Klasteryzacja

Celem laboratorium jest zapoznanie się z algorytmami klasteryzacji k-means oraz DBSCAN.
Na wynik działania algorytmu k-means wpływa głównie parametr

– liczba klastrów – `n_clusters`.

Z kolei na działanie algorytmu DBSCAN wpływają głównie dwa parametry, definiujące razem pojęcie gęstości:

– promień sąsiedztwa – `eps`,

– wymagana liczba punktów w sąsiedztwie – `min_samples`.

Zadanie 1. Celem zadania jest zbadanie, jak algorytmy k-means i DBSCAN radzą sobie ze zbiorami o różnej naturze. Wykorzystamy syntetycznie generowane zbiory:

- blobs
- circles
- moons
- ellipses

Zbadaj wpływ podanych parametrów algorytmów na kształt uzyskanych klastrów.

Zadanie 2. W zadaniu należy wykorzystać zbiór `banknotes`. Pierwsza kolumna zbioru wskazuje, czy banknot jest oryginalny czy sfałszowany. Pozostałe sześć kolumn opisuje cechy konkretnych banknotów, takie jak długość czy szerokość. Zastosuj algorytmy k-means i DBSCAN do pogrupowania banknotów w klastry. Ponieważ oba algorytmy wykorzystują pojęcie odległości euklidesowej, na początku znormalizuj wartości cech.

W przypadku algorytmu k-means oszacuj optymalną liczbę klastrów, korzytając z:

- metody *elbow*,
- wyniku profilu (ang. silhouette score),

- kryterium informacyjnego Akaikego (ang. Akaike Information Criterion).

W przypadku algorytmu DBSCAN sprawdź, czy wszystkie banknoty zostały przypisane do klastrów.

Czy wyniki klasteryzacji są zgodne z podziałem na banknoty oryginalne i sfałszowane? Oceń jakość otrzymanych klastrów, korzystając z miar [1]:

- homogeniczności (ang. homogeneity),
- zupełności (ang. completeness),
- V-miary (ang. V-measure).

[1] `https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation`