Predicting Bus Arrival Times Using Neural Networks

Cooper Sloan, csloan@mit.edu

Advisor: Jiahao Chen, jiahao@mit.edu

Abstract

Current public transport systems have unpredictable schedules, making them unreliable and inconvenient for passengers. Research has shown than by modeling bus networks, it is possible to deliver passengers useful information about bus arrival times. This research will focus on using larger neural networks with mixture models to generate more accurate predictions than are possible with smaller models. I hope to accurately predict bus arrival times in order to increase network efficiency and reduce transit operating costs.

Problem Statement

Commute times in the US have steadily increased over the past several decades, from a 22 minute average in 1980 to over 25 minutes is 2000, despite an increase in transport capacity [1]. Part of this problem stems from the fact that works primarily commute alone in private automobiles, with on 10.8 percent using public transportation. Ironically, Americans support increased public transit spending at a rate of 70% according to an 2013 American Public Transportation Association survey. Increasing usage of public transport is the key to decreased pollution, reduced traffic, and overall lower costs for consumers. In order to close the gap, public transport needs to be as appealing as driving a private automobile. One of the major obstacles standing in the way is the unpredictability of public transit arrival times. Commuters need a reliable schedule in order to rely on public transit.

[1] McKenzie, Brian, and Melanie Rapino. *Commuting in the united states: 2009*. US Department of Commerce, Economics and Statistics Administration, US Census Bureau, 2011.

In my project I will focus on predicting bus arrival times to decrease wait times for passengers. Current research shows that prediction models can be valuable to passengers. My research will tackle the problem of predicting network traffic. This has applications outside of public transportation including in communication and information networks.

Related Work

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[2]. Studies have been conducted across the world including several cities in the US and Canada[3] as well as China[4] and Brazil[5]. Even so, there is little research which analyzes the effectiveness of different models over the same data set. 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[6]. In general, machine learning models have shown the best accuracy over historical data based models and regression models[7]. A study using bus data from Houston Texas compared the performance of historical data based models with neural networks, and found that the neural networks performed better. This suggests that machine learning can certainly be applied, however, a combination of historic data as well as more complex models may produce the best accuracy. Existing work in the field fails to integrate these different types of models.

---

[2] Patnaik, Jayakrishna, et al. 2004. Estimation of Bus Arrival Times Using APC Data. Journal of Public Transportation, 7 (1): 1-20.
[3] Ibid. 2
[4] Chien, S., Ding, Y., and Wei, C. (2002). "Dynamic Bus Arrival Time Prediction with Artificial Neural Networks." J. Transp. Eng., 10.1061/(ASCE)0733-947X(2002)128:5(429), 429-438.
[5] Weigang, Li, et al. "Algorithms for estimating bus arrival times using GPS data." Intelligent Transportation Systems, 2002. Proceedings. The IEEE 5th International Conference on. IEEE, 2002.
[6] Ibid. 4
[7] Jeong, Ranhee, and Laurence R. Rilett. "Bus arrival time prediction using artificial neural network model." Intelligent Transportation Systems, 2004. Proceedings. The 7th International IEEE Conference on. IEEE, 2004.

The most famous study is out of Cuernavaca, Mexico, which found that the Gaussian Unitary Ensemble stemming from random matrix theory seems to fit bus arrival time data well[8]. Microscopic models have been used in order to explain the phenomenon[9]. These ideas however have not been incorporated into predictive models partly because current tools for analyzing complicated mixture models are insufficient.

Despite the breadth of research in the area, most of the studies use relatively small models. As computational power has increased in the past several decades, machine learning has allowed the use of larger and more complex models. The ability to use larger models has revolutionized entire fields of computer science, particularly computer vision. I would like to leverage the use of larger models, in particular neural networks with nonlinear mixed models. Additionally, the current work deals on the time scale of months, and the MBTA data set is unique in that it has several years worth of data.

Technical Approach

Given an accurate model of the transit system, passengers can be informed about bus arrival times and therefore decrease their wait time. The MBTA dataset contains several years of GPS data for all the buses in Boston. For my modeling, I will focus on the 1 bus which travels between Harvard Square and Dudley Station. By comparing the GPS data with the route information posted on the MBTA website, I will extract information including the arrival time at each stop, how late or early each bus was, the wait time at each stop, and the interval between adjacent stops. These will be my primary input features to the neural net (Figure 2). This

[8] Krbálek, Milan, and Petr Seba. "The statistical properties of the city transport in Cuernavaca (Mexico) and random matrix ensembles." *Journal of Physics A: Mathematical and General* 33.26 (2000): L229.
[9] Baik, Jinho, et al. "A model for the bus system in Cuernavaca (Mexico)."*Journal of Physics A: Mathematical and General* 39.28 (2006): 8965.

particular bus has 28 stops including 9 time check stops for which schedule times are posted.

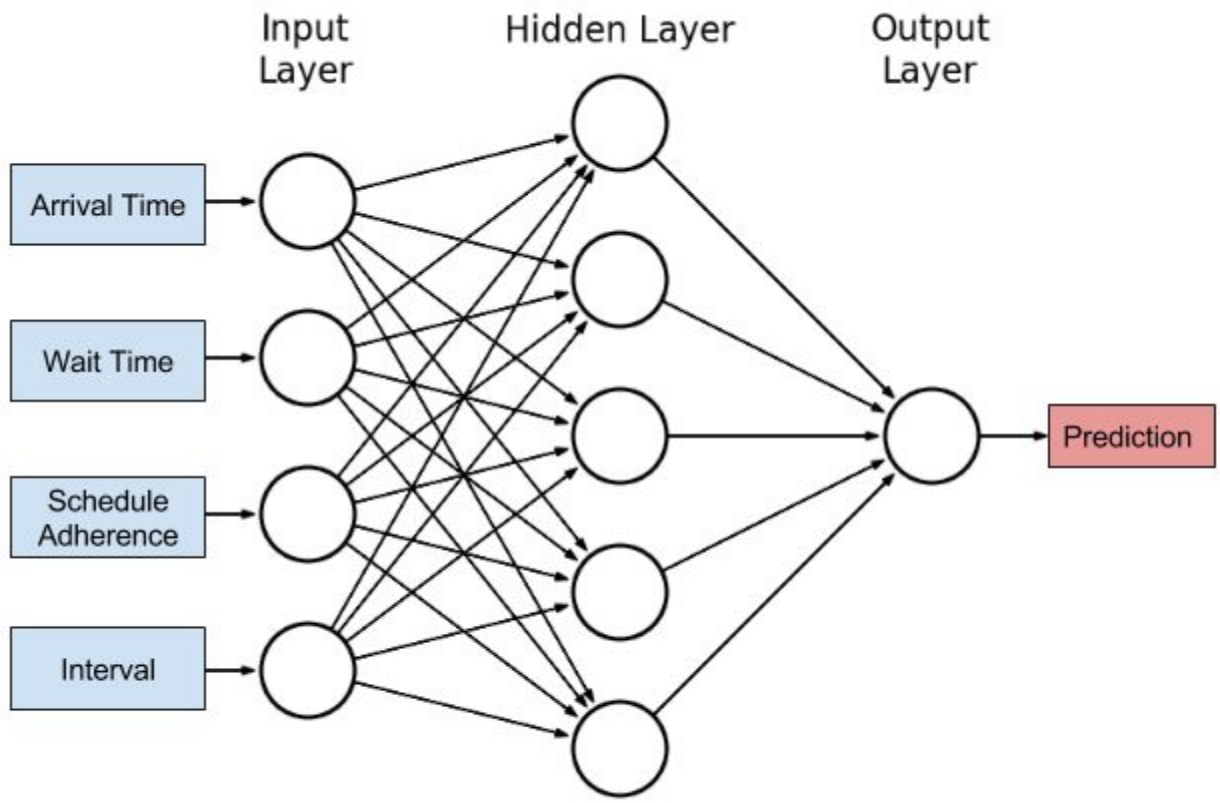

Figure 1: Neural net for predicting bus arrival times given data for previous stops

Model generation will consist of two stages. In the training stage the model will be initialized and will generate predictions for each of the input vectors. Predictions from the model can then be compared to the actual arrival time in order to backpropagate the signal and accordingly update the model. After the training stage is complete I will validate the model via another set of input vectors for which the actual arrival time is hidden. This stage will be used to estimate the accuracy of the model on novel data.

The following is a histogram (Figure 2) which I generated from the MBTA dataset. It visualizes the difference in arrival times of buses at the Massachusetts Ave and Newbury St 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 suggesting that simple statistical models may not be sufficient to explain bus arrival times.

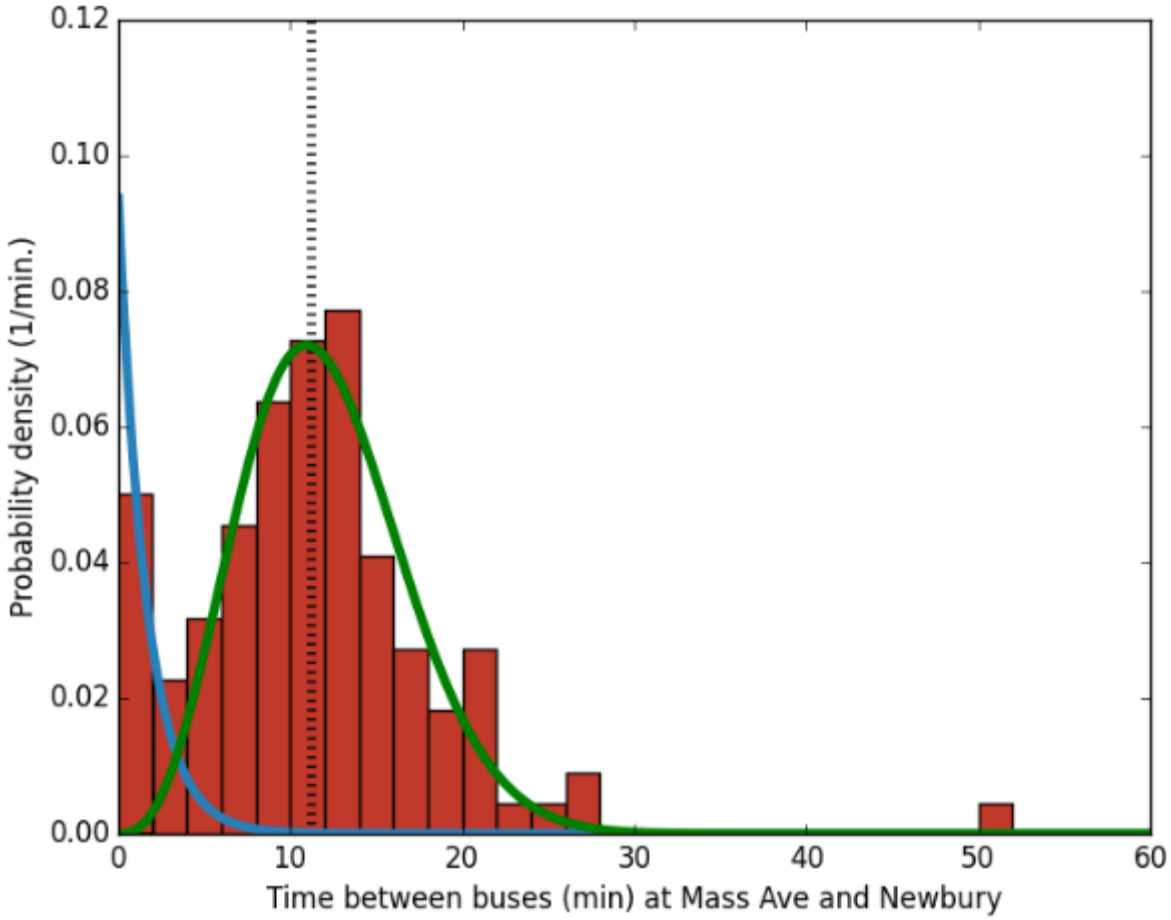

Figure 2: Histogram of the time interval between buses for the Mass Ave and Newbury stop. The red bars indicated frequency for each interval. Several possible statistical distributions are overlaid.

I would like to inject a few different statistical models into the neural net through mixed models. Specifically I will use Poisson distribution and Gaussian Unitary Ensemble (GUE) and evaluate their accuracy. The Poisson distribution is natural for a stochastic process like bus arrival times, and the GUE Ensemble shows promise as indicated by the Cuernavaca study. Nevertheless, the GUE has not be incorporated into a neural net by any current study. Part of the reason may be the lack of access to tools for these more complicated models. As a result, I will be implementing the algorithms myself in Julia.

I will cluster data points according to time of day, because the 1 bus has different schedules during peak hours. I will independently generate the models be each of the periods of the day, and compare this with the performance of the unclustered model. Due to the size of the MBTA data, I will be able to increase the number of parameters in my model with less fear of overfitting.

Evaluation Plan

My goal is to do accurate prediction on bus arrival times, so the key metric I would like to measure is error from the predicted to the actual values. Specifically I will use mean square error and mean absolute error which are appropriate in this case. I will also look into of other accuracy measure including AUC and F1 score, although this will require choosing cutoffs for prediction validity. Because this is a machine learning problem, I will measure the training as well as validation error, obviously with the emphasis being on validation error. The error will be a function of several factors, one of which is the model. I want to standardize a training and validation set so that I can compare models. Additionally I would like to measure how different sizes of training data affect the validation accuracy. Finally I want to look into the computation time for each of the models. In order to predict bus arrival times and pass these predictions on

to the passengers, model prediction needs to be fast. While there are many other factors I could measure I will limit my scope to these three factors.

In order to measure the error, I will need to frame my models in a way which generate predictions for bus arrival times in a structured way. In order for the error comparisons to be valid I am going to identify a subset of the MBTA data for training and another subset for validation. I want to make these sets rather large in order to avoid overfitting. As long as my models can generate predictions for the input set in a consistent way, it will be simple to measure the error for each model and compare them. Some of my models may not require the training data, and in that case I will compare their error to the validation error of the other models.

Timeline

- September and October: Collect schedule data from MBTA and process GPS dataset to extract input features
- November and December: Begin designing model and implementing learning algorithms in Julia to be able to generate models
- February and March: Evaluate different modeling techniques and collect accuracy/performance data
- April and May: Assess utility of models based on accuracy on the validation set and computational performance and draw conclusions

Conclusion

Ultimately, being able to accurately predict bus arrival times would enable more effective and enjoyable transportation. The current research fails to utilize larger models and does not incorporate 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.