MATH/BIOINF/STATS 547: Mathematics of Data

Date: February 6, 2020

Due date: February 18, 2020

Problem Set 2: Least Square and Spectral Clustering Algorithms

For this assignment you will be performing linear least squares and the spectral clustering algorithms discussed in class. Sample data is provided for the least squares problem and data in the form of a weighted adjacency matrix will be used for spectral clustering. To complete this assignment, you will need to be familiar with least squares and the spectral clustering algorithms described in "A tutorial on spectral clustering" [1]. For more information on least squares, please see "Matrix Methods in Data Mining and Pattern Recognition". The full book can also be found on the SIAM website [2].

1. Assume that we want to determine the elasticity properties of a spring by attaching different weights to it and measuring its length. From Hooke's law we know that the length l depends on the force F according to

$$l = e + \kappa F$$

where e and κ are constants to be determined. Assume that we have performed an experiment and obtained the following data

F	1	2	3	4	5
l	7.97	10.2	14.2	16.0	21.2

Calculate the values of the constants e and κ by

   (a) Minimizing the sum of squares equation

   (b) Using singular-value decomposition (SVD)

   Remark: This example was obtained from "Matrix Methods in Data Mining and Pattern Recognition", Example 3.9 (page 31) [2].

2. Download the data available on Canvas in the Problem Set 2 folder. This data is described in further detail in Section "MATLAB data description" (below). Perform the 3 spectral clustering algorithms outlined in "A tutorial on spectral clustering" (page 399). Note that this matrix is in the form of a weighted adjacency matrix W.

   (a) Plot the results of this clustering (See Figure 1 for an example)

   (b) Which algorithm performed the best on this data? Which algorithm was the fastest?

   (c) What values for k did you choose and why?

   (d) Propose an alternate clustering method other than k-means and plot the results

MATLAB data description

This matrix is derived from data, and has been normalized to turn it into a weighted adjacency matrix. MATLAB variables available on Canvas are described below:

MATLAB variables:

- data_mat: a 777×777 weighted adjacency matrix derived from real data. To ensure the matrix is connected, rows and columns where more than 10% of the entries were zeros were removed from the matrix.

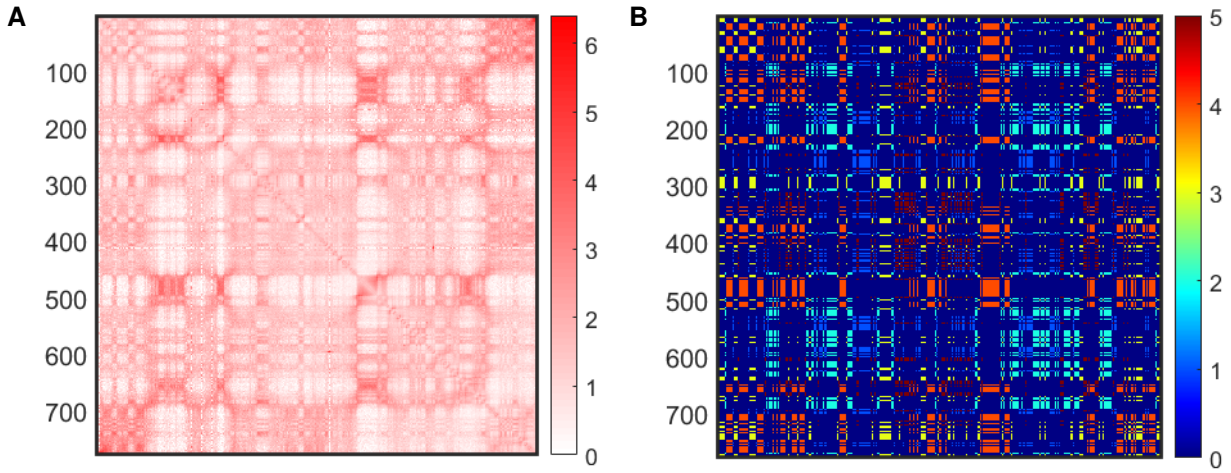

Figure 1: Weighted adjacency matrix and spectral clustering. (A) Weighted adjacency matrix plotted using `imagesc`. (B) Ng clustering algorithm results with $k = 5$, plotted using `imagesc`

References

[1] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[2] Lars Eldén. *Matrix methods in data mining and pattern recognition*, volume 4. SIAM, 2007.