Finding Communities within the Citibike Network

Shay Lehmann and Michelle Ho

CUSP-GX-5006
Prof. Gustavo Nonato
Machine Learning Assignment # 4

Abstract

In this assignment, the authors explore different clustering algorithms as a way to find communities within the Citibike network. In this directed network, the nodes are station locations and the edge weights are the count of trips over the given time period of data. The clustering techniques used are K-means, Spectral Clustering, and Hierarchical Clustering.

1 Introduction

Detecting communities within networks is one application of machine learning clustering algorithms. From making recommendations within a social networks to predicting the spread of disease, there are many applications for community detection.

2 Methods and Data Sets

The dataset chosen for this assignment is Citibike trip data from July 2013 to February 2014, as well as the station locations for that same time period.

- 'tripduration',
- 'starttime',
- 'stoptime',
- 'start station id',
- 'start station name',
- 'start station latitude',
- 'start station longitude',
- 'end station id',
- 'end station name',
- 'end station latitude',
- 'end station longitude',
- 'bikeid',
- 'usertype',
- 'birth year',
- 'gender'

The steps taken for this assignment:

1. A simple k-means algorithm is used to cluster the stations purely on station location information
2. The network graph is then used to generate spectral clusters
3. Hierarchal clusters are also generated based on the network graph

4. Parameters for the distance metric, number of clusters, and tree pruning are altered

5. The results are visualized on a map by coloring the station locations by their community labels and with corresponding dendrograms for hierarchical clusters

3 Results

4 Conclusions

In summary...

Python code used to generate the results, tables, and figures for this assignment can be found at `https://github.com/michellemho/machine_learning_for_cities`.

Figure 1: Model 7