Praca domowa 3

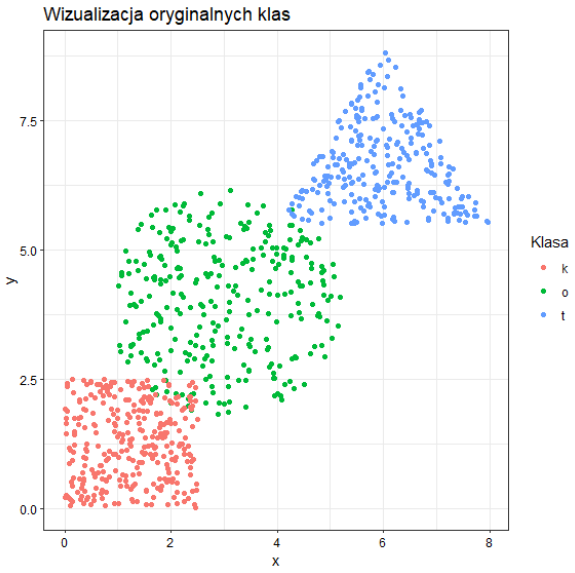
Termin oddania: 22.01.2025

1 Wstęp

Celem pracy domowej jest porównanie metod analizy skupień: *k-średnich* z pakietu `scikit-learn` i własnej implementacji *Algorytmu PAM (ang. Partitioning Around Medoids)*, czyli metody *k-medoidów*.

2 Zbiór danych

Przygotowaną implementację *Algorytmu PAM* oraz metodę *k-średnich* przetestuj na dwóch różnych zbiorach danych. Jednym z nich jest zbiór $\mathcal{K} = (X, y)$, gdzie X `= X.csv,` y `= y.csv`. Drugi zbiór danych przygotuj samodzielnie.



Rysunek 1: Wizualizacja zbioru danych $\mathcal{K}$.

3 Oczekiwany wynik

Praca domowa składa się z trzech elementów. Pierwszym z nich jest implementacja algorytmu PAM. Drugim elementem jest przygotowanie nowego zbioru danych do oceny jakości rozpatrywanych modeli. Ostatni punkt pracy domowej polega na przetestowaniu metody *k-średnich* oraz algorytmu PAM na dwóch zbiorach danych.

3.1 Implementacja algorytmu PAM *(8 punktów)*

Przygotuj funkcję `cluster_PAM(X, k)`, gdzie `X` to zbiór wejściowy (bez etykiet) i `k` to liczba skupień oraz sprawdź poprawność implementacji. Zadbaj o obsługę wyjątków, dokumentację oraz jakość i złożoność kodu. Krótki opis algorytmu PAM poniżej.

Algorytm PAM

Faza budowy

1. Podziel zbiór danych na k skupień z przypisanymi k medoidami.
2. Oblicz macierz odległości pomiędzy medoidami oraz pozostałymi obserwacjami.
3. Przypisz każdą z obserwacji (nie będącą medoidem) do najbardziej zbliżonego skupienia.

Faza zmiany

4. Przy użyciu iteracji zastąp jeden z medoidów jednym z niemedoidów i sprawdź, czy odległości wszystkich elementów niebędących medoidami od najbliższych im medoidów są mniejsze.
5. Jeśli nastąpiła przynajmniej jedna zmiana medoidów, przejdź do punktu 3. Jeśli nie, zakończ algorytm.

3.2 Zbiór danych *(3 punkty)*

Przygotuj drugi zbiór danych do testowania algorytmów. Zbiór danych powinien być z R^2 lub R^3 z ciekawymi zależnościami. Do rozwiązania pracy domowej dołącz plik z wygenerowanymi danymi o nazwie `NUMERINDEKSU_data.csv`, gdzie pierwszą kolumną będzie kolumna etykiet `y`, a kolejne kolumny będą opisane `X1, X2, ...`. Zawrzyj w pliku Jupyter Notebook graficzną reprezentację swojego zbioru.

3.3 Eksperyment (*4 punkty*)

Wykorzystując swoją implementację algorytmu PAM oraz algorytm *k-średnich* z pakietu `scikit-learn` sprawdź ich działanie na dwóch zbiorach danych (jeden, który jest podany w Sekcji 2 i drugi, który jest wynikiem Sekcji 3.2. Rozważ różne wartości hiperparametru k. Co dzieje się, gdy źle dobierzemy k? Czy obie metody dobrze identyfikują skupienia? Jaki jest czas działania metod?

4 Szczegóły rozwiązania

Rozwiązanie powinno zawierać pliki:

- `Algorytm_PAM.py` skrypt zawierający implementację algorytmu PAM,
- `NUMERINDEKSU_wyniki.ipynb` zawierający opis generowania danych i jego graficzną reprezentację oraz eksperymenty dotyczące porównania metod.

5 Ocena

Łączna liczba punktów do zdobycia jest równa 15, w tym:

- 3.1 Implementacja Algorytmu PAM (*8 punktów)*
  - implementacja oraz testy poprawności algorytmu - 5 punktów,
  - obsługa wyjątków - 1 punkt,
  - dokumentacja - 1 punkt,
  - jakość kodu, złożoność - 1 punkt.
- 3.2 Zbiór danych (*3 punkty)*
  - opis generowania zbioru danych - 2 punkty,
  - dołączenie zbioru danych oraz jego wizualizacji - 1 punkt.

- 3.3 Eksperyment (*4 punkty)*
  - testy algorytmu *k-średnich* - 1 punkt,
  - testy algorytmu *PAM* - 1 punkt,
  - testowanie wyboru hiperparametru k - 2 punkty.

6 Oddanie pracy domowej

Wszystkie punkty z sekcji *Szczegóły rozwiązania* należy umieścić w katalogu ZIP o nazwie `NAZWISKO_IMIE_GR_PD3` (bez polskich znaków), gdzie

$$\texttt{GR} = \begin{cases} 1 & \text{dla środa, 12:15,} \\ 2 & \text{dla środa, 14:15,} \\ 3 & \text{dla środa, 16:15.} \end{cases}$$

Tak przygotowany katalog należy przesłać na adres *anna.kozak@pw.edu.pl* do dnia 22.01.2025 do godziny 23:59. Tytuł wiadomości: *[WUM][PD3] Nazwisko Imię, Numer grupy: GR.*