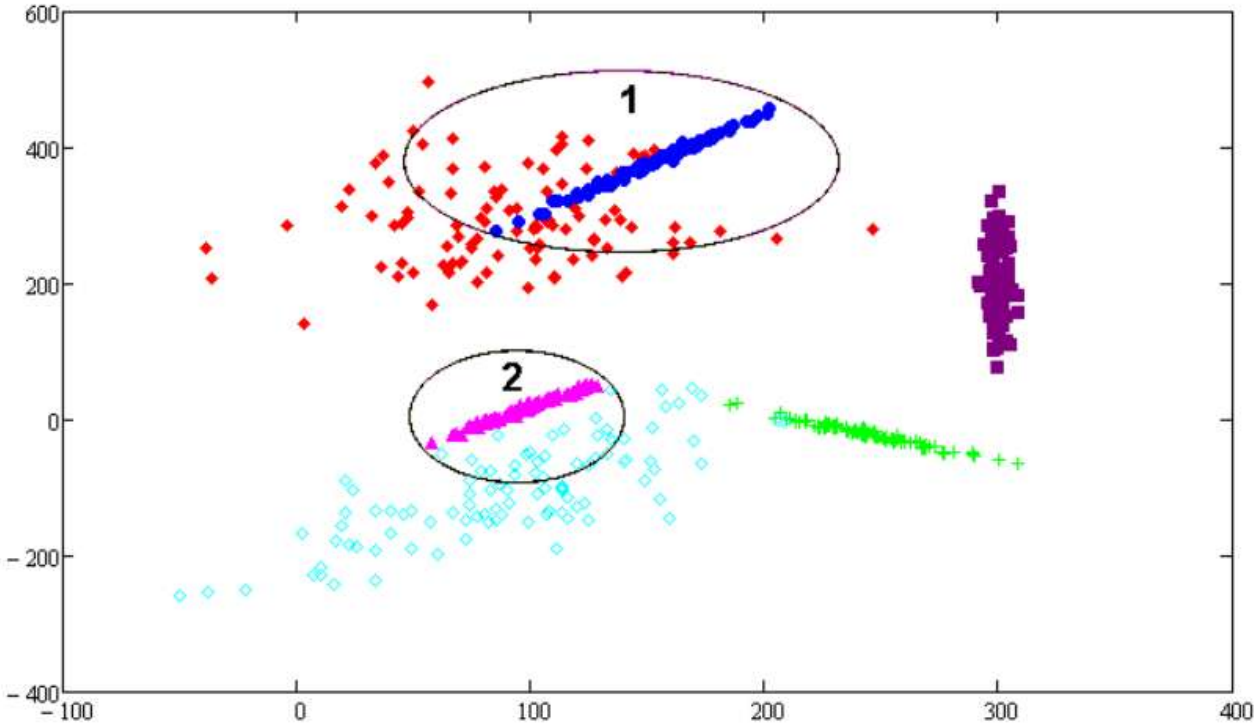
Ponieważ $\Delta LL > \delta$, wracamy do ***Kroku* 2**.

Jeszcze jednym dodatkowym ograniczeniem jest podanie maksymalnej liczby iteracji.

Po wykonaniu 4 iteracji powstał następujący podział:

Obiekty $X^{(1)}, X^{(4)}, X^{(5)}$ należą do klasy 1; a $X^{(2)}, X^{(3)}, X^{(6)}$ do klasy 2.

Przykład 3 (1)

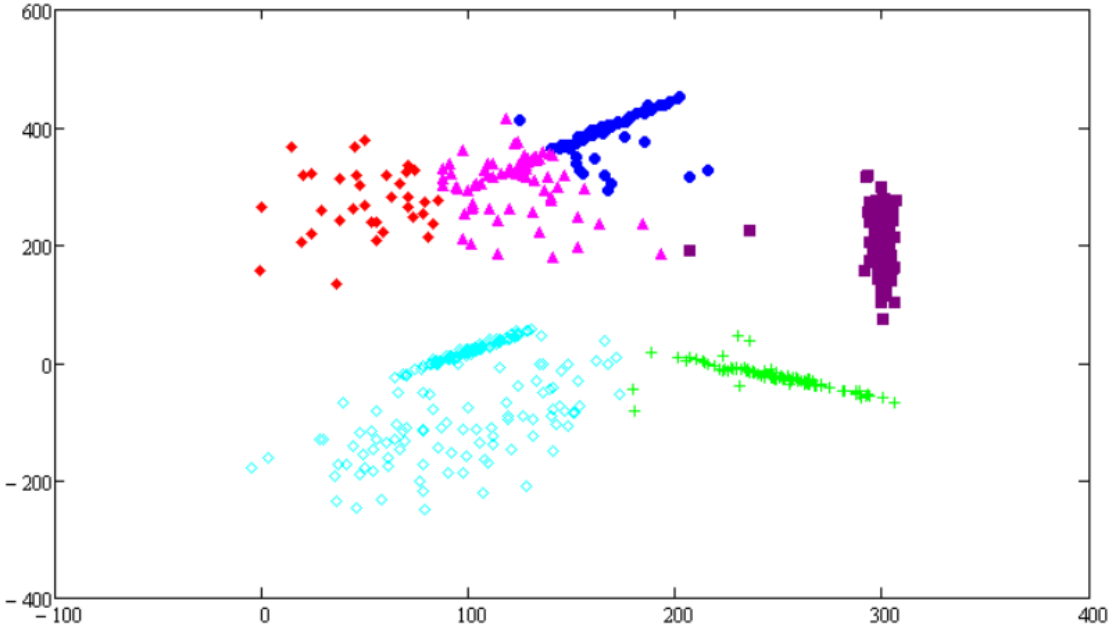


klasa 1
klasa 2
klasa 3
klasa 4
klasa 5
klasa 6

Podstawowy zbiór danych nie jest prosty z punktu widzenia zadania klasyfikacji, ponieważ jest w nim oczywiste pokrycie klas (strefa 1 i strefa 2).

Przykład 3 (2)

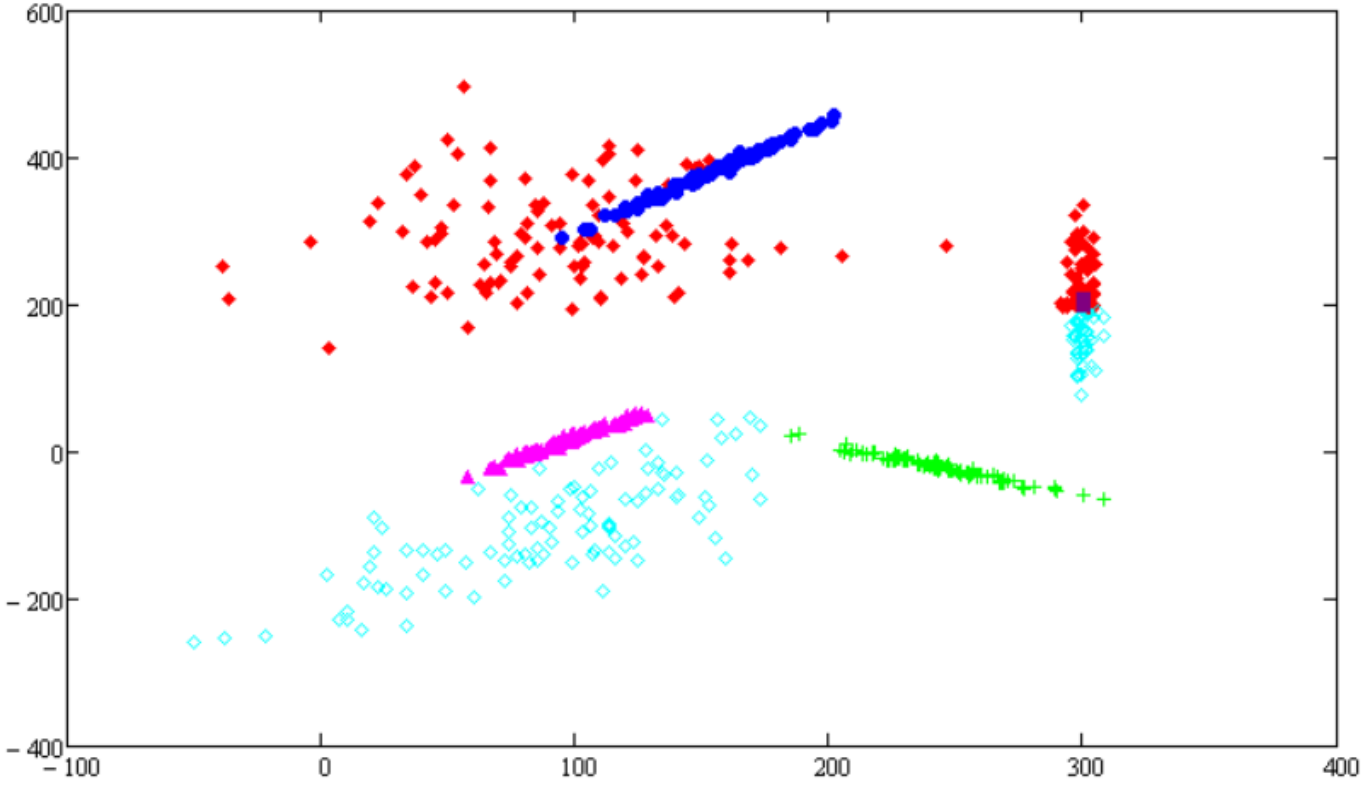
Dla algorytmu *k*-średnich trudności muszą być w miejscu pokrycia klas, co potwierdza się rysunkiem.



W miejscach pokrycia klas możemy obserwować największą liczbę błędów, ale z drugiej strony klasy 3 i 6 zostały rozpoznane bezbłędnie.

Przykład 3 (3)

Algorytm EM bezbłędnie rozpoznał pokrywające się klasy, ale prawie nie rozpoznał klasę 6.



III. Metoda Warda (1)

Metoda Warda zmierza do minimalizacji sumy kwadratów odchyleń wewnątrz skupień. W metodzie tej na każdym etapie spośród wszystkich możliwych do łączenia par skupień wybiera się taka para, która w rezultacie łączenia daje skupienie o minimalnym zróżnicowaniu. Marą takiego zróżnicowania jest wyrażenie *ESS*, zwane też błędem sumy kwadratów:

$$ESS = \sum_{i=1}^{k} \left(x_i - \bar{x} \right)^2 ,$$

gdzie x_i - wartość zmiennej będącej kryterium segmentacji dla i-tego obiektu; k - liczba obiektów w skupieniu.

Metoda Warda (2)

Metoda ta jest traktowana jako efektywna, chociaż zmierza do tworzenia skupień o małej wielkości.

14, 18, 17, 18, 14, 14, 14, 14, 12, 12

Wartości zróżnicowania dla jednego skupienia (wszystkie wyniki) ze środkiem 14,7 jest równa $ESS = 44{,}1$. Z drugiej strony, jeśli utworzymy cztery skupienia:

$S_1 = \{12,12\}, S_2 = \{14,14,14,14,14\}, S_3 = \{17\}, S_4 = \{18,18\}$

to wartość zróżnicowania jest równa:

$$ESS = ESS_1 + ESS_2 + ESS_3 + ESS_4 = 0 + 0 + 0 + 0 = 0.$$

Oznacza to, że utworzenie czterech powyższych grup jest najlepszym grupowaniem.

Przykład 4 (1)

Dane są wyniki badań: 2,5,9,10,15. Wykorzystując metodę Warda, wykonać podział na grupy.

Krok 1. Tworzymy pierwsze skupienia z par obiektów i wybieramy parę, która tworzy skupienie o najmniejszym ESS.

obiekty	1 i 2	1 i 3	1 i 4	1 i 5	2 i 3	2 i 4	2 i 5	3 i 4	3 i 5	4 i 5
ESS	4,5	24,5	32	84,5	8	12,5	50	0,5	18	12,5

Dla trzeciego i czwartego obiektów $ESS = 0{,}5$. Pierwsze skupienie S_1 tworzą te dwa obiekty.

Krok 2. Dla zredukowanego zbioru obiektów i nowo powstałego skupienia tworzy się ponownie wszystkie możliwe skupienia i wylicza się wartość ESS.

obiekty	1 i 2	1 i 5	2 i 5	S_1 i 1	S_1 i 2	S_1 i 5
ESS	4,5	84,5	50	38	14	20,66

Kolejne skupienie tworzą obiekty 1 i 2. Wartość $ESS = 4{,}5$.

Przykład 4 (2)

Krok 4. Ponownie obliczamy wartości ESS dla nowego układu skupień. Otrzymamy następujące wartości:

obiekty	S_1 i 5	S_2 i 5	S_1 i S_2
ESS	20,66	92,66	41

Tym razem dołączamy obiekt piąty do pierwszego skupienia. Powstaje nowe skupienie S_3, zawierające trzy obiekty $S_3 = \{O_3, O_4, O_5\}$.

Krok 5. Jest to krok ostatni. Pozostały tylko dwa obiekty S_3 i S_2, które łączymy w jedno skupienie, obejmujące wszystkie 5 obiektów. Możemy wyróżnić dwie grupy obiektów $\{O_1, O_2\}$ i $\{O_3, O_4, O_5\}$.

IV. Metoda aglomeracji

Algorytm składa się z następujących kroków:

- Buduje się macierz odległości (na przykład, euklidesowych).
- Wybieramy najmniejsze wartości w macierzy odległości (poza główną przekątną) i tworzymy nowe skupienie z obiektów, których ta najmniejsza odległość dotyczy.
- Ponownie wyznaczamy macierz odległości dla nowego, zredukowanego układu obiektów. Odległości utworzonego skupienia musimy obliczyć.
- Wykorzystując nową macierz odległości, znajdujemy kolejną najmniejszą odległość, i tak do końca dopóki nie zostanie 1 skupienie. Utworzone skupienie zawiera już wszystkie obiekty. Kończymy zatem proces grupowania.

Przykład 5 (1)

Wykonać klasyfikację obiektów $O_1, O_2, \ldots, O_6$, objętych badaniem, w rezultacie którego otrzymane są następujące wyniki:

	X_1	X_2	X_3	X_4
O_1	39,8	38	22,2	23,2
O_2	53,7	37,2	18,7	18,5
O_3	47,3	39,8	23,3	22,1
O_4	41,7	37,6	22,8	22,3
O_5	44,7	38,5	24,8	24,4
O_6	47,9	39,8	22,0	23,3

Przykład 5 (2)

Algorytm składa się z następujących kroków:

Krok 1. Buduje się macierz odległości (na przykład, euklidesowych):

$$\begin{array}{c} \\ O_1 \\ O_2 \\ O_3 \\ O_4 \\ O_5 \\ O_6 \end{array} \begin{array}{c} \begin{array}{cccccc} O_1 & O_2 & O_3 & O_4 & O_5 & O_6 \end{array} \\ \begin{pmatrix} 0,0 & 4,08 & 2,35 & 0,75 & 1,78 & 2,31 \\ 4,08 & 0,0 & 3,93 & 3,68 & 4,70 & 3,89 \\ 2,35 & 3,93 & 0,0 & 2,30 & 1,87 & 0,88 \\ 0,75 & 3,68 & 2,30 & 0,0 & 1,75 & 2,43 \\ 1,78 & 4,70 & 1,87 & 1,75 & 0,0 & 2,00 \\ 2,31 & 3,89 & 0,88 & 2,43 & 2,00 & 0,0 \end{pmatrix} \end{array}.$$

Dla analizy grupowania tych obiektów zastosujemy, na przykład, metodę najbliższego sąsiada. W metodzie tej odległość między dwoma skupieniami to najmniejsza odległość spośród wszystkich odległości pomiędzy obiektami.

Przykład 5 (3)

Krok 2. Wybieramy najmniejsze wartości w macierzy odległości (poza główną przekątną) i tworzymy nowe skupienie z obiektów, których ta najmniejsza odległość dotyczy. W naszym przykładzie jest to odległość między obiektem pierwszym a czwartym.

	O_1	O_2	O_3	O_4	O_5	O_6
O_1	0,0	4,08	2,35	0,75	1,78	2,31
O_2	4,08	0,0	3,93	3,68	4,70	3,89
O_3	2,35	3,93	0,0	2,30	1,87	0,88
O_4	0,75	3,68	2,30	0,0	1,75	2,43
O_5	1,78	4,70	1,87	1,75	0,0	2,00
O_6	2,31	3,89	0,88	2,43	2,00	0,0

Łączymy więc obiekty O_1,O_4 w nowe skupienie S_1.

Przykład 5 (4)

Krok 3. Ponownie wyznaczamy macierz odległości dla nowego, zredukowanego układu obiektów. Odległości utworzonego skupienia musimy obliczyć. Przykładowo odległość między skupieniem S_1 i obiektem O_2 wyznaczamy wzorem:

$$d(S_1,O_2)=\min\{d(O_1,O_2),d(O_4,O_2)\}=\min\{4{,}08;\ 3{,}68\}=3{,}68.$$

Wyliczając w podobny sposób pozostałe odległości, otrzymujemy nową macierz odległości postaci:

$$\begin{array}{c} \\ S_1 \\ O_2 \\ O_3 \\ O_5 \\ O_6 \end{array} \begin{array}{c} \begin{array}{ccccc} S_1 & O_2 & O_3 & O_5 & O_6 \end{array} \\ \begin{pmatrix} 0{,}0 & 3{,}68 & 2{,}30 & 1{,}75 & 2{,}31 \\ & 0{,}0 & 3{,}93 & 4{,}70 & 3{,}89 \\ & & 0{,}0 & 1{,}87 & \underline{0{,}88} \\ & & & 0{,}0 & 2{,}00 \\ & & & & 0{,}0 \end{pmatrix} \end{array}$$

Część macierzy została niewypełniona, ponieważ macierz jest symetryczna.

Przykład 5 (5)

Krok 4. Wykorzystując nową macierz odległości, znajdujemy kolejną najmniejszą odległość. Jest to odległość między obiektem O_3 i O_6. Łączymy te obiekty w nowe skupienie S_2 i powtarzamy krok 3.

Krok 5. Zgodnie z metodą najbliższego sąsiada odległość między skupieniami S_1 i S_2 to najmniejsza odległość między obiektami, tworzącymi te skupienia: $S_1 = \{O_1, O_4\}, S_2 = \{O_3, O_6\}$

$$d(S_1, S_2) = \min\{d(O_1, O_3), d(O_1, O_6), d(O_4, O_3), d(O_4, O_6)\} =$$

$$= \min\{2{,}35;\ 2{,}31;\ 2{,}30;\ 2{,}43\} = 2{,}30.$$

$$\begin{array}{c} \\ S_1 \\ S_2 \\ O_2 \\ O_5 \end{array} \begin{array}{c} \begin{array}{cccc} S_1 & S_2 & O_2 & O_5 \end{array} \\ \begin{pmatrix} 0{,}0 & 2{,}30 & 3{,}68 & \underline{1{,}75} \\ & 0{,}0 & 3{,}89 & 1{,}87 \\ & & 0{,}0 & 4{,}70 \\ & & & 0{,}0 \end{pmatrix} \end{array}.$$

Przykład 5 (6)

Krok 6. Najmniejszy element 1,75 pomiędzy skupieniem S_1 i obiektem O_5. Tworzymy nowe skupienie $S_3 = \{S_1, O_5\} = \{O_1, O_4, O_5\}$.

Krok 7. Nowa macierz odległości:

$$\begin{array}{c} \\ S_3 \\ S_2 \\ O \end{array} \begin{array}{c} \begin{array}{ccc} S_3 & S_2 & O_2 \end{array} \\ \begin{pmatrix} 0,0 & 1,87 & 3,68 \\ & 0,0 & 3,89 \\ & & 0,0 \end{pmatrix} \end{array}.$$

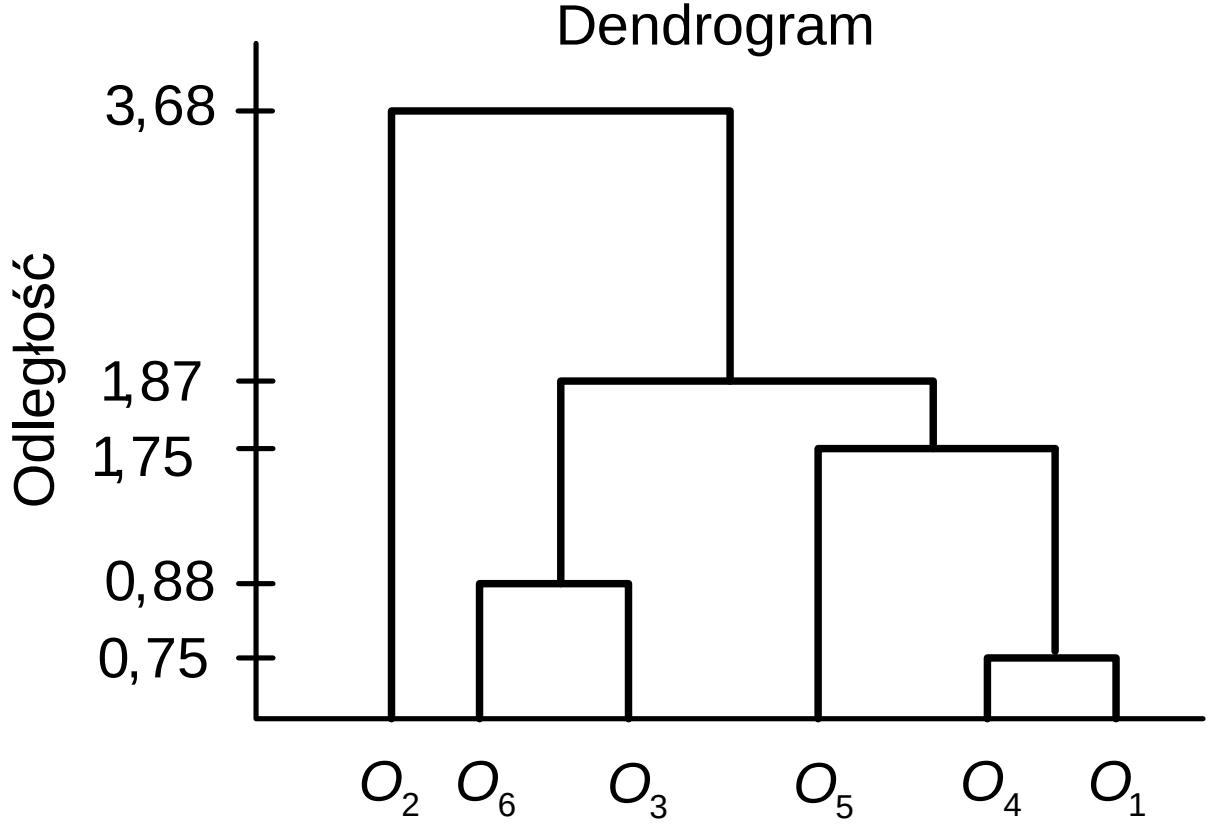
Krok 8. Łączymy skupienia S_2 i S_3, zawierające 5 obiektów $S_{23} = \{S_2, S_3\} = \{O_1, O_3, O_4, O_5, O_6\}$.

Krok 9. Ostatnia macierz odległości:

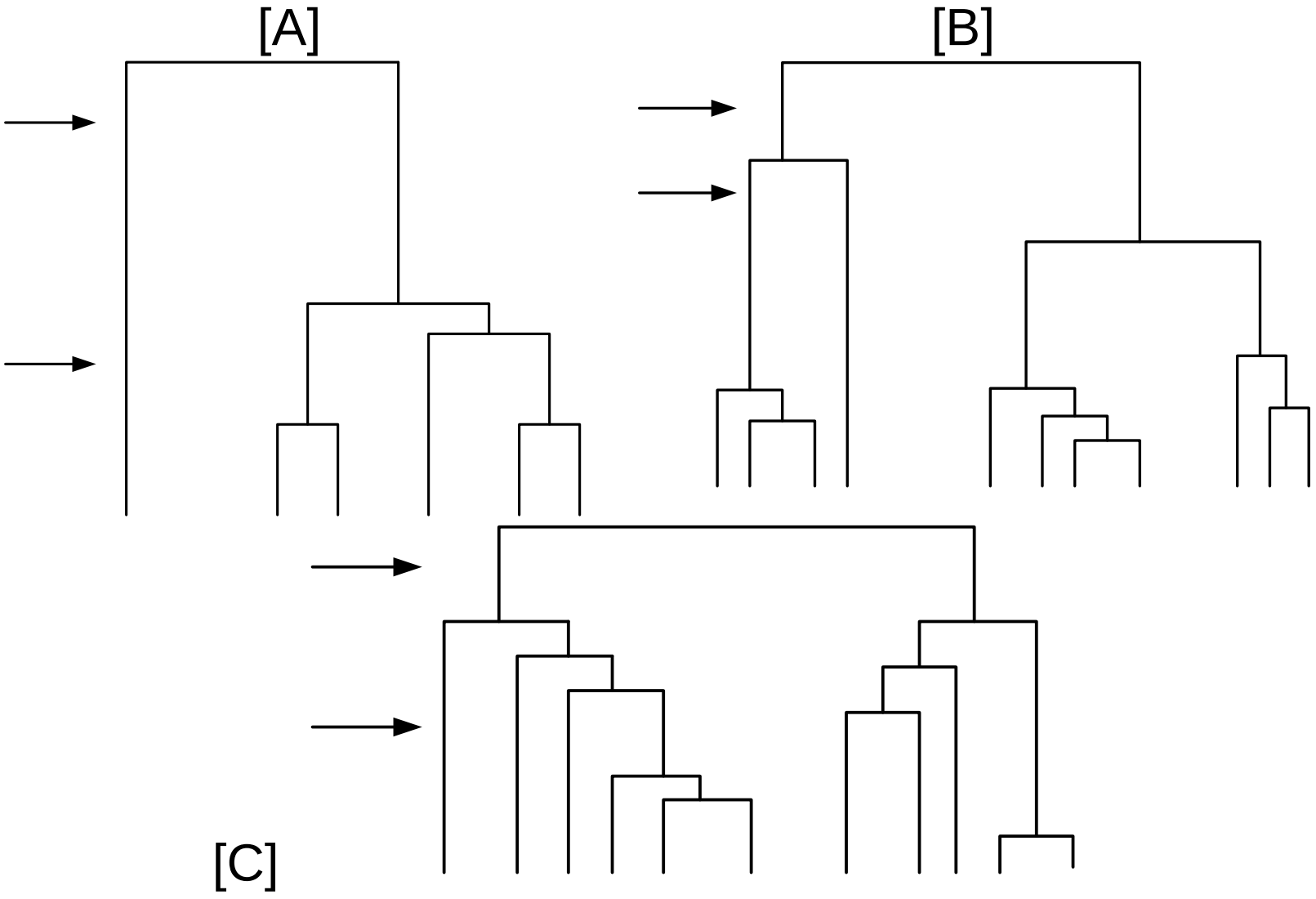
$$\begin{array}{c} \\ S_{23} \\ O_2 \end{array} \begin{array}{c} \begin{array}{cc} S_{23} & O_2 \end{array} \\ \begin{pmatrix} 0,0 & 3,68 \\ & 0,0 \end{pmatrix} \end{array}.$$

Dendrogram

Otrzymane rezultaty możemy przedstawić za pomocą ***dendrogramu*** (wykresu drzewkowego) ilustrującego hierarchiczną strukturę zbioru obiektów ze względu na zmniejszające się podobieństwo między nimi.



Grupowanie zmiennych



Wskazówki ustalenia liczby skupień

- należy analizować dendrogram pod względem różnić odległości między kolejnymi węzłami. Duża wartość różnic oznacza, że skupienia są odległe (odległość między kolejnymi węzłami jest duża) i w tym miejscu dokonujemy podziału. Przykładowo, kierując się tą sugestią, wybieramy dolną strzałkę na wykresie [A];

- wykorzystać różne mierniki, takie jak: miernik Mojena, miernik Grabińskiego itp.

Reguła Mojena

Punktem odcięcia jest odległość wiązania, dla której spełniona jest nierówność:

$$d_i > \bar{d} + k \cdot s_d, \text{ gdzie}$$

$d_0, d_1, \ldots, d_{n-1}$ są odległościami wiązania dla etapu $n, n-1, \ldots, 1$;

$\bar{d}, s_d$ - średnia oraz odchylenie standardowe d_i;

k - pewna stała.

Mojen zasugerował, że wartość $k \in [2{,}75;\ 3{,}50]$ daje zadowalające wyniki.

Z kolej Miligan i Cooper (1985) sugerują, że wartość $k = 1{,}25$ daje najlepsze wyniki.

Reguła Grabińskiego

Największa wartość $q_i = \frac{d_i}{d_{i-1}}$ wskazuje miejsce podziału.

Często zdarza się, że q_i przyjmuje wartość najwyższą dla pierwszych odległości – jest to podstawowa wada tej reguły.

Przykład 6 (1)

	X_1	X_2	X_3	X_4
O_1	39,8	38	19,2	23,2
O_2	47,6	39,8	22,4	22,1
O_3	41,7	37,6	21,0	22,3
O_4	40,7	38,5	24,8	23,4
O_5	47,9	39,8	22,0	23,3
O_6	39,7	38,0	20,0	22,3
O_7	48,0	39,9	23,3	22,1
O_8	39,5	37,9	20,2	23,3
O_9	47,7	39,7	22,7	23,0
O_{10}	47,8	39,8	22,0	23,3
O_{11}	47,9	39,9	22,4	22,7
O_{12}	39,4	37,6	19,8	22,5
O_{13}	39,6	38,1	18,8	23,2
O_{14}	48,1	39,7	23,0	22,3

	X_1	X_2	X_3	X_4
O_1	-1,02	-0,90	-1,34	0,82
O_2	0,89	0,94	0,49	-1,36
O_3	-0,55	-1,30	-0,31	-0,96
O_4	-0,80	-0,39	1,87	1,22
O_5	0,96	0,94	0,26	1,02
O_6	-1,04	-0,90	-0,88	-0,96
O_7	0,99	1,04	1,01	-1,36
O_8	-1,09	-1,00	-0,77	1,02
O_9	0,92	0,84	0,66	0,43
O_{10}	0,94	0,94	0,26	1,02
O_{11}	0,94	1,04	0,49	-0,17
O_{12}	-1,11	-1,30	-1,00	-0,56
O_{13}	-1,06	-0,79	-1,57	0,82
O_{14}	1,01	-0,84	0,83	-0,96

Przykład 6 (2)

Przebieg aglomeracji:

	d														
1	0,10	O_5	O_{10}												
2	0,42	O_7	O_{14}												
3	0,46	O_1	O_{13}												
4	0,51	O_9	O_{11}												
5	0,57	O_6	O_{12}												
6	0,68	O_9	O_{11}	O_2											
7	0,72	O_9	O_{11}	O_2	O_5	O_{10}									
8	0,77	O_9	O_{11}	O_2	O_5	O_{10}	O_7	O_{14}							
9	0,95	O_6	O_{12}	O_8											
10	1,05	O_6	O_{12}	O_8	O_1	O_{13}									
11	2,27	O_6	O_{12}	O_8	O_1	O_{13}	O_3								
12	4,17	O_6	O_{12}	O_8	O_1	O_{13}	O_3	O_4							
13	6,45	O_6	O_{12}	O_8	O_1	O_{13}	O_3	O_4	O_9	O_{11}	O_2	O_5	O_{10}	O_7	O_{14}

Przykład 6 (3)

Reguła Mojena:

Wyznaczamy średnią $\bar{d}$ oraz odchylenie standardowe s_d dla zmiennej Odległość:

$$\bar{d} = 1{,}47 \qquad s_d = 1{,}84$$

$$d_i > \bar{d} + 1{,}25 \cdot s_d = 1{,}47 + 1{,}25 \cdot 1{,}84 = 3{,}77,$$

czyli sugerowany jest podział na kroku dwunastym, gdzie $d_{12} = 4{,}17 > 3{,}77$.

Podział na grupy na kroku 12:

$$\{O_9, O_{11}, O_2, O_5, O_{10}, O_7, O_{14}\},$$

$$\{O_6, O_{12}, O_8, O_1, O_{13}, O_3, O_4\}$$

Przykład 6 (4)

	d_i	d_{i-1}	$q_i = d_i / d_{i-1}$
1	0,10	-	
2	0,42	0,10	4,20
3	0,46	0,42	1,08
4	0,51	0,46	1,11
5	0,57	0,51	1,13
6	0,68	0,57	1,18
7	0,72	0,68	1,07
8	0,77	0,72	1,06
9	0,95	0,77	1,22
10	1,05	0,95	1,11
11	2,27	1,05	**2,16**
12	4,17	2,27	1,84
13	6,45	4,17	1,54

Reguła Grabińskiego: Jeżeli nie brać pod uwagę kilka pierwszych wartości q_i (ponieważ podstawową wadą tej reguły jest to, że często zdarza się, że q_i przyjmuje wartość najwyższą dla pierwszych odległości), to można zauważyć, że iloraz $q_i = \frac{d_i}{d_{i-1}}$ przyjmuje wartość największą na kroku 11:

$$\{O_9, O_{11}, O_2, O_5, O_{10}, O_7, O_{14}\},$$

$$\{O_6, O_{12}, O_8, O_1, O_{13}, O_3\},$$

$$\{O_4\}$$

Przykład 6 (5)

	d_i	d_{i-1}	$d_i - d_{i-1}$
1	0,10	-	
2	0,42	0,10	0,32
3	0,46	0,42	0,03
4	0,51	0,46	0,05
5	0,57	0,51	0,06
6	0,68	0,57	0,10
7	0,72	0,68	0,05
8	0,77	0,72	0,05
9	0,95	0,77	0,17
10	1,05	0,95	0,10
11	2,27	1,05	1,21
12	4,17	2,27	1,9
13	6,45	4,17	2,27

Reguła maksimum: maksymalna wartość $d_i - d_{i-1}$ jest równa 2,27, krok 13, ale ponieważ na kroku 13 wszystkie obiekty zostały połączone w jedną grupę, to bierzemy krok podział na kroku 12:

$$\{O_9, O_{11}, O_2, O_5, O_{10}, O_7, O_{14}\},$$

$$\{O_6, O_{12}, O_8, O_1, O_{13}, O_3, O_4\}$$

Przykład 6 (6)

Obiekt O_4, który dołączony został do jednego ze skupień w ostatni moment wskazuje na wartość nietypową lub nową grupę. Badanie należy więc powtórzyć z o wiele większą grupą obiektów.