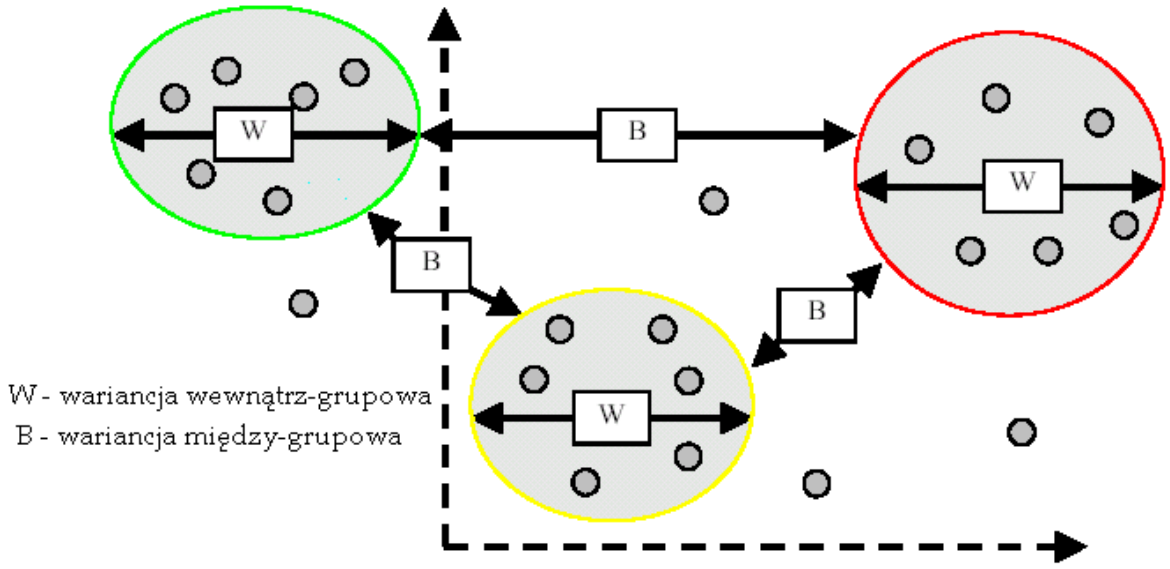
Dodatek B. Klasteryzacja.

Klasteryzacja (grupowanie) jest jedną z metod nie nadzorowanej (bez dostępnej a priori wiedzy) analizy danych. Głównym celem klasteryzacji jest podział rozpatrywanego zbioru obiektów na grupy (klastry), w ten sposób, aby każda z grup była możliwie jednorodna (tzn. zawierała elementy podobne do siebie), a jednocześnie poszczególne klastry były jak najbardziej zróżnicowane między sobą (rys. B1).



Rys.B1. Idea klasteryzacji. Dążymy do takiego podziału zbioru danych aby wariancja wewnątrz-grupowa (w każdym z klastrów była możliwie mała) a jednocześnie wariancja między-grupowa możliwie duża.

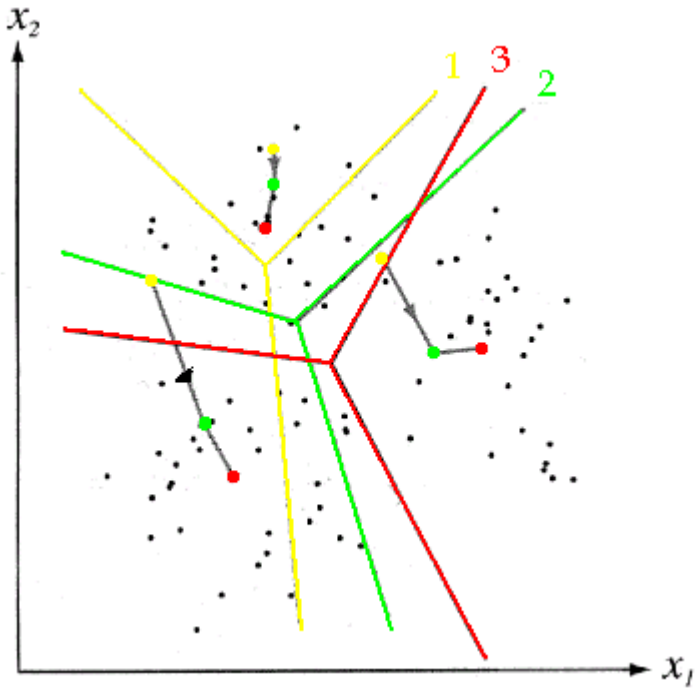
KLASTERYZACJA METODĄ K – ŚREDNICH

Klasteryzacja metodą k-średnich (MacQueen, 1967) jest jedną z prostszych metod dokonujących podziału zbioru danych na grupy. Polega ona na utworzeniu k grup (centroidów) i wyznaczeniu „wejścia" do nich bazującego na pewnej mierze podobieństwa. Jeśli dwa obiekty są względnie podobne, umieszczamy je w tej samej grupie, i uaktualniamy kryteria podobieństwa uwzględniając nowy zbiór elementów w grupie. Dane wejściowe mogą być w formie wektorów, a każda z grup jest początkowo wyznaczona przez jeden, losowo wybrany wektor. W każdym kroku algorytm porównuje wszystkie wektory wejściowe, do każdej z grup, a następnie wektor o najmniejszej odległości do jednej z grup jest do niej przypisywany poczym przeliczana jest wartość centroidu tej grupy. Ogólnie algorytm k-średnich może być przedstawiony w następujących punktach:

1. Wyznaczamy k punktów w przestrzeni reprezentowanej przez obiekty które będą grupowane. Punkty te będą stanowić inicjalne centroidy grup.
2. Przypisujemy, każdy z obiektów do grupy, dla której odległość między jej centroidem a danym obiektem jest najmniejsza (wg określonej miary).
3. Gdy wszystkie punkty zostaną przypisane obliczamy nowe pozycje centoidów w powstałych w ten sposób klastrach.
4. Powtarzamy kroki 2 i 3 dopóki zmiany pozycji centroidów (w kolejnych krokach) będą wystarczająco małe, lub dopóki nie zostanie wykonana określona liczba iteracji.

Algorytm k-srednich jest stosunkowo prosty i szybki, ale ma pewne ograniczenia. Po pierwsze liczba klastrów k na jaką ma być podzielony zbiór wejściowy musi być określona z góry. To może być największą wadą, ponieważ, w niektórych przypadkach liczba klastrów jest dokładnie tym czego szukamy. Aby poradzić sobie z tym problemem często stosuje się rozwiązanie polegające na wielokrotnym przeprowadzeniu klasteryzacji dla różnych wartości k, a następnie wybraniu najlepszych wyników. Innym problemem jest inicjalny wybór centroidów, możemy wyróżnić tu kilka metod:

1. Metoda losowa: w sposób losowy dzielimy dane wejściowe na k klastrów (wybieramy losowo k centroidów)
2. Metoda Forgy'a:
3. Metoda Macqueen'a
4. Metoda Kaufman'a

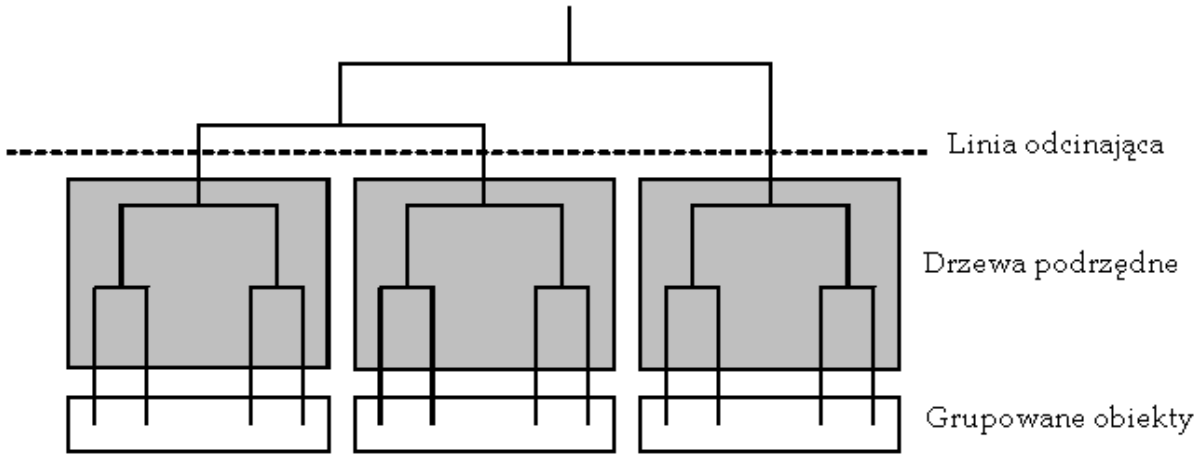


Rys. B2. Algorytm k-średnich w dwuwymiarowej przestrzeni danych, liczba klastrów k = 3. Trzy inicjalne wartości centroidów (kolor żółty) wybrane losowo z zestawu danych. Trzy diagramy Voronoi odpowiadają trzem kolejnym iteracjom algorytmu.

KLASTERYZACJA HIERARCHICZNA

Techniki klasteryzacji hierarchicznej można podzielić na dwie grupy: skupiające (gromadzące) i dzielące. Przy technice dzielącej rozpoczynamy od zdefiniowania jednego klastra, do którego należą wszystkie dane wejściowe. W kolejnych krokach dokonujemy podziału dotąd, aż każdy element wejściowy (każdy obiekt) sam będzie stanowił klaster (będzie jedynym elementem należącym do tego klastra). Stosując technikę skupiającą zaczynamy od pojedynczych obiektów tworzących klastry (w których same są jedynym elementem), a następnie w każdym kroku, łączymy dwa klastry, aż do momentu uzyskania jednej grupy skupiającej wszystkie obiekty.

Wynik klasteryzacji hierarchicznej przedstawiany jest zazwyczaj w postaci drzewa zwanego dendrogramem. Dendrogram jest wykresem dendrytowym pokazującym jak poszczególne klastry są ze sobą związane. Liście dendrogramu są elementami wejściowymi, a korzeń jest końcowym wynikiem klasteryzacji łączącym wszystkie liście. Rozgałęzienie w tym drzewie występuje w punkcie, w którym są łączone dwa klastry (lub jeden klaster podzielony na dwa – przy klasteryzacji dzielącej). Przez „ucięcie" dendrogramu na określonym poziomie, otrzymujemy zbiór rozłącznych grup (klastrów) rys. B3.



Rys. B3 Wytwarzanie rozłącznych klastrów poprzez „obcięcie" dendrogramu.

Klasteryzacja hierarchiczna przeprowadzana jest na podstawie pewnej miary podobieństwa obiektów. W pierwszym etapie algorytm buduje macierz zawierającą odległości pomiędzy kolejnymi parami obiektów (stosuje się tu różne miary odległości, część

z nich opisana jest w końcowej części tego dodatku). Mała wartość w tabeli odległości pozwala przypuszczać, że te dwa klastry/obiekty są bardziej podobne do siebie niż inne klastry/obiekty z większą wartością miary odległości. Podczas wykonywania klasteryzacji techniką skupiania skanujemy macierz odległości w poszukiwaniu najmniejszej odległości. Element a(i,j), będący najmniejszą wartości w macierzy odległości, wyznacza nam dwa klastry (pierwszy określony przez numer rzędu 'i', a drugi określony przez numer kolumny 'j'), które zostaną złączone. W ten sposób powstanie nowa grupa, składająca się z dwóch połączonych ze sobą klastrów. Kolejnym etapem jest uaktualnienie macierzy odległości. Rzędy i kolumny odpowiadające połączonym klastrom są usuwane a na ich miejsce wprowadzony jest jeden rząd i jedna kolumna odpowiadające nowo powstałej grupie. Konieczne jest w tym miejscu określenie w jaki sposób będziemy określać odległość pomiędzy dwoma klastrami (skupiskami obiektów). Najczęściej stosowane metody to (rys.B4):

- Pojedyncze wiązanie (*single linkage*)

  W tej metodzie łączenie grup opiera się na odległości pomiędzy najbliższymi elementami należącymi do łączonych klastrów. Grupy z najmniejszą odległością pomiędzy ich najbliższymi elementami są łączone jako pierwsze. Rozpoczynamy od grup o rozmiarze 1 (każdy obiekt tworzy jeden klaster), a następnie, po każdym połączeniu zmniejszamy liczbę grup o 1.

- Pełne wiązanie (*complete linkage*)

  W tej metodzie używamy odległości pomiędzy najbardziej odległymi elementami należącymi do grup, ażeby zdecydować które z dwóch klastrów połączyć jako pierwsze.

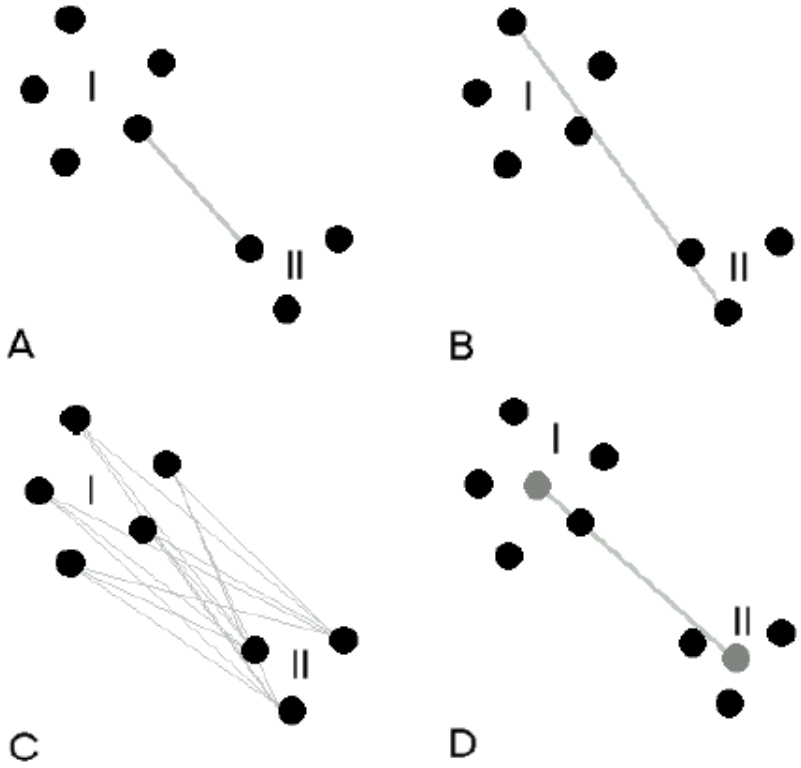
- Wiązanie średnich (*average linkage*)

  W tym wypadku, odległość między klastrami definiujemy jako średnią odległość pomiędzy wszystkimi parami elementów należących do obu grup metodę tą określa się skrótem UPGMA (*unweighted pair group method with arthmetic mean*). Jeśli użyjemy wag w postaci wielkości klastrów (liczby znajdującej się w nich elementów) otrzymamy metodę WPGMA (*weighted pair group method with arithmetic mean*)

- Metoda centroidów

  Dla każdej z grup obliczamy centroid – jako wartość średnią wszystkich obiektów (wektorów) należących do danej grupy. Odległość między klastrami jest definiowana jako odległość między centroidami tych klastrów. Metodę tą określa się skrótem

UPGMC (*unweighted pair group method centroid*). Istnieje również metoda uwzględniająca, w postaci odpowiednich wag, wielkość klastrów. Określa się ją skrótem WPGMC (*unweighted pair group method centroid*).



Rys B4. Wizualna reprezentacja odległości pomiędzy klastrami (grupami): a)pojedyncze wiązanie b)pełne wiązanie c)wiązanie średnich d)odległość między centroidami grup