Ćwiczenie 4

Grupowanie danych

Część teoretyczna

Wykład 9: Grupowanie 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 zmiennej `geny`, czasy pomiaru ekspresji zamieszczone są w zmiennej `czas`, a poziomy ekspresji genów dla siedmiu punktów czasowych (profile ekspresji genów) zamieszczone są w zmiennej `ekspresja`.

Zadania pomocnicze

Zapoznaj się z funkcjami `kmeans`, `subplot` , `linkage`, `cluster` i `clustergram` (help Matlaba).

Zadania do wykonania

Zadanie polega na pogrupowaniu profili ekspresji genów za pomocą metody *K*-średnich i grupowania hierarchicznego.

1. Wczytaj i zmodyfikuj zbiór danych:

```
load dane_genetyczne; %załadowanie danych
rand('state',nr_gr*r_k);
ekspresja=ekspresja + rand(614,7)*0.1-0.05; %modyfikacja danych
```

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

   Podejrzyj zmienne.

2. Pogrupuj profile ekspresji genów za pomocą metody *K*-średnich na 16 grup:

```
[grupy1, srednie] = kmeans(ekspresja, 16);
figure
for c = 1:16 %wykresy pogrupowanych profili
    subplot(4,4,c);
```

[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.

```
    plot(czas,ekspresja((grupy1 == c),:)');
    axis tight
end
suptitle('Grupowanie profili metodą k-średnich');
```

gdzie:
- `grupy1` to wektor, który zawiera numery grup, do których zostały przydzielone poszczególne przykłady
- `srednie` to wektory średnie reprezentujące środki grup (centroidy)

3. Narysuj centroidy reprezentujące poszczególne grupy (wykorzystaj kod z poprzedniego punktu).

4. Pogrupuj profile ekspresji genów za pomocą metody hierarchicznej na 16 grup:

```
grupy2 = clusterdata(ekspresja,'linkage','average','maxclust', 16)
figure
for c = 1:16 %wykresy pogrupowanych profili
    subplot(4,4,c);
    plot(czas,ekspresja((grupy2 == c),:)');
    axis tight
end
suptitle('Grupowanie hierarchiczne profili');
```

gdzie:
- `'average'` oznacza metodę średniego wiązania jako miarę odległości pomiędzy grupami
- `'maxclust'` oznacza liczbę grup
- `grupy2` to wektor, który zawiera numery grup, do których zostały przydzielone poszczególne przykłady

5. Narysuj dendrogram:

```
clustergram(ekspresja,'RowLabels',geny,'ColumnLabels',czas);
```

6. Na podstawie zmiennych `grupy1` i `grupy2` wyznacz liczebności poszczególnych grup. Wypisz nazwy wszystkich genów, które tworzą grupy jednoelementowe.

7. 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.
   Wskazówka: aby znaleźć grupę profilu x wyznacz odległości tego profilu od wszystkich centroidów (`d=dist(x,srednie')`). Centroid najbliższy wskazuje grupę, do której przypisany będzie x.

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 Matlaba z wynikami i komentarzem (zachowaj numerację zadań).
E) Wnioski końcowe.

Zadania dodatkowe dla ambitnych

1. Wyznacz rozrzut wewnątrzgrupowy *W* oraz rozrzut międzygrupowy *B* dla grup otrzymanych w ćwiczeniu. Jak liczba grup *K* wpływa na te rozrzuty?

2. Zaimplementuj grupowanie metodą *K*-średnich w innym środowisku, np. C/C++C#, Python, ...

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.

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.