DRP Program Writeup

Guanhua Chen

Networks have become a common data structure with which to understand many of the concepts of everyday life. In the research field, we also encounter data with network structure, such as modelling communities of neurons in the brain. Thus, I have two goals for attending this DRP Program: 1) learn about what is the Network Machine Learning and what is the difference between network-valued data and tabular structure data. 2) Find an interesting dataset in the network structure and then do some hands-on machine learning towards it through Jupyter Notebook.

The textbook we used is "Hands-on Network Machine Learning with Scikit-Learn and Graspologic". Its goal is to give us the concepts, the intuition, and the tools we need to actually implement programs capable of learning from network data. It covers the fundamentals of network machine learning, focusing on developing intuition on networks, and doing while paired with relevant Python tutorials.

What is network machine learning? The network has information (given by the edges), but traditional statistics and machine learning are not meant for networks. I learned that the problem we run into is that networks are not, in their most naive form, tabular data, but network-valued data. Loosely, network machine learning is machine learning for network-valued data. One solution is that we can adapt the network to a more traditional format: tabular structure. Then we can then apply regular techniques from statistics and machine learning.

Here are some key concepts for the network machine learning that I have learned. Adjacency Matrix is a square matrix with a row and column for each node. A 1 in the matrix at row i and column j means an edge in the graph between nodes i and j. We use embedding to convert the network to a tabular format. Spectral Embedding estimates a tabular representation (called latent positions). We embed the adjacency matrix into a low dimensional representation. This is done using linear algebra (singular value decomposition). The k-means clustering algorithm is a partitioning method used to group a set of data points into k clusters based on their similarity.

I found network data about the US Airline in 1997. The first two columns are nodes, which indicate there is an edge between those two edges. The third column indicates the weight of the edge, which might be the frequency. I convert the data into an adjacency matrix. And then embedded the adjacency matrix and use network plot for first two dimensions of this embedding to see the position of these airports. I also used K-means on the network adjacency matrix embeddings to cluster the airports into two communities.

I am honored to be a part of this program and to have the opportunity to talk with and ask advice from my mentor, Ronan. I will continue to read the book and learn more deeply in this direction.