Praca domowa 4

Termin oddania: 24.01.2024

1 Wstęp

W pracy domowej rozpatrujemy 3 różne metody analizy skupień: *k–średnich*, *metodę hierarchiczną* oraz *DBSCAN*. Bazując na załączonych zbiorach danych chcemy odpowiedzieć na postawione pytania badawcze.

2 Dane

W tej pracy domowej posłużymy się danymi sztucznie generowanymi. Poniższe instrukcje pozwolą na uzyskanie rozpatrywanych zbiorów danych. Graficzna reprezetnacja została zamieszczona na Rysunku 1.

```
import numpy as np
from sklearn.datasets import make_blobs, make_circles, make_moons

## Dane 1
n = 1500
X, y = make_blobs(n_samples = n)

## Dane 2
noise = 0.05
X, y = make_circles(n_samples = n, factor = 0.5, noise = noise)

## Dane 3
noise = 0.05
X, y = make_moons(n_samples = n, noise = noise)

## Dane 4
X = np.random.rand(n, 2)
```

3 Pytania badawcze

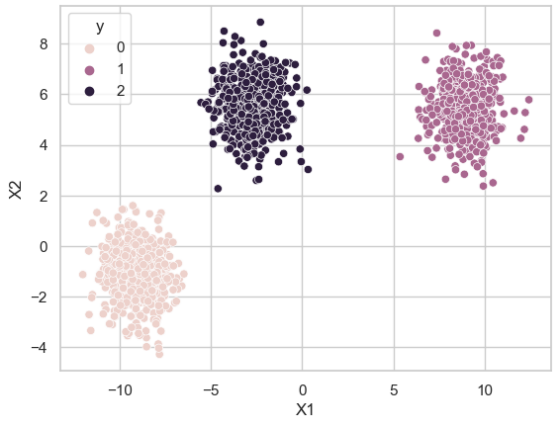
1. Które algorytmy wymagają liczby klastrów na wejściu? Jak wybór tego parametru wpływa na wyniki?
2. Zbadaj wpływ parametru szumu (`noise`) dla zestawów danych 2 i 3.
3. Wygeneruj wykres pokazujący, jak całkowita suma odległości między punktami w klastrach zależy od liczby klastrów.

4 Szczegóły rozwiązania

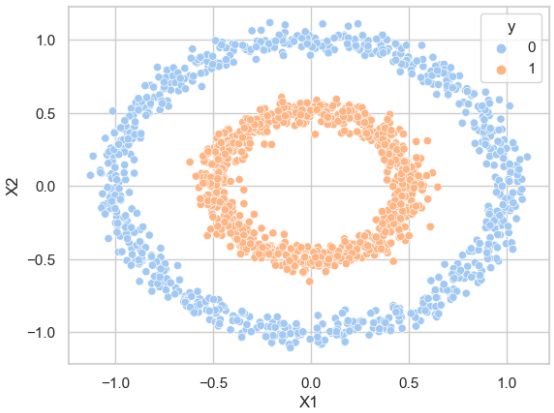
Rozwiązanie powinno zawierać pliki:

- folder `Kody` zawierający wszystkie potrzebne kody do przygotowania rozwiązania zadania domowego,
- plik `NUMERINDEKSU_raport.pdf` opisujący wyniki (maksymalnie 3 strony, wykresy nie wliczają się do limitu stron).

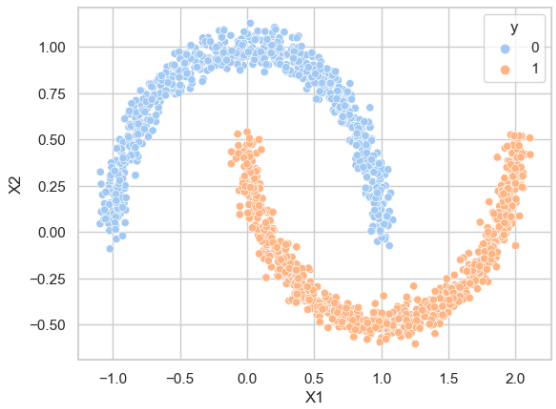
Rysunek 1: Reprezentacja graficzna zbiorów z Sekcji 2.



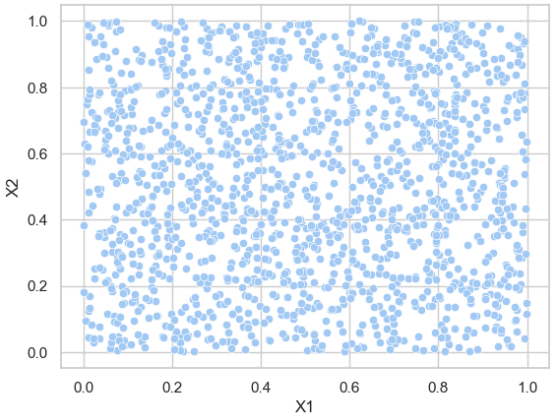
(a) Dane 1



(b) Dane 2



(c) Dane 3



(d) Dane 4

5 Ocena

Łączna liczba punktów do zdobycia jest równa 10, w tym:

Pytanie badawcze 1 (5 punktów)

- przetestowanie metod - 2 punkty,
- raport - 2 punkty,
- jakość kodu (porządek, czytelność) - 1 punkt.

Pytanie badawcze 2 (2,5 punkta)

- testy parametru - 1,5 punkta,
- raport - 1 punkt.

Pytanie badawcze 3 (2,5 punkta)

- analiza sumy odległości - 1,5 punkta,
- raport - 1 punkt.

6 Oddanie pracy domowej

Wszystkie punkty z sekcji *Szczegóły rozwiązania* należy umieścić w katalogu ZIP o nazwie `NUMERINDEKSU_GR_PD4`, gdzie

$$\texttt{GR} = \begin{cases} 1 & \text{dla środa, 12:15,} \\ 2 & \text{dla środa, 16:15.} \end{cases}$$

Tak przygotowany katalog należy przesłać na adres *anna.kozak@pw.edu.pl* do dnia 24.01.2024 do godziny 23:59. Tytuł wiadomości: *[WUM][PD4] Nazwisko Imię, Numer grupy: GR*.