Procesy i systemy Business Intelligence

Grupowanie

Wprowadzenie

- Celem procesu grupowania jest podział zbioru obiektów, fizycznych lub abstrakcyjnych, na klasy obiektów o podobnych cechach, nazywane *klastrami* lub *skupieniami*
- Klaster

  1. Zbiór obiektów, które są "podobne"

  2. Zbiór obiektów, takich, że odległość pomiędzy dwoma dowolnymi obiektami należącymi do klastra jest mniejsza aniżeli odległość pomiędzy dowolnym obiektem należącym do klastra i dowolnym obiektem nie należącym do tego klastra

  3. Spójny obszar przestrzeni wielowymiarowej, charakteryzujący się dużą gęstością występowania obiektów

Przykłady

- **Zbiór dokumentów** – zbiór punktów w przestrzeni wielowymiarowej, w której pojedynczy wymiar odpowiada jednemu słowu z określonego słownika. Współrzędne dokumentu w przestrzeni są zdefiniowane względną częstością występowania słów ze słownika. Klastry dokumentów odpowiadają grupom dokumentów dotyczących podobnej tematyki
- **Zbiór sekwencji stron WWW** – pojedyncza sekwencja opisuje sekwencję dostępów do stron WWW danego serwera realizowaną w ramach jednej sesji przez użytkownika. Klastry sekwencji odpowiadają grupom użytkowników danego serwera, którzy realizowali dostęp do tego serwera w podobny sposób

Niepodobieństwo obiektów

- Niepodobieństwo (podobieństwo) obiektów opisujemy za pomocą macierzy niepodobieństwa (podobieństwa)
- Danych jest N obiektów, z których każdy jest opisany wartościami p atrybutów A_1, …, A_p (nazywanych zmiennymi)
- Macierz niepodobieństwa obiektów D, typu N × N opisuje niepodobieństwo pomiędzy każda parą obiektów:

$$D = \begin{bmatrix} 0 & D(x_1, x_2) & \dots & D(x_1, x_N) \\ D(x_2, x_1) & 0 & \dots & D(x_2, x_N) \\ \vdots & \vdots & & \vdots \\ D(x_N, x_1) & D(x_N, x_2) & \dots & D(x_N, x_N) \end{bmatrix}$$

gdzie $D(x_i, x_j)$ oznacza niepodobieństwo obiektów x_i i x_j

Miary odległości (1)

- Dyskusja dotycząca podobieństwa, lub odległości, dwóch obiektów wymaga przyjęcia miary odległości pomiędzy dwoma obiektami x i y reprezentowanymi przez punkty w przestrzeni wielowymiarowej:
- Klasyczne aksjomaty dla miary odległości będącej metryką:
  - $D(x, y) \geq 0$
  - $x=y \Leftrightarrow D(x, y) = 0$
  - $D(x, y) = D(y, x)$
  - $D(x, y) \leq D(x, z) + D(z, y)$ (nierówność trójkąta)

Atrybuty ciągłe (1)

- Dana jest k-wymiarowa przestrzeń euklidesowa, odległość pomiędzy dwoma obiektami $x=[x_1, x_2, ..., x_p]$ i $y=[y_1, y_2, ..., y_p]$ można zdefiniować następująco:
- odległość euklidesowa: („norma L_2 ") $\sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$
- odległość Manhattan : („norma L_1 ") $\sum_{i=1}^{p} | x_i - y_i |$
- odległość max z wymiarów: („norma L_∞") $\max_{i=1}^{p} | x_i - y_i |$

Atrybuty ciągłe (2)

- Odległość Minkowskiego: $$\left(\sum_{i=1}^{p}\left(|x_i - y_i|\right)^q\right)^{1/q}$$

- Problem skalowalności i zmienności atrybutów:
  - Normalizacja wartości atrybutu A (atrybut ma wartość a):

$$a' = \frac{a - min}{\max - min}$$

  - Standaryzacja wartości atrybutu A (atrybut ma wartość a):

$$a' = \frac{a - \mu(A)}{\sigma(A)}$$

gdzie μ(A) oznacza średnią wartość atrybutu A, natomiast σ(A) oznacza odchylenie standardowe

Miary odległości (2)

- W przypadku, gdy obiekty nie poddają się transformacji do przestrzeni euklidesowej, proces grupowania wymaga zdefiniowania innych miar odległości (podobieństwa): sekwencja dostępów do stron WWW, sekwencje DNA, sekwencje zbiorów, zbiory atrybutów kategorycznych, dokumenty tekstowe, XML, grafy, itp.

Atrybuty binarne (1)

- W jaki sposób obliczyć podobieństwo (lub niepodobieństwo) pomiędzy dwoma obiektami opisanymi zmiennymi binarnymi:
- Podejście: konstruujemy macierz niepodobieństwa

q – liczba zmiennych przyjmujących wartość 1 dla obu obiektów
r – ... 1 dla obiektu i, i wartość 0 dla j
s – ... 0 dla obiektu i, i wartość 1 dla j
t - ... 0 dla obu obiektów

		obiekt j		
		1	**0**	**Sum**
obiekt i	**1**	q	r	q+r
	0	s	t	s+t
	Sum	q+s	r+t	p

Atrybuty binarne (2)

- Zmienne binarne symetryczne: zmienną binarną nazywamy symetryczną jeżeli obie wartości tej zmiennej posiadają tą samą wagę (np. płeć). Niepodobieństwo pomiędzy obiektami i i j jest zdefiniowane następująco:

$$d(i,j) = \frac{r+s}{q+r+s+t}$$

- Zmienne binarne asymetryczne : zmienną binarną nazywamy asymetryczną jeżeli obie wartości tej zmiennej posiadają różne wagi (np. wynik badania EKG). Niepodobieństwo pomiędzy obiektami i i j jest zdefiniowane następująco

$$d(i,j) = \frac{r+s}{q+r+s}$$

Atrybuty binarne (3)

imię	pleć	gorączka	katar	test1	test2	test3	test4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	Y	N	N	N	N
…	…	…	…	…	…	…	…

- Dana jest tablica zawierająca informacje o pacjentach:

$$d_{asym}(jack, mary) = \frac{0+1}{2+0+1} = 0.33 \qquad d_{sym}(jack, mary) = \frac{0+1}{2+0+1+3} = 0.17$$

$$d_{asym}(jack, jim) = \frac{1+1}{1+1+1} = 0.67 \qquad d_{sym}(jack, jim) = \frac{1+1}{1+1+1+3} = 0.33$$

$$d_{asym}(jim, mary) = \frac{1+2}{1+1+2} = 0.75 \qquad d_{sym}(jim, mary) = \frac{1+2}{1+1+2+2} = 0.5$$

Atrybuty kategoryczne (1)

- Zmienna kategoryczna jest generalizacją zmiennej binarnej: może przyjmować więcej niż dwie wartości (np. dochód: wysoki, średni, niski)
- Zmienne kategoryczne:
  - nominalne
  - porządkowe
- Niepodobieństwo (podobieństwo) pomiędzy obiektami i, j, opisanymi zmiennymi kategorycznymi nominalnymi, można zdefiniować następująco:

$$d(i, j) = \frac{p - m}{p} \qquad sim(i, j) = \frac{p - n}{p} = sim(i, j) = \frac{m}{p}$$

gdzie p oznacza łączną liczbę zmiennych, m oznacza liczbę zmiennych, których wartość jest identyczna dla obu obiektów, n oznacza liczbę zmiennych, których wartość jest różna dla obu obiektów.

Obiekty opisane atrybutami różnych typów

- Najpopularniejszym podejściem do problemu określenia niepodobieństwa obiektów opisanych atrybutami różnych typów jest podejście oparte na agregacji niepodobieństw poszczególnych typów atrybutów opisujących obiekty i uzyskaniu jednej zagregowanej miary niepodobieństwa, będącej średnia ważona poszczególnych miar niepodobieństwa pojedynczych atrybutów
- Ogólna postać takiej zagregowanej miary niepodobieństw obiektów jest najczęściej definiowana w następujący sposób:

$$D(x_i, x_j) = \sum_{k=1}^{p} w_k \cdot d_k(x_{ik}, x_{jk}) \qquad \sum_{k=1}^{p} w_k = 1$$

gdzie w_k oznacza wagę k-tego atrybutu A_k, a $d_k(x_{ik}, x_{jk})$ niepodobieństwo wartości k-tego atrybutu obiektów x_i i x_j

Cosinusowa miara podobieństwa

- Miara odległości:

$$D(\vec{x}, \vec{y}) = \frac{\vec{x} \circ \vec{y}}{\|\vec{x}\| \cdot \|\vec{y}\|} \text{ lub}$$

$$D(\vec{x}, \vec{y}) = 1 - \frac{\vec{x} \circ \vec{y}}{\|\vec{x}\| \cdot \|\vec{y}\|}$$

- Załóżmy x=[a1, a2, a3, a4] i y=[2a1, 2a2, 2a3, 2a4] (strona y jest kopią x), wówczas D(x, y) = 0
- Inna miara odległości stron WWW to miara Tanimoto:

$$\text{sim}(\vec{x}, \vec{y}) = \frac{\vec{x} \circ \vec{y}}{\|\vec{x}\|^2 \cdot \|\vec{y}\|^2 - \vec{x} \circ \vec{y}}$$

Miara podobieństwa sekwencji

- Sekwencje DNA, sekwencje dostępu do stron WWW: definicja odległości (podobieństwa) sekwencji symboli, powinna uwzględniać fakt, że sekwencje mogą mieć różną długość oraz różne symbole na tych samych pozycjach, np.:

  x= abcde y= bcdxye

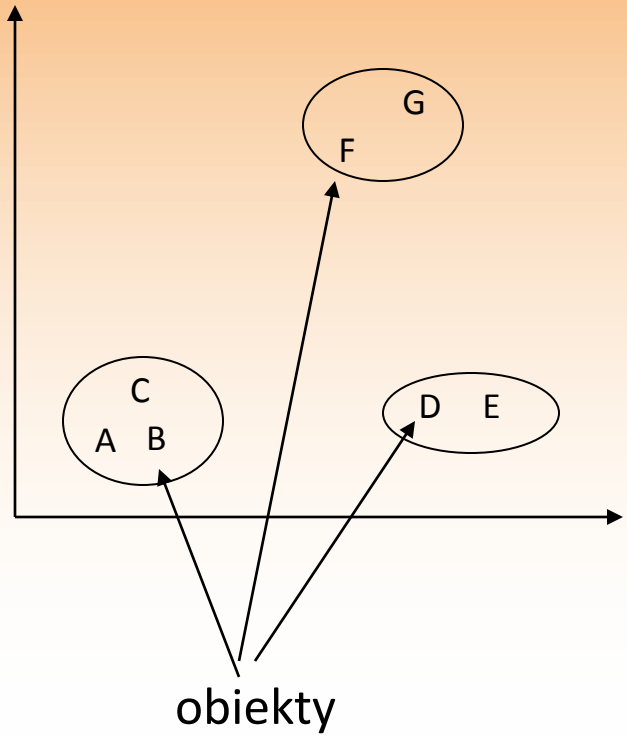
- Miara odległości: $\mathbf{D(x, y) = |x| + |y| - 2|LCS(x, y)|}$

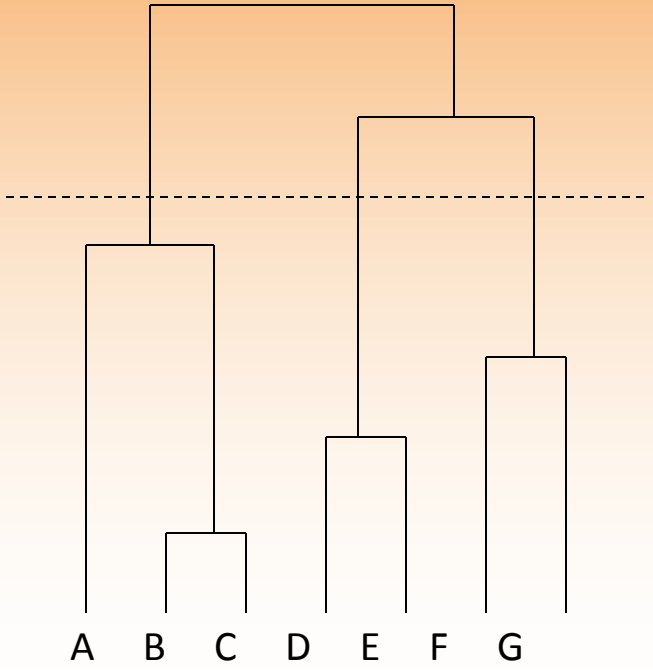
  gdzie LCS oznacza najdłuższą wspólną podsekwencję (ang. longest common subsequence) (LCS(x,y) = bcde). Stąd, D(x, y) = 3

Metody hierarchiczne - wprowadzenie

- **Metody hierarchiczne:** generują zagnieżdżoną sekwencję podziałów zbiorów obiektów w procesie grupowania
- Metoda grupowania hierarchicznego polega na sekwencyjnym grupowaniu obiektów - drzewo klastrów (tzw. dendrogram).
  - Podejście podziałowe (top-down): początkowo, wszystkie obiekty przypisujemy do jednego klastra; następnie, w kolejnych iteracjach, klaster jest dzielony na mniejsze klastry, które, z kolei, dzielone są na kolejne mniejsze klastry
  - Podejście aglomeracyjne (bottom-up): początkowo, każdy obiekt stanowi osobny klaster, następnie, w kolejnych iteracjach, klastry są łączone w większe klastry

Metody hierarchiczne - wprowadzenie





dendrogram

Miary odległości (1)

- W obu podejściach, aglomeracyjnym i podziałowym, liczba klastrów jest ustalona z góry przez użytkownika i stanowi warunek stopu procesu grupowania
- 4 podstawowe (najczęściej stosowane) miary odległości pomiędzy klastrami są zdefiniowane następująco, gdzie

  $|p - p'|$ oznacza odległość pomiędzy dwoma obiektami (lub punktami) p i p', m_i oznacza średnią wartość klastra C_i, i n_i oznacza liczbę obiektów należących do klastra C_i

Miary odległości (2)

minimalna odległość:

$$d_{\min}(C_i, C_j) = \min_{p \in C_i, p' \in C_j} \|p - p'\|$$

maksymalna odległość:

$$d_{\max}(C_i, C_j) = \max_{p \in C_i, p' \in C_j} \|p - p'\|$$

odległość średnich:

$$d_{mean}(C_i, C_j) = \|m_i - m_j\|$$

średnia odległość:

$$d_{ave}(C_i, C_j) = 1/(n_i n_j) \sum_{p \in C_i} \sum_{p' \in C_j} \|p - p'\|$$

Ogólny hierarchiczny aglomeracyjny algorytm grupowania

- Wejście: baza danych D obiektów (n - obiektów)
- Wyjście: dendrogram reprezentujący grupowanie obiektów

1. umieść każdy obiekt w osobnym klastrze;
2. skonstruuj macierz odległości pomiędzy klastrami;
3. dla zadanej wartości niepodobieństwa d_k (d_k może się zmieniać w kolejnych iteracjach)
4. **Repeat**
5. utwórz graf klastrów, w którym każda para klastrów, której wzajemna odległość jest mniejsza niż d_k, jest połączona krawędzią;
6. **until** wszystkie klastry utworzą graf spójny;
7. **return** dendrogram

Hierarchiczny aglomeracyjny algorytm grupowania (1)

Wejście: baza danych D n obiektów.

Wyjście: dendrogram reprezentujący sekwencję grupowania obiektów

1: umieść każdy obiekt w osobnym klastrze;

2: skonstruuj macierz odległości międzyklastrowej dla wszystkich par klastrów;

3: korzystając z macierzy odległości międzyklastrowych, znajdź najbliższą parę klastrów i połącz znalezione klastry, tworząc nowy klaster;

4: uaktualnij macierz odległości międzyklastrowych po operacji połączenia;

Hierarchiczny aglomeracyjny algorytm grupowania (2)

5: **if** wszystkie obiekty należą do jednego klastra **then**
6: zakończ procedurę grupowania;
7: **else**
8: przejdź do kroku 3;
9: **end if**
10: **return** dendrogram reprezentujący sekwencje grupowania obiektów;

Algorytm k-średnich

Wejście: liczba klastrów k, baza danych n obiektów

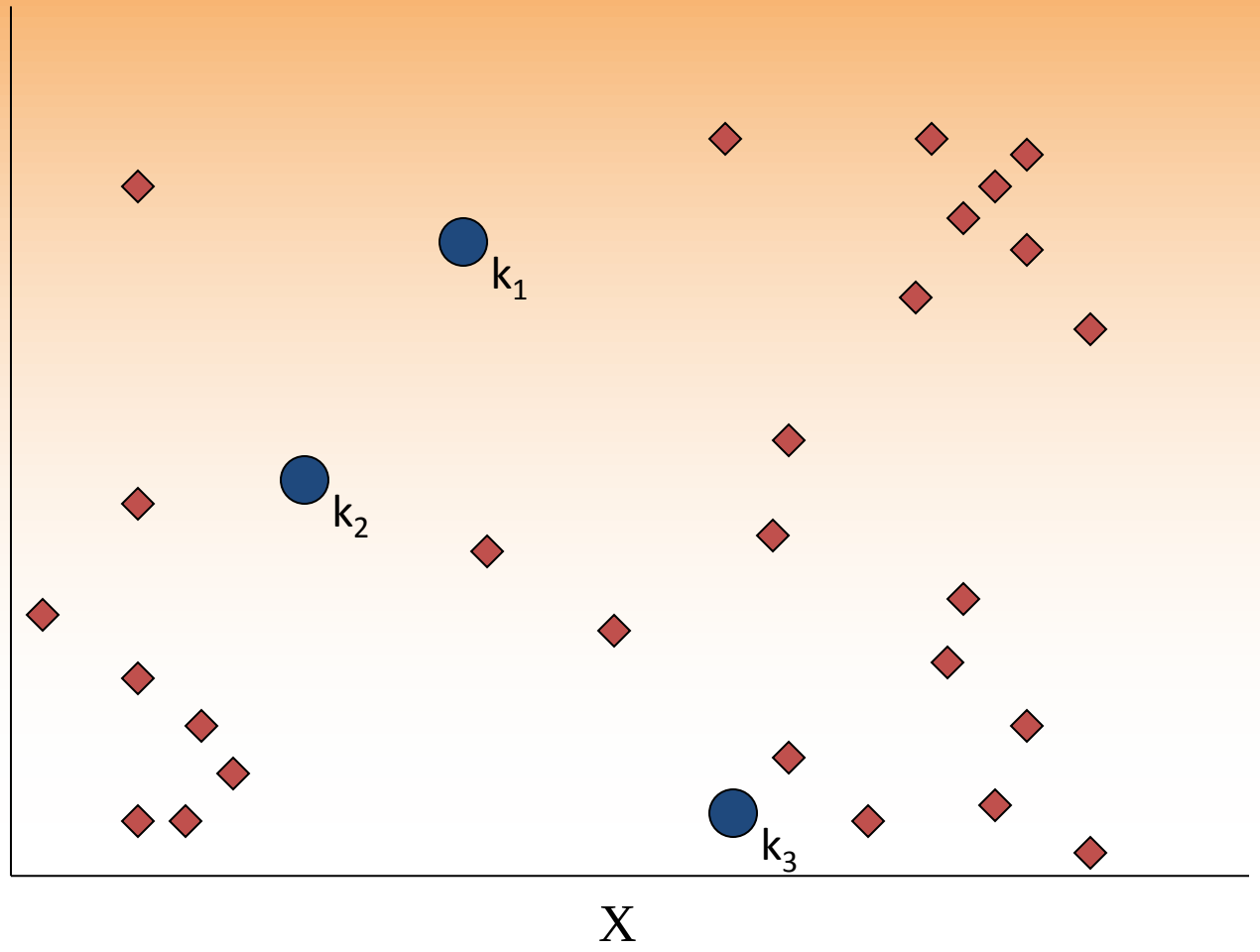
Wyjście: zbiór k klastrów minimalizujący kryterium błędu średniokwadratowego

1. wybierz losowo k obiektów jako początkowe środki k klastrów;
2. **while** występują zmiany przydziału obiektów do klastrów **do**
3. dla każdego obiektu $p_i \in D$ przydziel obiekt p_i do tego klastra C_i, dla którego odległość obiektu p_i od środka klastra C_i jest najmniejsza;
4. uaktualnij środki klastrów – środkiem klastra jest wartość średniej danego klastra;
5. **end while**;

Przykład (1) – krok 1

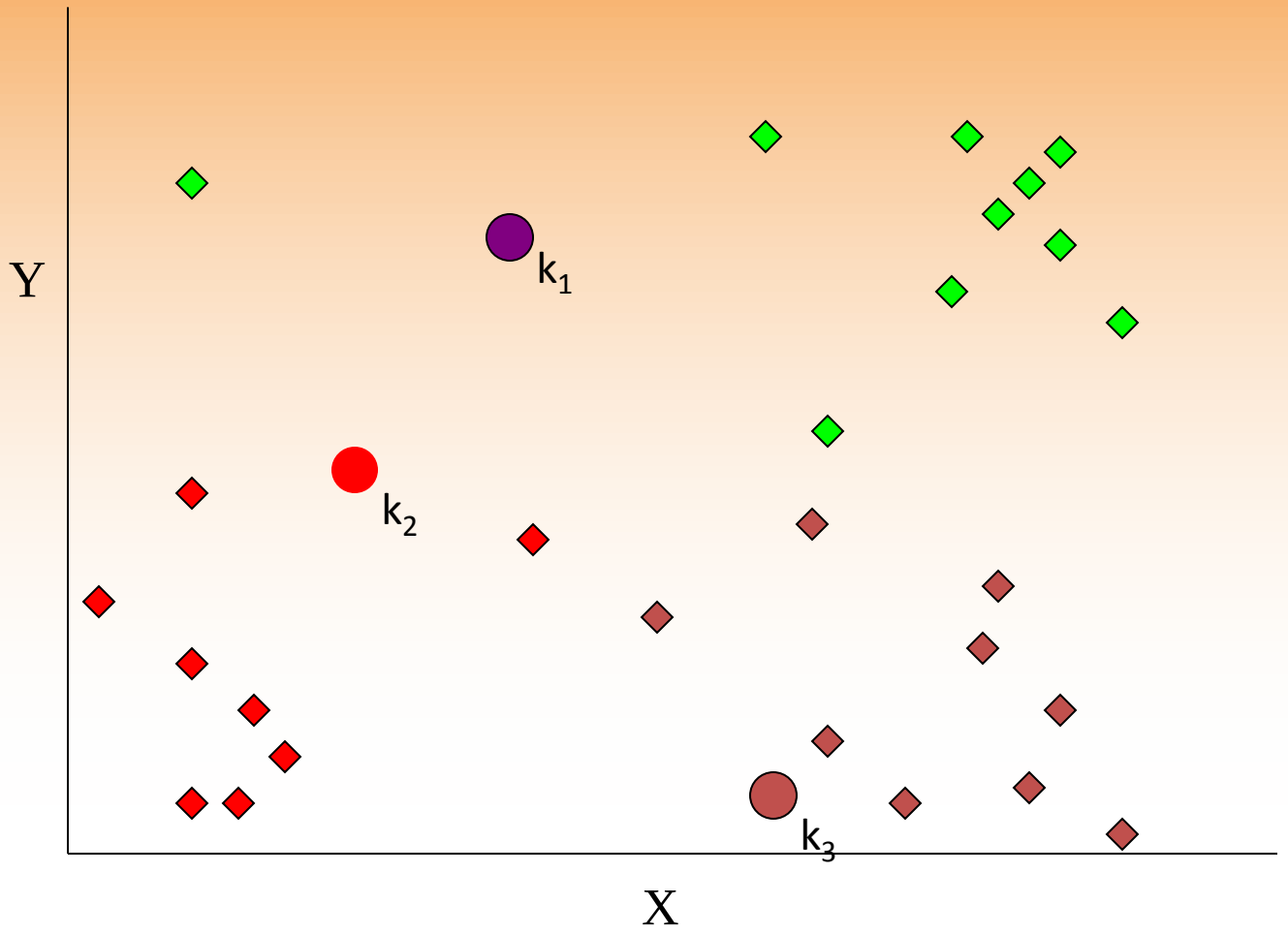
Założenie:
k = 3

wybierz 3 początkowe środki klastrów (losowo)



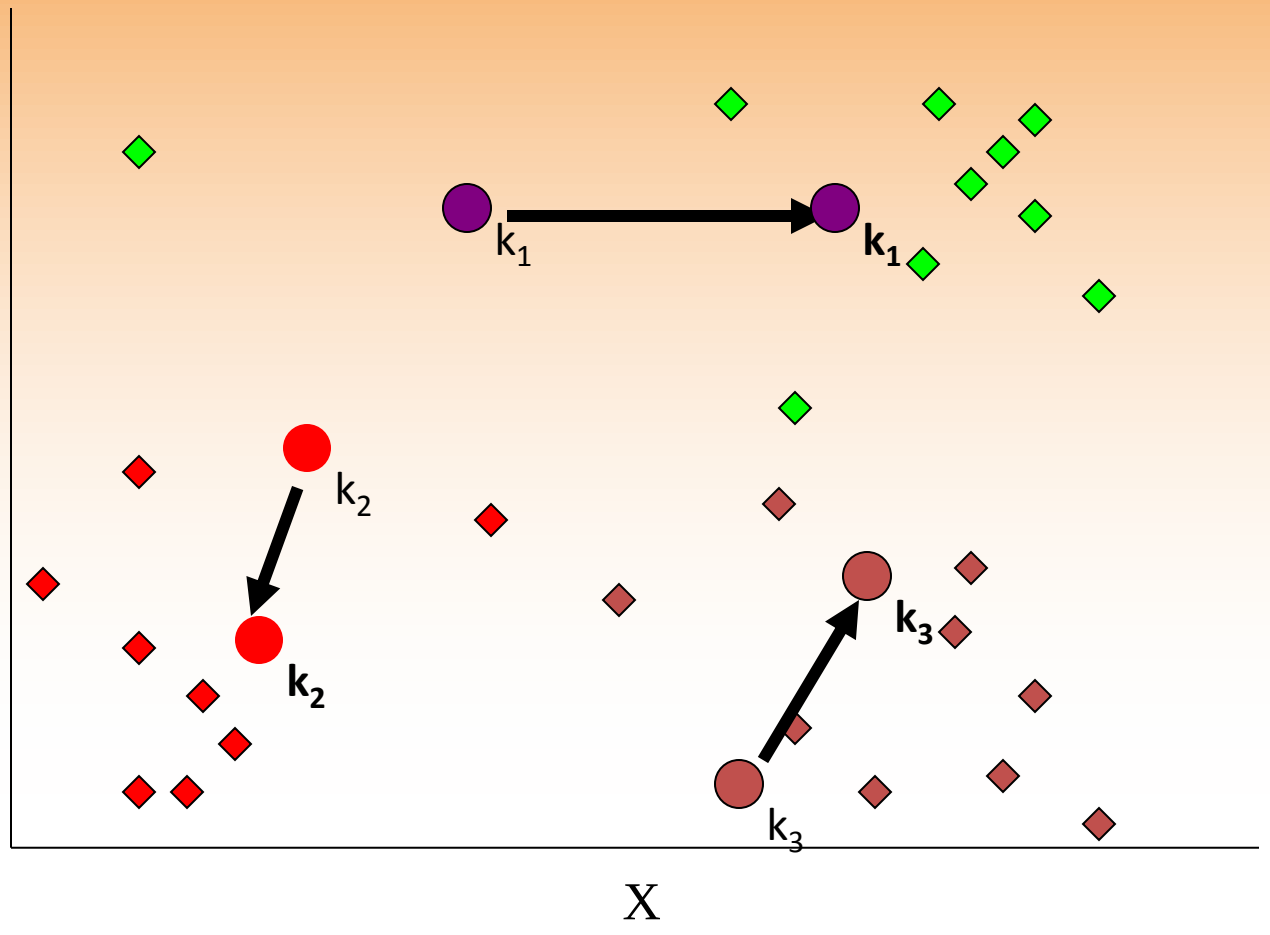
Przykład (2) – krok 2

Przydziel każdy obiekt do klastra w oparciu o odległość obiektu od środka klastra



Przykład (3) – krok 3

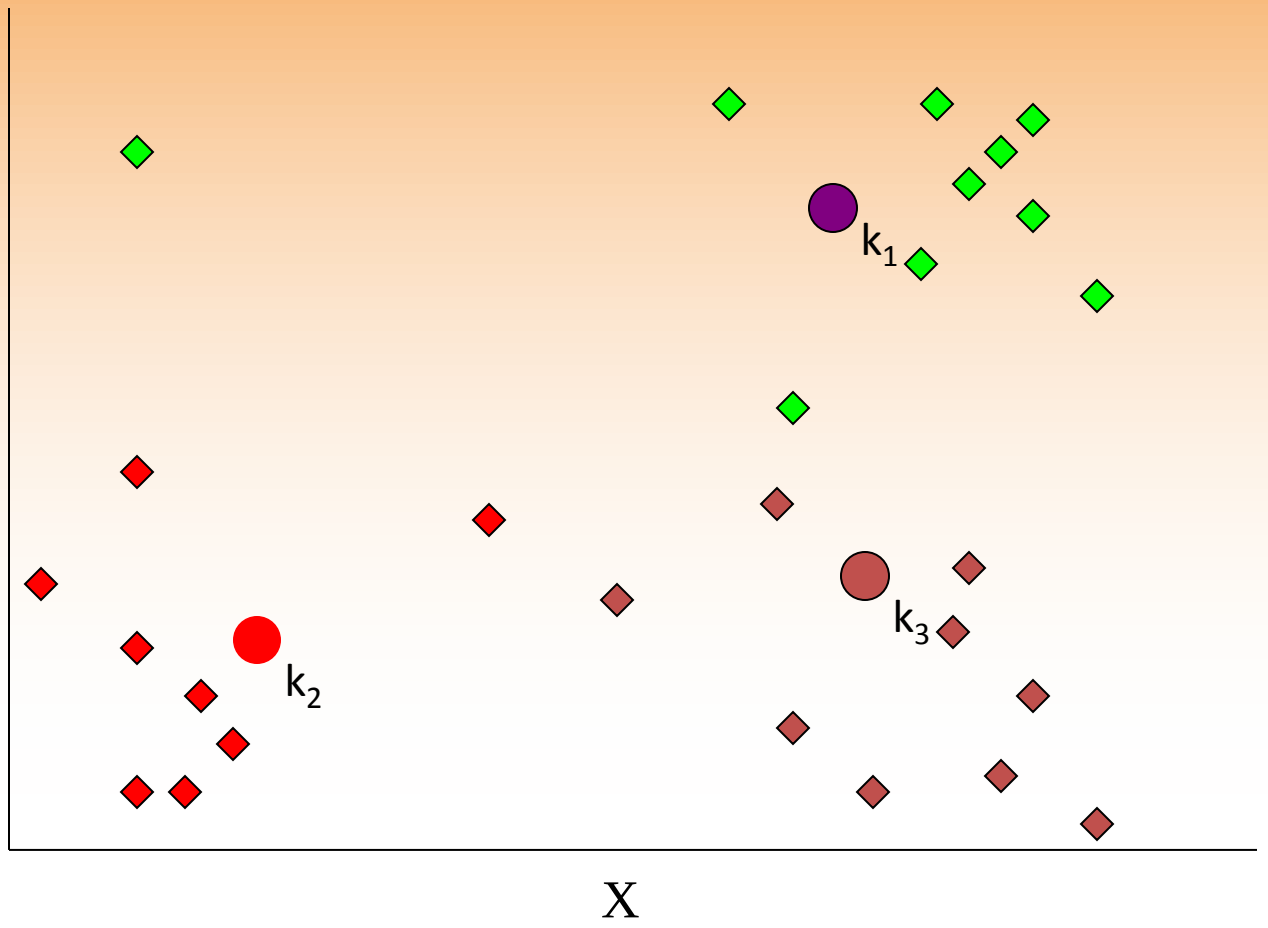
Uaktualnij środki (średnie) wszystkich klastrów



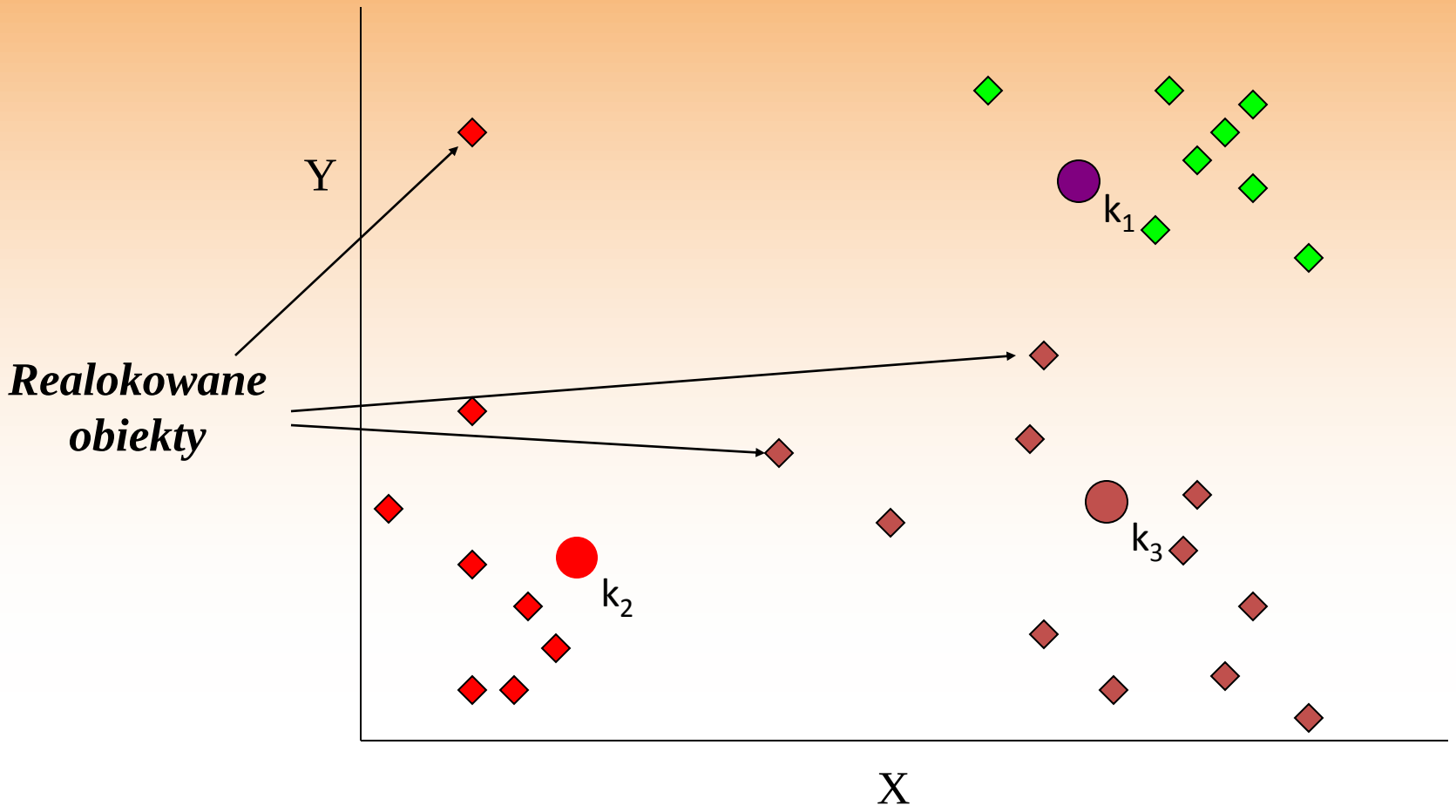
Przykład (4) – krok 4

Realokuj obiekty do najbliższych klastrów

Q: Jakie obiekty zostaną realokowane?

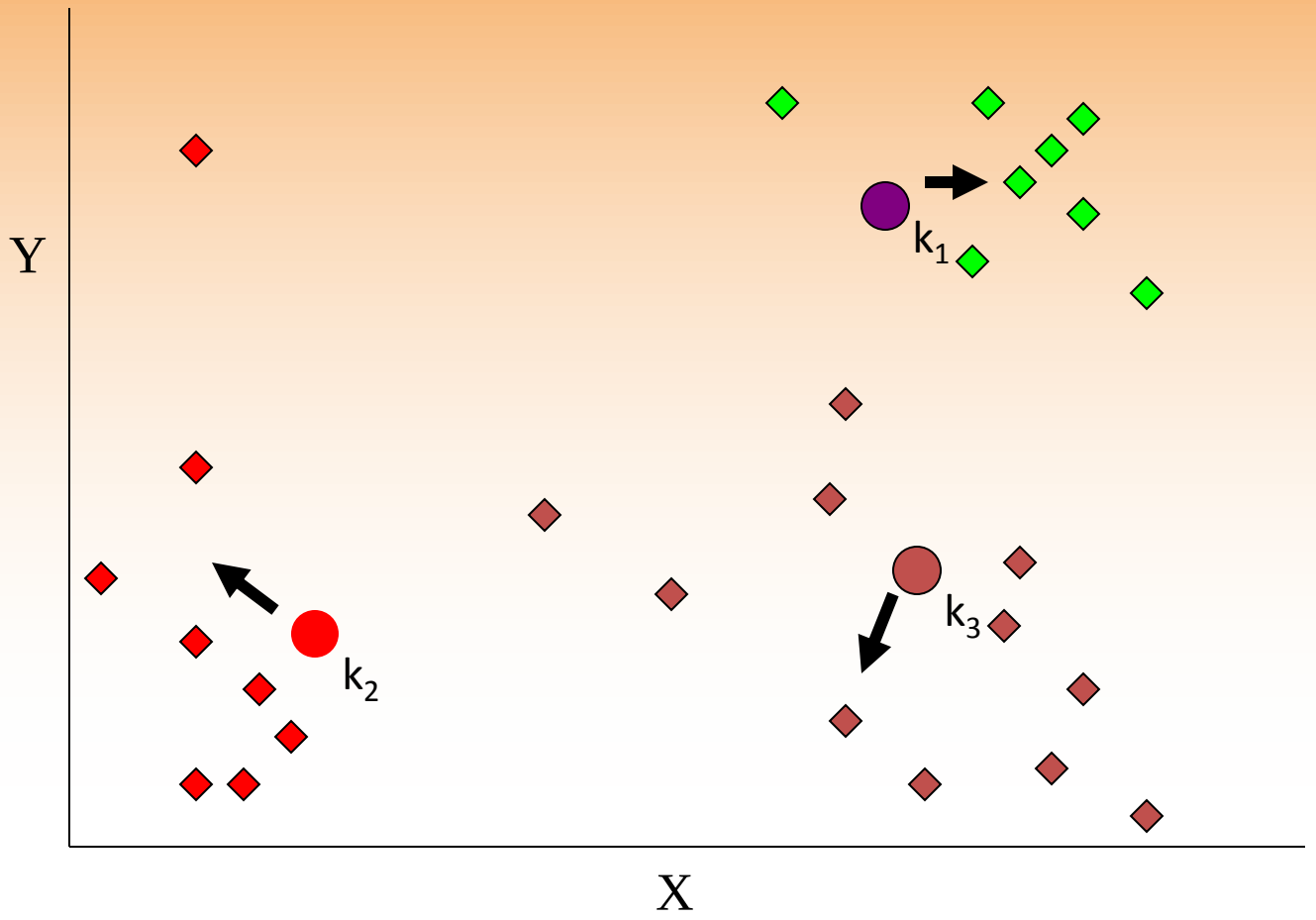


Przykład (5) – krok 4 ...



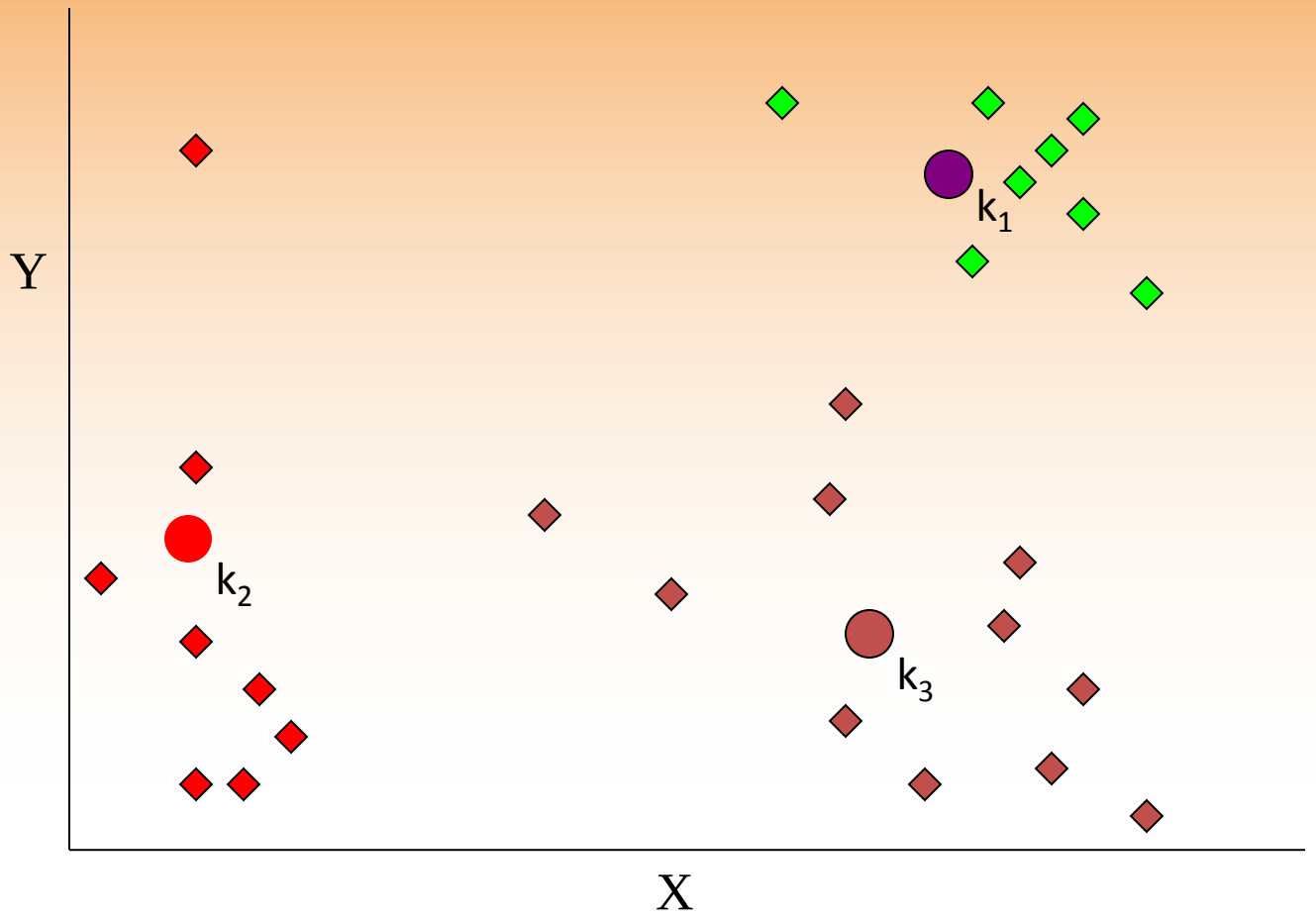
Przykład (6) – krok 4b

oblicz nowe średnie klastrów



Przykład (7) – krok 5

Przesuń środki klastrów do nowo obliczonych średnich



Algorytm k-średnich (2)

- Złożoność algorytmu k-średnich wynosi **O(knI),** gdzie I oznacza liczbę iteracji
- Dla danego zbioru środków klastrów m_k, w ramach jednokrotnego przeglądu bazy danych można obliczyć wszystkie K*n odległości $d(m_k, x)$ i dla każdego obiektu x wybrać minimalną odległość; obliczenie nowych środków klastrów można wykonać w czasie O(n)
- Algorytm bardzo czuły na dane zaszumione lub dane zawierające punkty osobliwe, gdyż punkty takie w istotny sposób wpływają na średnie klastrów powodując ich zniekształcenie

Algorytm k-średnich (3)

- Wynik działania algorytmu (tj. ostateczny podział obiektów pomiędzy klastrami) silnie zależy od początkowego podziału obiektów
- Algorytm może „wpaść” w optimum lokalne
- W celu zwiększenia szansy znalezienia optimum globalnego należy kilkakrotnie uruchomić algorytm dla różnych podziałów początkowych lub spróbować poprawić jakość grupowania metoda postprocessingu

Problem punktów osobliwych

- Wadą algorytmu k-średnich jest jego czułość na występowanie punktów osobliwych
- Punktami osobliwymi nazywamy obiekty, które znacząco różnią się od pozostałych grupowanych obiektów
- Punkty osobliwe reprezentują najczęściej dwa przypadki:

  (1) błędy w danych

  (2) obiekty o bardzo specyficznych wartościach atrybutów
- Najprostsze rozwiązanie polega na usunięciu obiektów, które znacząco odbiegają od środków klastrów

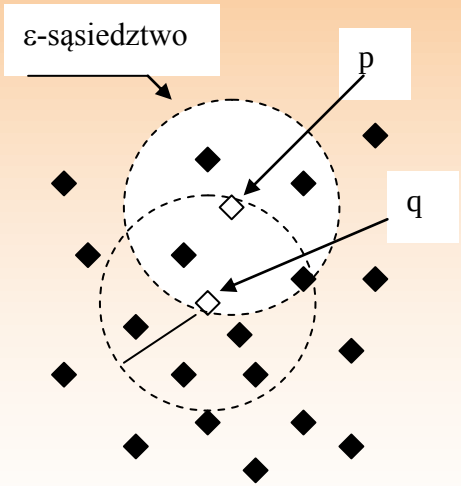
Inne metody grupowania (1)

- Metody oparte o analizę gęstości: dany klaster jest rozszerzany o obiekty należące do jego sąsiedztwa, pod warunkiem, że gęstość obiektów w danym sąsiedztwie przekracza zadaną wartość progową (algorytmy DBSCAN, OPTICS, DENCLUE)
- Metody oparte o strukturę gridową: przestrzeń obiektów jest dzielona na skończona liczbę komórek, które tworzą strukturę gridu; cały proces grupowania jest wykonywany na tej strukturze (algorytmy STING, CLIQUE).
- Metody oparte o konstrukcję modelu: metody te zakładają pewien model dla każdego z klastrów, a następnie, przypisują obiekty do klastrów zgodnie z przyjętymi modelami (algorytm EM)

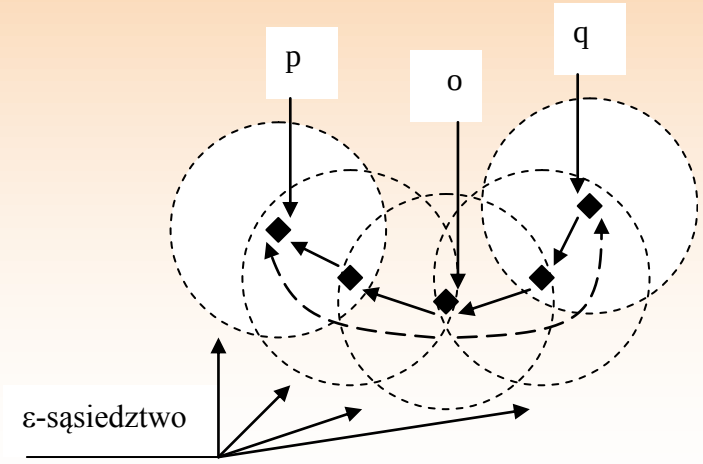
Grupowanie gęstościowe

- Proces grupowania w metodach grupowania gęstościowego jest oparty na pojęciu *gęstości* (ang. *density*)
- Klastrem obiektów jest obszar w przestrzeni obiektów charakteryzujący się dużą gęstością obiektów; klastry obiektów są odseparowane od siebie obszarami o małej gęstości obiektów
- Podstawowa idea metod grupowania gęstościowego polega na przyrostowej konstrukcji klastra, do którego dołączane są obiekty należące do najbliższego sąsiedztwa tego klastra, pod warunkiem że gęstość najbliższego sąsiedztwa klastra jest większa od pewnej zadanej wartości progowej

Gęstościowa osiągalność



(a)



(b)

Klaster

- Niech D oznacza zbiór obiektów. Klastrem C, względem parametrów ε i MinPts, jest dowolny podzbiór obiektów zbioru D spełniający warunki:

  (1) dla dowolnych obiektów p, g $\in$ D: jeżeli p $\in$ C i q jest gęstościowo osiągalny z obiektu p, względem parametrów ε i MinPts, wówczas q $\in$ C,

  (2) dla dowolnych obiektów p, g $\in$ C: obiekt p jest gęstościowo połączony z obiektem q, względem parametrów ε i MinPts
- Niech C_1, …, C_k oznacza zbiór klastrów zbioru obiektów D względem parametrów ε i MinPts
- Zbiorem punktów osobliwych jest taki podzbiór obiektów zbioru D, który nie należy do żadnego klastra C_i, $i = 1, \ldots, k$

Algorytm DBSCAN

- Algorytm jest realizowany w trzech krokach:
  1. rozpoczynamy od dowolnego obiektu $p \in D$,
  2. jeżeli ε-sąsiedztwo obiektu p spełnia warunek minimalnej gęstości, to jest $|N\varepsilon(p)| \geq MinPts$, wówczas tworzony jest klaster C i wszystkie obiekty gęstościowo osiągalne z obiektu p są dołączane do klastra C; w przeciwnym razie (obiekt p nie jest obiektem centralnym), wracamy do kroku (1) i wybieramy następny obiekt zbioru D,
  3. proces grupowania jest kontynuowany tak długo, aż zostaną przetworzone wszystkie obiekty zbioru D
- Obiekty, które nie zostały zaklasyfikowane do żadnego z klastrów, tworzą zbiór punktów osobliwych