Laboratorium: Klasteryzacja

April 18, 2023

1 Cel/Zakres

- Klasteryzacja
- Znajdywanie parametrów dla algorytmów klasteryzacji.

2 Przygotowanie danych

```
from sklearn.datasets import fetch_openml
import numpy as np

mnist = fetch_openml('mnist_784', version=1, as_frame=False, parser='auto')
mnist.target = mnist.target.astype(np.uint8)
X = mnist["data"]
y = mnist["target"]
```

3 Ćwiczenie

Pozyskane dane (zmienna `X`) reprezentują zeskanowane znaki nieznanego alfabetu ;). Celem ćwiczenia jest identyfikacja ile tych znaków jest i jak mogą one wyglądać.

Zakładając, że możemy mieć do czynienia z 8–12 różnymi znakami uzyj metody centroidów do ich klasteryzacji.

1. Przeprowadź klasteryzację dla 8, 9, 10, 11 i 12 skupisk.
2. Wylicz wartośc wskaźnika sylwetkowego dla każdego z ww. skupisk. Zapisz wartość wszystkich wskaźników sylwetkowych jako listę w pliku Pickle o nazwie `kmeans_sil.pkl`.

   5 pkt.
3. Znany lingwista prof. Talent twierdzi, że w zbiorze `X` można zidentyfikować 10 różnych znaków. Czy wartości wskaźnika sylwetkowego potwierdzają tą obserwację?
4. Prof. Talent dostarczył swoich wyników klasyfikacji w postaci zbioru `y`. Policz macierz błędów pomiędzy danymi otrzymanymi z procesu klasteryazacji dla 10 skupisk i zbioru `y`.
5. Dla każdego wiersza ww. macierzy znajdź indeks o najwyższej wartości (np. `numpy.argmax()` albo `pandas.Series.argmax()`). Wartości umieść na posortowanej rosnąco liście bez duplikatów (użyj np. `set()`). Listę zapisz w pliku Pickle o nazwie `kmeans_argmax.pkl`.

2 pkt.

6. Znajdź sensowne wartości parametru `eps` dla DBSCAN. Heurystyka dla określenia wartości parametru `eps` oparta jest o odległość euklidesową pomiędzy instancjami. Policz odległości dla pierwszych 300 elementów ze zbioru `X` ze wszystkimi pozostałymi elementami w zbiorze `X` (użyj np. `numpy.linalg.norm(x1-x2)`, gdzie `x1` i `x2` to punkty w przestrzeni wielowymiarowej), pomiń odległości równe 0, a następnie wyświetl 10 najmniejszych. Ww. 10 wartości umieść na liście w kolejności rosnącej, a listę zapisz w pliku Pickle o nazwie `dist.pkl`.

   2 pkt.

7. Policz średnią `s` z 3 najmniejszych wartości z ww. listy. Przyjmij kolejno wartości `eps` od `s` do `s+10%*s` z krokiem co `4%*s` i wykonaj klasteryzacje.

8. Dla każdej klasteryzacji (dla kolejnych wartości `eps`) policz ile jest unikalnych etykiet zidentyfikowanych przez algorytm DBSCAN. Wartości umieść na liście i zapisz w pliku Pickle o nazwie `dbscan_len.pkl`.

   5 pkt.

4 Prześlij raport

Prześlij plik o nazwie `lab07/lab07.py` realizujący ww. ćwiczenia.

Sprawdzane będzie, czy skrypt Pythona tworzy wszystkie wymagane pliki oraz czy ich zawartość jest poprawna.