Ćwiczenie GD

Grupowanie danych

Część teoretyczna

Wykład na temat grupowania danych.
Wikipedia: Mikromacierz DNA.

Zadanie dotyczy grupowania profili ekspresji genów. Do badania ekspresji genów służą mikromacierze DNA. Mikromacierz zawiera znane fragmenty DNA, różniące się od siebie sekwencją kwasów nukleinowych (tzw. sondy). Sondy umieszczone są odpowiednich komórkach macierzy. Badany materiał (próbka DNA) jest wyznakowany znacznikiem fluorescencyjnym i umieszczany na macierzy. Cząsteczki tego materiału łączą się z komplementarnymi sondami (tzn. takimi, które mają analogiczne sekwencje nukleotydów). Komórki macierzy, które zawierają sondy z dołączonymi cząsteczkami badanej próbki dają jaśniejszy obraz. Obraz sczytuje się ilościowo (za pomocą lasera lub mikroskopu). Intensywność sygnału dla poszczególnych sond mikromacierzy jest proporcjonalna do ilości kwasu nukleinowego o danej sekwencji w próbce. Możemy więc określić skład genetyczny badanej próbki.

Dane analizowane w ćwiczeniu[1] zawierały poziomy ekspresji (intensywności obrazu) 6400 genów, mierzone po czasie $t = 0, 9.5, 11.5, 13.5, 15.5, 18.5$ i 20.5 godziny. Mikromacierz zawierała 6400 sond (różnych sekwencji DNA), z których każda reprezentowała inny gen. Po odrzuceniu genów niewystępujących w próbce (puste komórki macierzy), danych błędnych i genów, których poziom ekspresji nie zmienia się znacząco w czasie, pozostało 614 genów. Nazwy tych genów zamieszczone są w pierwszej kolumnie danych (plik GD_dane_genetyczne.csv), czasy pomiaru ekspresji zamieszczone są w pierwszym wierszu danych, a poziomy ekspresji genów dla siedmiu punktów czasowych (profile ekspresji genów) zamieszczone są w kolejnych wierszach.

Zadania pomocnicze

Ćwiczenie wykorzystuje algorytmy grupowania i PCA opisane w rozdz. 12.5 G. James, D. Witten, T. Hastie, R. Tibshirani, J. Taylor: An Introduction to Statistical Learning with Applications in Python (https://www.statlearning.com/). Zapoznaj się z tym materiałem.

Zadania do wykonania

Zadanie polega na pogrupowaniu profili ekspresji genów za pomocą metody *k*-średnich i grupowania hierarchicznego (aglomeracyjnego).

1. Zaimportuj niezbędne moduły (uwaga – należy zainstalować pakiet ISLP: pip install ISLP)

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import (KMeans , AgglomerativeClustering)
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.metrics import silhouette_score
from scipy.cluster.hierarchy import (dendrogram , cut_tree)
from ISLP.cluster import compute_linkage
from statsmodels.datasets import get_rdataset
```

[1] Dane i eksperyment opisane są w: DeRisi, J.L., Iyer, V.R., Brown, P.O. (1997). *Exploring the metabolic and genetic control of gene expression on a genomic scale*. Science 24, 278(5338), 680-686.

```
from ISLP import load_data
```

2. Wczytaj i zmodyfikuj zbiór danych:

```
data= pd.read_csv("../content/drive/MyDrive/Colab
Notebooks/GD_dane_genetyczne.csv")
geny = data.iloc[:,0].tolist()
ekspresja = data.iloc[:,1:].to_numpy()
czas = data.columns[1:].to_numpy()
czas = list(map(float, czas))

#modyfikacja danych
nr_gr = ?
r_k = ?
np.random.seed(int(nr_gr*r_k))
ekspresja = ekspresja + np.random.rand(614, 7) * 0.1 - 0.05
```

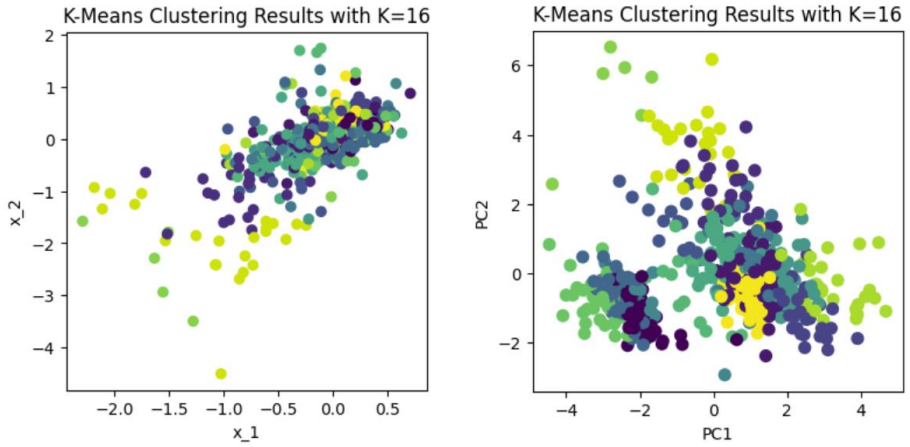
gdzie za `nr_gr` wstaw numer swojej sekcji a za `r_k` aktualny rok kalendarzowy.

Podejrzyj zmienne.

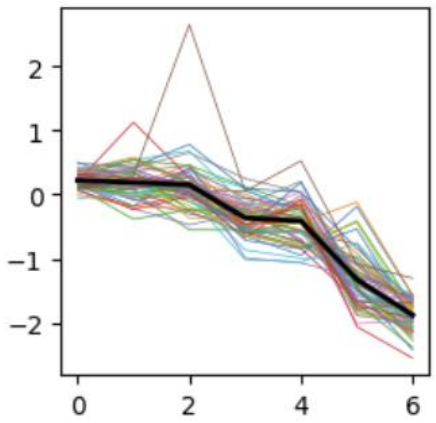
3. Zwizualizuj dane wykorzystując metodę PCA i pokaż wykresy wariancji wyjaśnianej przez poszczególne składowe PCA (wykorzystaj kod z komórek [50]-[52] rozdz. 12.5.4)

4. Pogrupuj profile ekspresji genów za pomocą metody *k*-średnich na 16 grup (patrz komórka [33] rozdz. 12.5.3).

5. Zwizualizuj wyniki w postaci wykresów punktowych – we współrzędnych oryginalnych (wybrane atrybuty, np. x_1 i x_2) oraz we współrzędnych PCA (np. pierwsza i druga składowa główna) ([35], [50], [51]).



6. Zwizualizuj profile należące do poszczególnych grup i centroidy tych grup.
16 wykresów postaci:



7. Wykreśl zależność wariancji wewnątrzgrupową (`kmeans.inertia_`) od liczby grup (`n_clusters`).

8. Wykreśl zależność wariancji wewnątrzgrupową (`kmeans.inertia_`) od liczby uruchomień algorytmu (`n_init`).

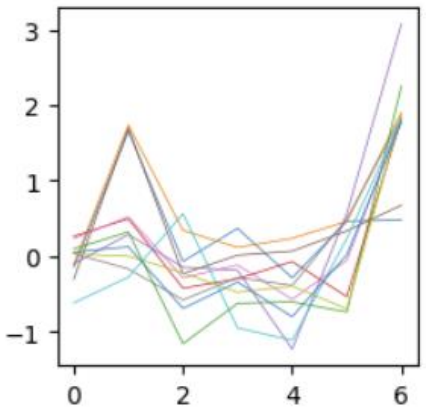
9. Wykreśl wykresy obrazujące wariancję wewnątrzgrupową (`kmeans.inertia_`) w kolejnych iteracjach dla pięciu różnych stanów początkowych (`random_state`). Przyjmij `n_init = 1` (pojedyncze uruchomienie algorytmu).

10. Pogrupuj profile ekspresji genów za pomocą metody hierarchicznej (patrz komórka [38] rozdz. 12.5.3).

11. Wyznacz przynależność do grup profili ekspresji, zakładając 16 grup ([43]).

12. Zobrazuj wynik dendrogramem (patrz komórki [41] i [42] rozdz. 12.5.3). Ustaw odpowiedni poziom odcięcia (`color_threshold`), zapewniający podział na 16 grup.

13. Zwizualizuj profile należące do poszczególnych grup.
    16 wykresów postaci:



14. Zobrazuj na wykresie słupkowych liczebności grup dla metody k-means i aglomeracyjnej.

15. Zobrazuj na wykresie wartości metryki silhouette (`silhouette_score`) dla obu algorytmów i różnej liczby grup.

16. Do której grupy utworzonej przez algorytm k-średnich trafi profil x = [-0.2205, -0.0041, 0.3821, 0.3680, 0.4918, 1.6983, 1.9820]? Narysuj ten profil na tle wszystkich profili z grupy, do której został przypisany. Wypisz nazwy wszystkich genów z tej grupy.

Co powinno znaleźć się w sprawozdaniu

A) Cel ćwiczenia.
B) Treść zadania.
C) Opis używanych w ćwiczeniu metod grupowania (nie kopiuj treści wykładu, poszukaj w literaturze i Internecie).
D) Metodyka rozwiązania – poszczególne instrukcje z wynikami i komentarzem (zachowaj numerację zadań).
E) Wnioski końcowe.

Zadania dodatkowe dla ambitnych

1. Zbadaj działanie algorytmu aglomeracyjnego przy różnych metodach obliczania odległości pomiędzy grupami. Powtórz p. 16 dla algorytmu aglomeracyjnego.
2. Wykonaj to ćwiczenie w innym środowisku, np. R, Matlab, ...

Przykładowe zagadnienia i pytania zaliczeniowe

1. Cel i plan ćwiczenia.
2. Materiał ze sprawozdania.
3. Problem grupowania danych.
4. Typy algorytmów grupowania danych.
5. Kroki algorytmu grupowania danych.
6. Miary podobieństwa/niepodobieństwa.
7. Funkcja celu w grupowaniu.
8. Algorytm *k*-średnich.
9. Grupowanie hierarchiczne.
10. Algorytmy grupowania oparte na gęstościach.
11. Metryka Silhouette.
12. Metoda analizy głównych składowych.

Do przygotowania na następne zajęcia

1. Zapoznać się z instrukcją do kolejnego ćwiczenia.
2. Zapoznać się z częścią teoretyczną do kolejnego ćwiczenia.
3. Wykonać zadania pomocnicze do kolejnego ćwiczenia.