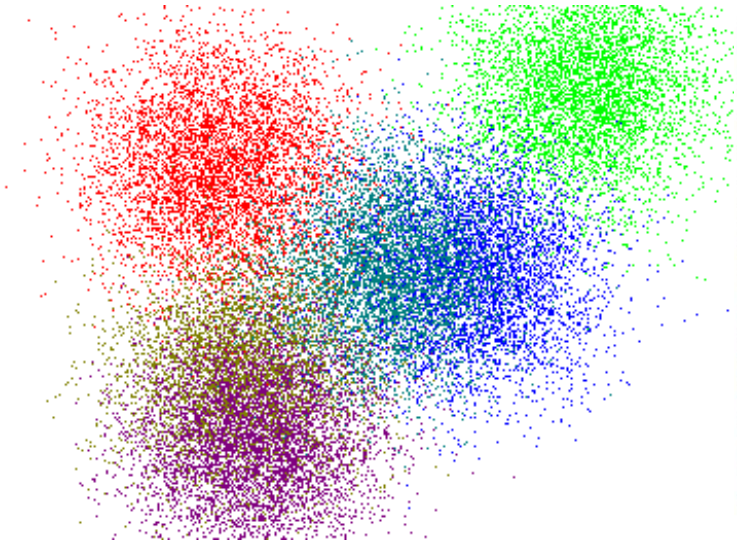
UCZNIE NIENADZOROWANE

ALGORYTM KMEANS

Jednym z zadań uczynienia nienadzorowanego jest tzw. proces analizy skupisk. Proces, w trakcie którego odnajdujemy w nieoznakowanych danych obiekty podobne i ‘zamykamy’ je w klastry. W trakcie tzw. ostrej klasteryzacji obiekt może należeć tylko do jednego klastra. W klastrze znajdują się obiekty podobne sobie względem przyjętej metryki (np. Euklidesowa). Obiektem może być np.: pacjent w klinice onkologicznej, klient sklepu internetowego a celem procesu uczenia odnalezienie podobnym obiektów. Przykład klasteryzacji dla danych dwuwymiarowych został przedstawiony na rys 1a. Jednym z prostszych algorytmów klasteryzacji jest algorytm kMeans, jego pseudokod został przedstawiony na rys 1b.



```
N:  number of data objects
K:  number of clusters

objects[N]:  array of data objects
clusters[K]:  array of cluster centers
membership[N]:  array of object memberships

kmeans_clustering( )
1   while  δ/N > threshold
2      δ ← 0
3      for i ← 0 to N-1
4         for j ← 0 to K-1
5            distance ← | objects[i] - clusters[j] |
6            if distance < dmin
7               dmin ← distance
8               n ← j
9         if membership[i] ≠ n
10           δ ← δ + 1
11           membership[i] ← n
12        new_clusters[n] ← new_clusters[n] + objects[i]
13        new_cluster_size[n] ← new_cluster_size[n] + 1
14     for j ← 0 to K-1
15        clusters[j][*] ← new_clusters[j][*] / new_cluster_size[j]
16        new_clusters[j][*] ← 0
17        new_cluster_size[j] ← 0
```

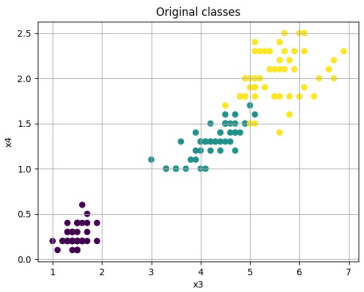
Rys 1 a) Przykład klasteryzacji dla danych dwuwymiarowych; b) pseudokod algorytmu kMeans

OCENA KLASTERYZACJI

Jakoś klasteryzacji można ocenić z wykorzystaniem tzw. indeksów klasteryzacji. W zależności czy posiadamy informacje o prawdziwej przynależności obiektów (w praktyce rzadko) czy nie możemy skorzystać z takich wskaźników –wymieniając tylko kilka z wielu- jak homogeniczności, przyległości Randa[1], indeks Dunna[2] czy współczynnik Silhouette[3].

ZADANIA DO WYKONANIA

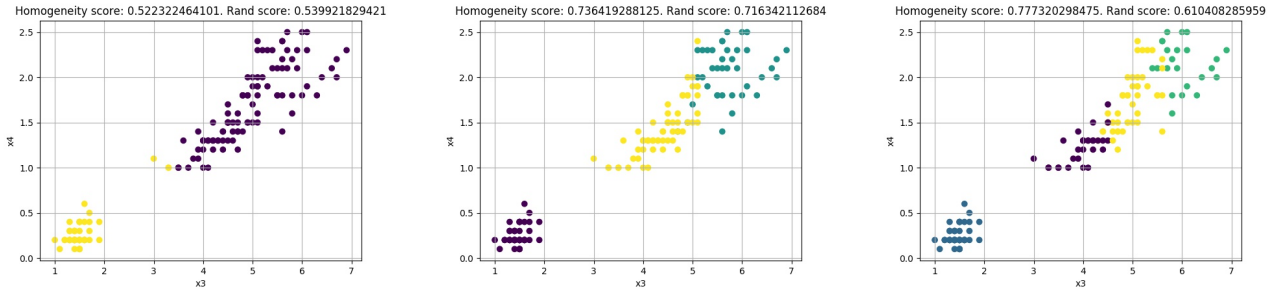
Korzystając z biblioteki **`sklearn`** (http://scikit-learn.org) przetestuj działanie algorytmu kMeans na danych ”Iris”. Sprawdź wyniki dla k=2, k=3, k=4 oraz oceń jakość uzyskanych klastrów korzystając z wskaźnika homogeniczności oraz przyległości Randa.



W celu realizacji zadania należy skorzystać z modułów:

- sklearn.cluster.KMeans
- sklearn.metrics.homogeneity_score
- sklearn.metrics.adjusted_rand_score

Rys 2. Dane z zaznaczonymi prawdziwymi klasami



Rys 3. Przykładowe rozwiązanie dla k=2, k=3, k=4 oraz wskaźnik homogeniczności i Randa

[1] http://faculty.washington.edu/kayee/pca/supp.pdf
[2] http://www.sciencedirect.com/science/article/pii/S0031320303002838
[3] http://www.sciencedirect.com/science/article/pii/0377042787901257?via%3Dihub