AMATH 482/582: HOMEWORK 4

SATHVIK CHINTA

Abstract.

1. Introduction and Overview

2. Theoretical Background

We will use spectral clustering in the beginning to see if we can get reliable results. Spectral clustering is a method for deeply rooted in graph theory. In our case, each of the data points would get treated as a node in our graph. In order to perform spectral clustering, we will use a Laplacian matrix on our set of points. A laplacian matrix can be defined as

$$L_{i,j} = \begin{cases} \deg(v_i) \text{ if } i = j \\ -1 \text{ if } i \neq j \text{ and } v_i \text{ is adjacent to } v_j \\ 0 \text{ otherwise} \end{cases}$$

Equivalently, we can define two matrices: the degree matrix (D) and the adjacency matrix (A) and define the laplacian matrix with the equation:

$$L = D - A$$

We can then check the closely connected nodes which, in theory, should be closely related to each other as well. The components of the eigenvectors that correspond to the smallest eigenvalues of our constructed laplacian matrix, then, can be used to find clusters in our graph.

The second eigenvalue of the laplacian matrix is called the Fiedler value. The magnitude of this reflects how connected the graph is as a whole. We can select the Fiedler vector (the second eigenvector) and use the components whose value is positive to signify us predicting the +1 case (either democrat or republican) and the value who is negative to signify us predicting the -1 case (the other party).

3. Algorithm Implementation and Development

4. Computational Results

Plotting the sigma values (1000 values between 0 and 4) against accuracy for spectral clustering yields the following plot:

Date: March 11, 2022.

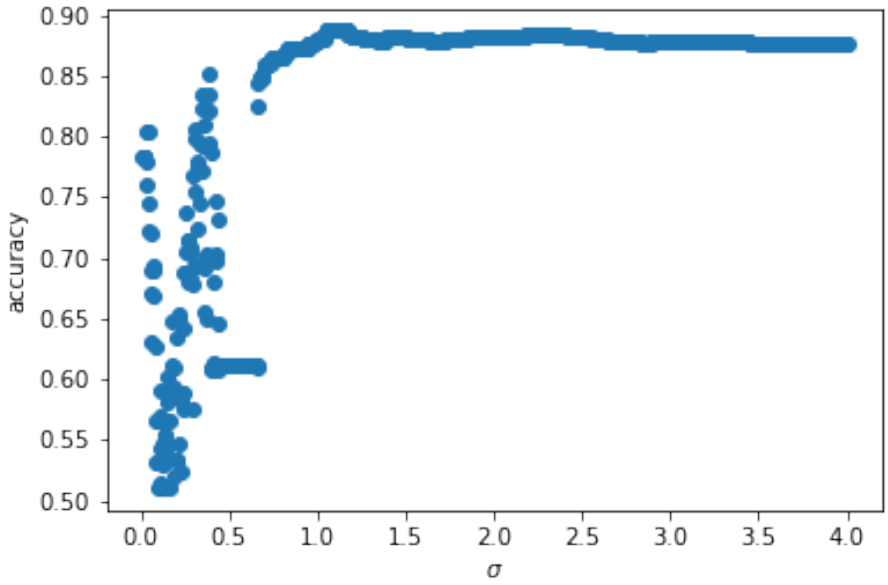

FIGURE 1. Plot of sigma values against accuracy for spectral clustering

We can see that the accuracy of the spectral clustering increases drastically after about $\sigma = 0.75$. Until that point, the accuracy seems to fall drastically at the smaller sigma values, then increase as well. After $\sigma = 0.75$, the accuracy seems to stabilize in the high 0.8's. The maximum accuracy occurs when $\sigma = 1.\overline{053}$. The accuracy at this point is 0.887, which indicates that even though our spectral clustering seems to be unstable, we are still getting a good accuracy.

When using the semi-supervised learning technique, we get the following results for different values of M and J.

J \ M	2	3	4	5	6
5	0.8827	0.8827	0.8620	0.8827	0.8827
10	0.8781	0.8919	0.8873	0.8873	0.8160
20	0.8850	0.8873	0.8988	0.8735	0.7816
40	0.8850	0.8942	0.9011	0.9264	0.9310

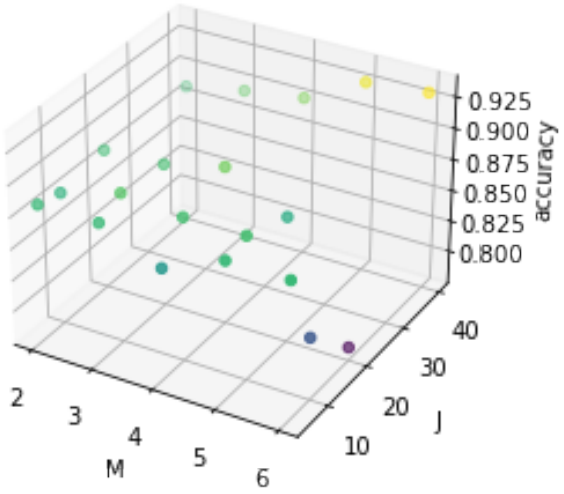

FIGURE 2. Plot of M and J values against accuracy for semi-supervised learning

We can see that the accuracy of the semi-supervised learning algorithm is almost as good, if not better, than that of the spectral clustering algorithm in most cases. The best accuracy occurs when $M = 6$ and $J = 40$. At this point, the accuracy is 0.9310. Curiously, the lowest accuracy occurs when $M = 5$ as well (though $J = 20$ in this case). This indicates we might be seeing behavior similar to that of spectral clustering, where even the slightest variation in our hyperparameters can lead to wildly different results. To test this theory, I plotted many values of M and J against eachother in order to see if there is any correlation between the accuracy and the values of M and J. The plot below shows the accuracy of the semi-supervised learning algorithm in this case

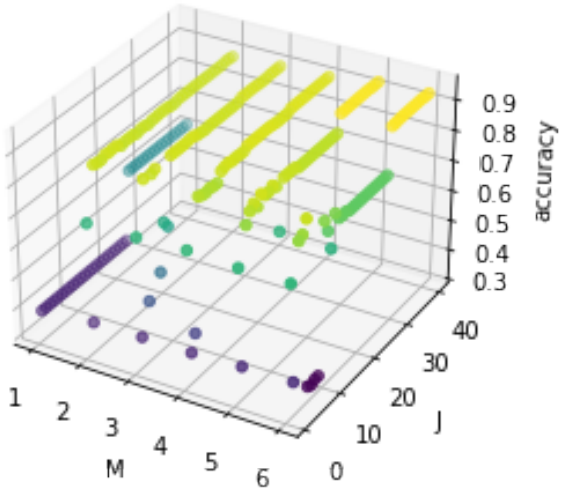

FIGURE 3. Plot of M and J values against accuracy for semi-supervised learning for a wide range of points

We can see that our hypothesis is almost correct. There does seem to be wild drifts in the accuracy, but not as erratic as we saw in the spectral clustering case. Instead, these seem to be linked to a drop of directly correlated to the J values. As J increases by a factor of 10, it seems that accuracy increases as well. As such, this methodology does seem to be much more stable (no more rapid oscillation between varying accuracies).

5. Summary and Conclusions

Acknowledgements