

Introduction to Density-based Clustering and its Applications

Name: Dongfeng Li

Mentor: Yikun Zhang

In this DRP I did my first unsupervised learning project. It is modeled based on the data from my local power plant desulfurization system. In this project, I built two models of clustering. One of them is the most classic method in centroid-based clustering; the other is the most classic method in density-based clustering. Also I used the variable selection technique to clean the data and deduce the dimensionality of data points. To visualize the result, multidimensional scaling(MDS) and pairplot are introduced and utilized. The modeling was written from scratch while the following techniques are using packages from different resources of Python.

Those would be the methods I utilized in my project in general. For knowledge I learned in this process would be much more. Many possible concepts that are related to this project were checked but not drawn. When considering dimensionality reduction, principal component analysis was also considered but not used since we want to have more scattered data patterns. The variable selection technique could select those data with larger variance which means more scattered. Also for visualization, there is an interesting technique I found in a paper that utilized a minimum spanning tree as a coloring tool.

During this program I not only learned about knowledge, but also how to do the research from scratch. Since this topic is quite different from my mentor's original project, many specific technical details could not be directly answered by Dr. Zhang. Online resources and papers are essentially important. I read through 5 papers, over 50 blogs, and hundreds of web pages to figure out some coding technique or mathematical reasoning. Also, reviewing contents from stat 416 helps a lot. These are great harvests from this project.

Finally, I would like to thank Dr. Zhang. He is indeed a knowledgeable and kind person that gave me plenty of useful information and reading materials. And the final presentation provides me with a great platform to learn from my peer's project.