Ćwiczenie 4. Analiza skupień przy użyciu algorytmu k-means

Głównym celem ćwiczenia jest zbadanie zdolności uogólniania wybranego modelu (MLP, CART lub SVM) w klasyfikacji danych. Dane wejściowe zostaną przedstawione za pomocą `k` środków wyznaczonych z rekordów wejściowych badanego zbioru. Aby umożliwić realizację klasyfikacji na wyznaczonych środkach, konieczne będzie przeprowadzenie klasteryzacji wektorów wejściowych dla każdej klasy z osobna. Należy do tego celu wykorzystać funkcję `kmeans`; przykładowe wywołanie tej funkcji przedstawiono na Listingu 1 w liniach 13 (dla klasy o etykiecie 1 zbioru DBC) oraz 16 (dla klasy o etykiecie 2 zbioru DBC) dla określonej (na podstawie liczności danych – w tym wypadku 20%) liczby klastrów – linie odpowiednio 12 i 15. Do realizacji zadania liczba środków musi zostać dobrana indywidualnie do danego zbioru – dla zbiorów mniej licznych proponuje się zastosowanie 50%-80% danych do klasteryzacji.

```
1  clear all
2  clc

3  load('DBC');
4  X = DBC(:,1:end-1);
5  Y = DBC(:,end);

6  % Wyłuskanie rekordów dla każdej klasy
7  ind_1 = find(Y == 1);
8  X1 = X(ind_1,:);
9  ind_2 = find(Y == 2);
10 X2 = X(ind_2,:);

11 % Klastry dla danych klasy X1 i X2 - środki z 20% rekordów
12  k1   = round(0.2 * length(ind_1));
13 [k_1,Klastry_1] = kmeans(X1, k1, 'Distance', 'sqeuclidean');
14 Dane_klasy_1 = [Klastry_1 ones(k1, 1)];
15 k2   = round(0.2 * length(ind_2));
16 [k_2,Klastry_2] = kmeans(X2, k2, 'Distance', 'sqeuclidean');
17 Dane_klasy_2 = [Klastry_2 2*ones(k2, 1)];

18 Klastry = [Dane_klasy_1; Dane_klasy_2];
```

Listing 1. Klasteryzacja danych i wyznaczanie określonej liczby środków w aplikacji Matlab.

Tak jak w poprzednich ćwiczeniach zdolność uogólniania wybranego klasyfikatora należy wyznaczyć na podstawie kryterium dokładności. Tym razem, dane wejściowe będą przedstawione w formie zbioru `Klastry`.