Praca domowa 3

Termin oddania: 22.01.2026

1 Wstęp

W tej pracy domowej przyjrzymy się modelowi analizy skupień - algorytmowi aglomeracyjnemu z definicją odległości *Single Linkage*.

2 Część 1 *(9 punktów)*

Przygotuj implementację algorytmu hierarchicznego `hierarchical_clustering()`, który przyjmuje jako argumenty kolejno:

1. macierz rzeczywistą `X` typu $n \times p$, reprezentującą n punktów w R^p,
2. hiperparametr `n_clusters` oznaczający liczbę skupień.

Funkcja ma zwracać n – elementowy obiekt `c`, gdzie c_i reprezentuje etykietę skupienia dla obserwacji $\texttt{X}[i,]$.

Algorytm Aglomeracyjny

1. Zdefiniuj początkowy zbiór skupień $C = \{C_1, C_2, \ldots, C_n\}$, gdzie każde skupienie zawiera dokładnie jeden element $C_i = \{x_i\}$. Oblicz macierz odległości D, gdzie $D_{ij} = d(x_i, x_j)$.

2. Odległość między dwoma skupieniami A, B jest definiowana jako minimum odległości pomiędzy ich elementami
$$D(A, B) = \min_{a \in A, b \in B} d(a, b).$$

3. Dopóki liczba skupień jest większa niż wartość hiperparametru `n_clusters` wykonuj:
   - Znajdź parę skupień (C_i, C_j), dla których odległość jest minimalna
   $$(C_i, C_j) = \arg\min_{A,B \in C, A \neq B} D(A, B).$$
   - Połącz skupienia C_i i C_j w jedno nowe skupienie $C_* = C_i \cup C_j$.
   - Zaktualizuj zbiór skupień $C = (C \backslash \{C_i, C_j\}) \cup \{C_*\}$.
   - Zaktualizuj macierz odległości, obliczając odległość między nowym skupieniem C_* a pozostałymi skupieniami $S \in C$
   $$D(C_*, S) = min(D(C_i, S), D(C_j, S).$$
   .

Przeprowadź test poprawności implementacji funkcji `hierarchical_clustering()`.

3 Część 2 *(6 punktów)*

3.1 Zbiór danych *(3 punkty)*

Przygotuj zbiór danych do testowania algorytmu. 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.

3.2 Eksperyment (*3 punkty*)

Wykorzystując swoją implementację algorytmu hierarchicznego oraz algorytm `AgglomerativeClustering` z hiperparametrem `linkage="single"` z pakietu `scikit-learn` sprawdź ich działanie na utworzonym zbiorze danych. Rozważ różne wartości hiperparametru *n_clusters*. Czy obie metody dobrze identyfikują skupienia? Jaki jest czas działania metod?

4 Szczegóły rozwiązania

Rozwiązanie powinno zawierać pliki:

- `hierarchical_clustering.py` skrypt zawierający implementację algorytmu `hierarchical_clustering()`,
- `NUMERINDEKSU_wyniki.ipynb` zawierający testy poprawności implementacji, opis generowania danych i jego graficzną reprezentację oraz eksperymenty dotyczące porównania metod.

5 Ocena

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

- 2 Implementacja algorytmu `hierarchical_clustering()` (*9 punktów)*
  - implementacja oraz testy poprawności algorytmu - 6 punktów,
  - obsługa wyjątków - 1 punkt,
  - dokumentacja - 1 punkt,
  - jakość kodu, złożoność - 1 punkt.
- 3.1 Zbiór danych (*3 punkty)*
  - opis generowania zbioru danych - 2 punkty,
  - dołączenie zbioru danych oraz jego wizualizacji - 1 punkt.
- 3.2 Eksperyment (*3 punkty)*
  - testy algorytmu `hierarchical_clustering()` - 1 punkt,
  - testy algorytmu `AgglomerativeClustering` - 1 punkt,
  - testowanie wyboru hiperparametru `n_clusters` - 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

$$\texttt{GR} = \begin{cases} 1 & \text{dla środa, 12:15 (AK),} \\ 2 & \text{dla środa, 12:15 (KW),} \\ 3 & \text{dla środa, 14:15 (AK),} \\ 4 & \text{dla czwartek, 14:15 (AK).} \end{cases}$$

Tak przygotowany katalog należy przesłać na adres *anna.kozak@pw.edu.pl* do dnia 22.01.2026 do godziny 23:59. Prace przesłane po tym terminie będą miały odjęte 2 punkty za każdy rozpoczęty dzień spóźnienia. Tytuł wiadomości: *[WUM][PD3] Nazwisko Imię, Numer grupy: GR.*
Prace, które nie będą przygotowane w odpowiednim formacie nie będą oceniane.