Spectral Algorithms to Analyze scRNA-seq Data

Eden Cahill, eden.cahill@sps.edu
St. Paul's School, Class of 2023
USC Viterbi Department of Computer Science, SHINE 2022

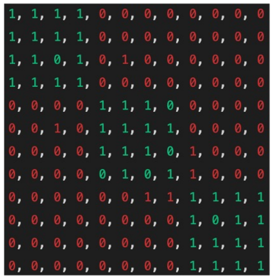
Introduction

Spectral algorithms such as PCA use the singular vectors of data to reduce dimensionality before clustering. In SHINE, I worked on understanding and implementing a dimensionality reduction algorithm called Principal Component Analysis (PCA). First, I used the Stochastic Block Model (SBM) to simulate data to understand the improvement of the clustering algorithm due to PCA. I specifically focused on the K-means++ algorithm. I did this to understand the power of PCA before moving on to the real life scRNA-seq data which is more complex.

Objective & Importance of Professor's Research

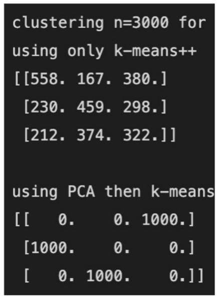
Professor Zhang's research is in the field of theoretical computer science. His research focuses on creating algorithms to cluster scRNA-seq data. Clustering this data can reveal rare cell populations, track cell development, and uncover relationships between genes. The knowledge provided by scRNA-seq analysis aids scientists in developing better treatments and in answering fundamental questions in biology. However, scRNA-seq data is difficult to cluster due to its complexity and large size. Professor Zhang's research is important because he is trying to create algorithms to address these challenges in order to be able to accurately cluster scRNA-seq data. Recently, Professor Zhang and my mentor Chandra Sekhar Mukherjee showed that PCA can be applied before clustering to significantly reduce the distance of data points belonging to the same cluster, while reducing inter-cluster distances mildly. Their research demonstrated that this improves clustering accuracy in scRNA-seq analysis.

How the Algorithms Work



```
1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0
1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0
1, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0
1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0
0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0
0, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0
0, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 0
0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0
0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1
0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1
0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1
0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1
```

Figure 1: SBM data
PC: Edie Cahill through VS Code



```
clustering n=3000 for
using only k-means++
[[558. 167. 380.]
 [230. 459. 298.]
 [212. 374. 322.]]

using PCA then k-means
[[   0.    0. 1000.]
 [1000.    0.    0.]
 [   0. 1000.    0.]]
```

Figure 2: Accuracy Matrices
PC: Edie Cahill through VS Code

I first generated data using the SBM. Figure 1 represents the SBM data for 12 nodes and 3 clusters. For this data, I chose the probability of intra-cluster connections to be 0.9 and inter-cluster connections to be 0.1. I found that k-means++ clustered the SBM data well when the probability of intra-cluster connections was much larger than the probability of inter-cluster connections.

However, as the probabilities got closer together (making the dataset harder to separate), k-means++ was less accurate. This is why I next implemented PCA to improve the accuracy of k-means++. Figure 2 shows the clustering results for SBM data for 3000 nodes and 3 clusters. Intra-cluster connection probability is 0.55, and inter-cluster probability is 0.45. The first matrix shows the clusters recovered without using PCA. This clustering is highly inaccurate. The following matrix shows the clusters recovered by using PCA then k-means++. This matrix shows a 100% clustering accuracy. The matrices show how PCA improves the accuracy of clustering when the probabilities of intra-cluster and inter-cluster connections are very close.

Skills Learned

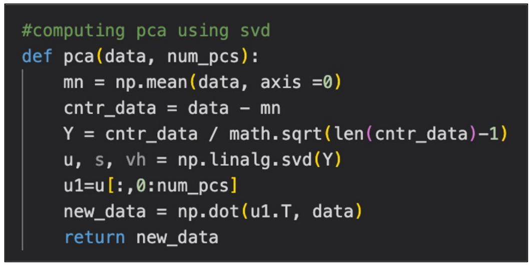
During SHINE, I have learned several skills that will help me in my future computer science pursuits. I have advanced my knowledge of Python. I gained experience in utilizing the Numpy library within Python which is helpful for matrix operations. I have also learned more about unsupervised Machine Learning and greedy algorithms with a specific focus on k-means++. Above all, I gained insight into how algorithms are created from theory and subsequently tested and improved.

Next Steps

In the future, I hope to study more greedy algorithms. My research focused on utilizing k-means++. I want to try this process again with a different greedy algorithm such as the Louvain algorithm. It will be interesting to examine if the results differ when using a different greedy algorithm on the same dataset.

How This Relates to My STEM Coursework

SHINE helped me have a renewed sense of the importance of my math coursework in school. I was introduced to the intersection of mathematics and computer science and how the two are inextricably connected. Mathematics is incredibly important for understanding and successfully implementing algorithms. Particularly, my SHINE research underscored the significance of linear algebra. I also learned how computer science can be applied to other STEM fields such as biology.



```
#computing pca using svd
def pca(data, num_pcs):
    mn = np.mean(data, axis =0)
    cntr_data = data - mn
    Y = cntr_data / math.sqrt(len(cntr_data)-1)
    u, s, vh = np.linalg.svd(Y)
    u1=u[:,0:num_pcs]
    new_data = np.dot(u1.T, data)
    return new_data
```

Figure 3: PCA Algorithm
PC: Edie Cahill through VS Code

Acknowledgements

I would like to thank Professor Zhang, my mentor Chandra Sekhar Mukherjee, my lab partner Elliot Bobrow, Michelle Emelle, Dr. Katie Mills, and the SHINE Team for their support and guidance.

References

[1] Kiselev, Vladimir Yu et al. "Challenges in unsupervised clustering of single-cell RNA-seq data." *Nature reviews. Genetics* vol. 20,5 (2019): 273-282. doi:10.1038/s41576-018-0088-9

[2] "PDL1 (Immunotherapy) Tests: Medlineplus Medical Test." *MedlinePlus*, U.S. National Library of Medicine, https://medlineplus.gov/lab-tests/pdl1-immunotherapy-tests/.

[3] VAN VU. A simple svd algorithm for finding hidden partitions. *Combinatorics, Probability and Computing*, 27(1):124–140, 2018.