

Politechnika Gdańska
Wydział Elektrotechniki i Automatyki
Katedra Inżynierii Systemów Sterowania

Monitorowanie i diagnostyka w systemach sterowania

Klasteryzacja danych

Zadania do ćwiczeń laboratoryjnych – zajęcia nr 5

Opracowanie:

dr inż. Michał Grochowski

Zadanie 1

Celem laboratorium jest zapoznanie się z podstawowymi algorytmami klasteryzacji danych:

- klasteryzacji hierarchicznej;
- klasteryzacji typu: k-means;
- klasteryzacji typu: Gaussian Mixture
- klasteryzacji typu: Soft Gaussian Mixture.

Należy:

1. Zapoznać się dokładnie z poleceniami:
 - `pdist`
 - `squareform`
 - `linkage`
 - `dendrogram`
 - `cophenet`
 - `inconsistent`
 - `cluster`
 - `kmeans`
 - `silhouette`
 - `gmdistribution.fit`
 - `clusterdata`
2. Przeanalizować (ze zrozumieniem !!) procesy klasteryzacji przedstawiony w załączonym pliku: `DataSetsClustering_tutorial.m`
3. Zbadać wpływ różnych miar odległości pomiędzy elementami zbioru danych na proces klasteryzacji.
4. Zbadać wpływ ilości klastrów na stopień dopasowania elementów zbioru do danego klastra.
5. Zbadać wpływ poziomu "przecięcia" dendrogramu na stopień dopasowania elementów zbioru do danego klastra.
6. Wskazać cechy wspólne i różnice w sposobach klasteryzacji danych.

Zadanie 2

Zadanie polega na dokonaniu klasteryzacji zbioru danych zawierających informacje charakteryzujące kwiaty lrysy. Baza danych (`irysy.mat`) zawiera następujące elementy opisu kwiatów:

- długość liści (w cm);
- szerokość liści (w cm);
- długość płatków kwiatowych (w cm);
- szerokość płatków kwiatowych (w cm).

Należy:

1. Za pomocą 2 metod: klasteryzacji hierarchicznej oraz klasteryzacji typu k - means (lub innej), dokonać grupowania danych opisujących kwiaty.
2. Zbadać ile rodzajów kwiatów jest opisanych w tej bazie.
3. Zbadać w jakim stopniu elementy pasują do wybranych klastrów w przypadku różnej ich liczby.
4. Oprócz funkcji przedstawionych w tutorialu, wykorzystać do grupowania danych funkcję: `clusterdata`. Szczegóły jej parametrów dostępne w `doc help clusterdata`.

Dane o kwiatach pochodzą z:

Murphy, P.M., Aha, D.W. (1994). UCI Repository of machine learning databases [<http://www.ics.uci.edu/~mlearn/MLRepository.html>]. Irvine, CA: University of California, Department of Information and Computer Science.
From R.A. Fisher.

Zadanie 3

Wykorzystując wiedzę o algorytmach klasteryzacji rozpatrz ponownie problem wykrywania uszkodzeń silnika przy pomocy metody PCA.

Przeanalizuj dane pochodzące z symulacji silnika i spróbuj zbudować 2 (lub więcej modeli) lokalne modele PCA, zależne od punktu (obszaru) pracy silnika. Porównaj skuteczność działania tych modeli z modelem PCA zaprojektowanym na poprzednich zajęciach.