

Eksploracja danych

Projekt 3: Grupowanie przy użyciu K-means

To zadanie skupia się na technice grupowania K-means. K-means to algorytm partycjonujący, który jest jedną z najczęściej używanych metod grupowania, ponieważ jest dość łatwy do zrozumienia i wdrożenia. Dzisiejsze zadanie polega na zaimplementowaniu algorytmu i zbadaniu jego cech na dwóch różnych 4-wymiarowych zbiorach danych.

Zestawy danych można znaleźć w pliku

<http://www.math.us.edu.pl/~pgladki/teaching/2021-2022/ed-p03.mat>

Kiedy załadujesz ten plik do MatLaba, znajdziesz dwie macierze:

1. *Patterns*: macierz 150×4 , gdzie każdy wiersz zawiera jeden wzorzec,
2. *Patterns2*: macierz 200×4 , gdzie każdy wiersz zawiera jeden wzorzec.

Zadania te należy wykonywać na obu zbiorach danych.

Użyj analizy głównych komponentów (PCA), aby rzutować 4-wymiarowe wektory wzorców na dwa główne komponenty. Umożliwi to przeglądanie wzorów w dwóch wymiarach. Pozwoli Ci to odgadnąć, ile jest klastrów, co z kolei będzie przydatne do podjęcia decyzji jakich parametrów użyć dla różnych algorytmów.

Zaimplementuj algorytm grupowania K-means i znajdź odpowiednią wartość dla k . Pamiętaj, że grupowanie zapewniane przez algorytm K-means zależy od początkowego rozmieszczenia klastrów, więc rozsądnie może być wykonać kilka przebiegów dla każdego k i wybranie grupowania, które daje najniższą średnią odległość od środka klastra.