

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 w odpowiednich komórkach macierzy. Badany materiał (próbka DNA) jest wyznakowany znacznikiem fluorescencyjnym i umieszczany na macierzy. Częsteczki 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 ćwiczeniu¹ 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

Zapoznaj się z 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/>)

Zadania do wykonania

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

...

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.