Cooper Sloan

May 19, 2016

Predictive Models for Bus Arrival Times

Have you ever found yourself sprinting to catch a bus, or waiting for 20 minutes in the snow for a bus which never shows up? Bus arrival times depend on countless factors, including weather, traffic accidents and Red ox games which make them unreliable for riders and difficult to predict in general. To that end, the ability to accurately predict arrival times and provide them in real time to transit users would result in more efficient transportation and higher consumer satisfaction.

The following is a histogram I generated from MBTA data scraped from Nextbus.com, which posts real time bus location data from every bus in the MBTA network. It visualizes the difference in arrival times of buses at the Massachusetts Ave and Newbury t stop in Boston for the week of March 28 to April 3, 2016. Most stops show the same general distribution. In a perfect network, busses would arrive at regular intervals, and the entire data would be concentrated at around 10 minutes (dotted black). At the other extreme, if buses were completely random and decoupled, the data would follow a Poisson distribution (shown in green). An exponential model (shown in blue) seems to capture the data near zero. However neither of these models fit the data as a whole. The fact that the data shows a spike at zero is largely unexplained, and suggests that simple statistical models may not be sufficient to explain bus arrival times.

Several statistical techniques have been used across various cities to predict bus arrival times. Research suggests that simple regression techniques can be used to reduce the average waiting time for passengers. Additionally, machine learning techniques including artificial neural networks with adaptive algorithms have shown to be very effective and predicting arrival times

based on recent travel data. tudies have been conducted across the world including several cities in the U and Canada as well as China and Brazil. In general, machine learning models have shown the best accuracy over historical data based models and regression models. Even so, there is little research which analyzes the effectiveness of different models over the same data set. Part of the problem is that current tools for analyzing complicated mixture models are insufficient.

I would like to explore both regression as well as machine learning methods with mixture models and determine how to best describe the bus network system. pecifically, I hope to compare several statistical models to fit the data and inject some knowledge of those models into machine learning algorithms to achieve better results than either method could do on its own. The MBTA data set is unique in both its size and the time interval (several years) of data available. ome factors which I would like to explore are historical as well as instantaneous data, to see if arrival time can be dynamically predicted given the most recent data from the network. Kalman filtering has been used in this capacity for car travel times with limited success, but a combination of historical and real time data may show more success. I would also like to develop tools necessary for using mixture models with existing machine learning algorithms. Using mixture models on transit data would be a novel contribution to the existing research. My overall goal would be to determine which factors are most useful in modeling and predicting arrival times, and analyzing the possible accuracy and benefits of such a model.

Ultimately, being able to accurately predict bus arrival times would enable more effective and enjoyable transportation. The current research fails to compare different modeling methods and does not utilize mixture models. Understanding the underlying model governing bus transit networks would allow transit agencies to design more efficient routes and may lead to insights about queueing networks in general.