Preserving Clusters with Dimensional Reduction

Carlos Manzano-Peralta [Cmanzano703@gmail.com]
STEM Academy of Hollywood, Class of 2023
USC Viterbi Department of Electrical Engineering, SHINE 2022

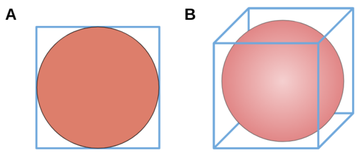
USC Viterbi School of Engineering

USC-Meta Center for Research and Education in AI and Learning

SHINE Summer High School Intensive in Next-Generation Engineering

Introduction

Visualizing high dimensional data is very important as it can illustrate the model being produced. Mapping high dimensional data-points into 2D and 3D while preserving the information between the high dimensional data points so that we can visualize them is an important research problem.



Objective

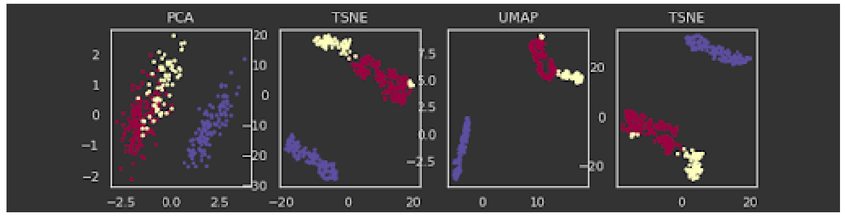
Single cell analysis (SCA) is an important technique to discover new cell types and gene expressions. Due to the multidimensionality of the datasets in SCA, it is hard to visualize and interpret.

Dimensionality reduction helps in easy visualization and interpretation of SCA data. Dimensionality reduction can also help in machine learning to embed objects in high dimensions (like multimedia data - images, audio, video, etc) into either 2D/3D for data exploration, clustering and discovering relationships in the data.

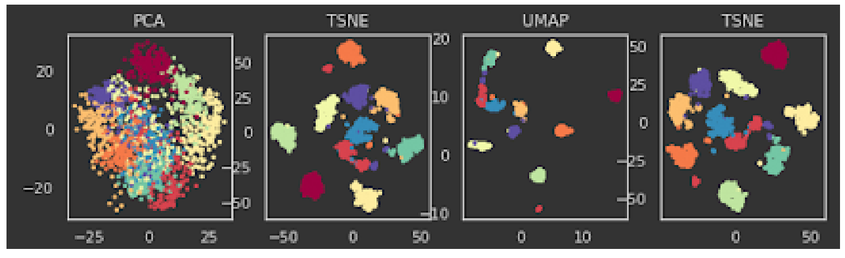
Skills Learned

Not only did I master my knowledge in the coding language, Python using the computational notebook named Jupyter but I also learned how to use programming libraries such as Numerical, Plotting, and ML (Machine Learning). Numerical libraries allow to use numpy and pandas commands which are used to assemble mathematical estimates on the graphs, ML libraries such as sklearn, UMAP, pacMAP are algorithms which produces variations of the dimensional reduction and Plotting libraries includes matplotlib and seaborn which embed the plot points.

Using the algorithms, PCA, TSNE , UMAP, and TSNE I was able to collect and produce two different data sets to compare and contrast,.



The image above is plot-points collected by the dataset, "penguin.data" provided by UMAP, the image below is plot-points collected by the dataset, "digits.data" which was produced by me and my merntor, Sampad.



Additional Info

Includes two examples of 16-dimensional plots being reduced down to 2-dimensional plots, the Jupyter code written in Python which produced the plot-points, and two research papers on why Dimensional Reduction is necessary.

Relations to STEM Coursework

Engineering / Architecture Elective - worked on soldering circuit boards throughout the school year which sparked my interest to explore the Electrical Engineering field specifically.

P.E.E.R Summer Program (In collaboration with CalTech & Kaiser Permanente) - learned Python, worked in a Bioengineering lab, and learned the process of Single Cell Analysis and how to transcribe cell clusters.

Next Steps

After having the chance to work within the lab setting, I've concluded of wanting to major in Computer Engineering as it's the best of both worlds: Computer Science and Electrical Engineering.

Acknowledgements

I want to first thank my mentor, Sampad Mohant for his tremendous help of furthering my knowledge of Single Cell Analysis and exploring the ideas of what it takes to become an Electrical Engineer, Professor Bhaskar Krishnamachari for giving me the opportunity to work within his lab, My advisory teacher and engineering teacher, David Trachtenberg and Lindsay Weitzel for motivating me to pursue SHINE, My center mentor Dr. Mills for her support throughout these 7 weeks, Professors Meisam Razaviyayn and Murali Annavaram for choosing me as a recipient of a USC Meta Scholarship for the generous opportunity to attend SHINE at no cost.