Optimizing Public Transit

Mindy Huang
Christopher Ling
CS229 with Andrew Ng

1 Introduction

Most applications of machine learning deal with technical challenges, while the social sciences have seen much less use and give more surprising results. Therefore, for this project we chose to study what comprises an optimal light rail transit system. We chose light rail systems (subways, bus rapid transit, etc.) because they are generally not tied to traffic, are costly to implement and therefore restricted to a limited number of stops at key destinations, and are the current focus in transit planning.

First, we model the ridership of a system in a city. Then, we use our model to generate a light rail transit system that optimizes for ridership.

2 Model

We use linear regression to predict ridership a transit system. First, we look at each individual station within the city and its features (demographics of the area, points of interest nearby, etc). Let x_{ij} be the j-th feature of the i-th station. Then we represent the i-th station like so:

$$\theta_1 x_{i1} + \theta_2 x_{i2} + \cdots + \theta_m x_{im} = \sum_j \theta_j x_{ij}$$

Then to get a "score" of how good the transit sytem of the city as a whole is, we sum each station together.

$$\text{score} = \sum_i \sum_j \theta_j x_{ij} = \sum_j \theta_j \sum_i x_{ij}$$

We then add in features of the transit sytem as a whole (average distance between each station, etc). Let y_i represent each feature of the transit system as a whole. So the model of the entire ridership system is:

$$\text{ridership} = \underbrace{\sum_j \theta_j \sum_i x_{ij}}_{\text{features of stations}} + \underbrace{\sum_i \phi_i y_i}_{\text{features of system}}$$

2.1 Features of each station

At each station, we have the following kinds of features:

- Demographics - we obtained demographics data on the area around each station from the 2010 Census. Features include median age, race, employment status, average transit duration to work, etc.
- Points of interest - we have features for the count of each place type within a 10 mile radius of each station. In other words, x_1 is the number of parks within 10 miles, x_2 is the number of sports stadiums with 10 miles, etc.

Some interesting findings - having airports, parks, and universities nearby correlates with high ridership. Also, working further away correlates with low ridership. We also found that having art galleries near stations correlates with higher ridership, possibly due to the influence of New York City.

2.2 Features of the system as a whole

Currently, we only have one feature for the system as a whole - average distance between each station. However, this is extremely important for the second step of our project - when we generate our own transit system that maximizes ridership. If we do not have a parameter that monitors the distance between each station, then maximizing ridership of the system is trivially putting an infinite number of stations on the same point.

Features we plan on implenting in the future include

- the number of stations in the system per unit of area the that the transit system covers
- the coverage of the city the system serves
- a cost function to control the jagged-ness of station placement, to mimic the routes of an actual transit system

2.3 Fitting the model

To determine our parameters θ and ϕ, we implemented both normal equations and gradient descent. As this was an ill-conditioned problem, the normal equations failed to converge, so we continued with gradient descent.

3 Optimizing the model

Our greatest obstacle was the lack of training data - there are only 25 cities in the US that have implemented light rail systems. As such, we implemented a lot of techniques to optimize our error.

- **Regularization** - Since we had such a small data set, it was imperative to regularize to prevent overfitting. We added a penalty to our model to smooth it out

  $$\text{ridership} = \theta X - \lambda||\theta||_2$$

  Regularizing brought down our error by about 10 percent.

- **Leave-one-out cross validation** - we used this to get a more accurate estimate for the generalization error of our model, as well as to optimize our regularization and gradient descent coefficients. We found that $\alpha = 1.38 \times 10^{-8}$ and $\lambda = 0.021$ optimized our general error. This improved our error by about 20 percent.

- **Feature selection** - we used feature selection to remove features that increased error. We found that having Hindu temples, locksmiths, and taxi-stands a part of our algorithm all increased general error, most likely because they occur randomly with no real correlation to ridership. This improved our error by about 10 percent.

- **Logging** - Our data set was extremely jagged - numbers ranged from less than 1 when we looked at percentage race to tens of thousands for median income. Therefore, to smooth it out we took the log of the numbers. This worked surprisingly well, and brought our error down by over 100 percent.

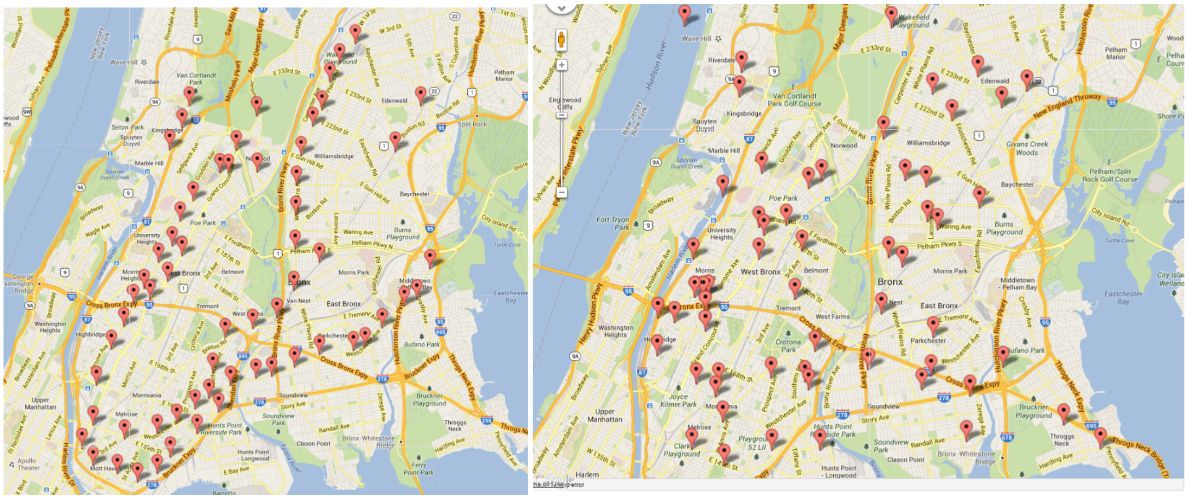
With all of our optimization techniques, our estimated generalization error decreased to 30 percent.

4 Generating a Transit System

After fitting a reasonably accurate model, we moved on to the second step of our project - given data of a city, find locations that optimize the ridership of the transit system if stations were built there. To do this, we discretized a city into blockgroups (a unit of geography used by the Census) and implemented a greedy algorithm that selects the blockgroups maximizing ridership according to our model. The algorithm runs until adding more stations begins to decrease ridership.

Below are comparisons between the actual transit systems and our optimal station locations according to our model. We generated models for The Bronx in New York City, and Austin. Since New York City has high ridership, when we run our model we would expect our algorithm to output something similar to the existing system. And since the system in Austin has low ridership, running our model in Austin should output a system that is fairly different.

Figure 1: Comparison of **The Bronx**. Approximately 320 rides/person/year



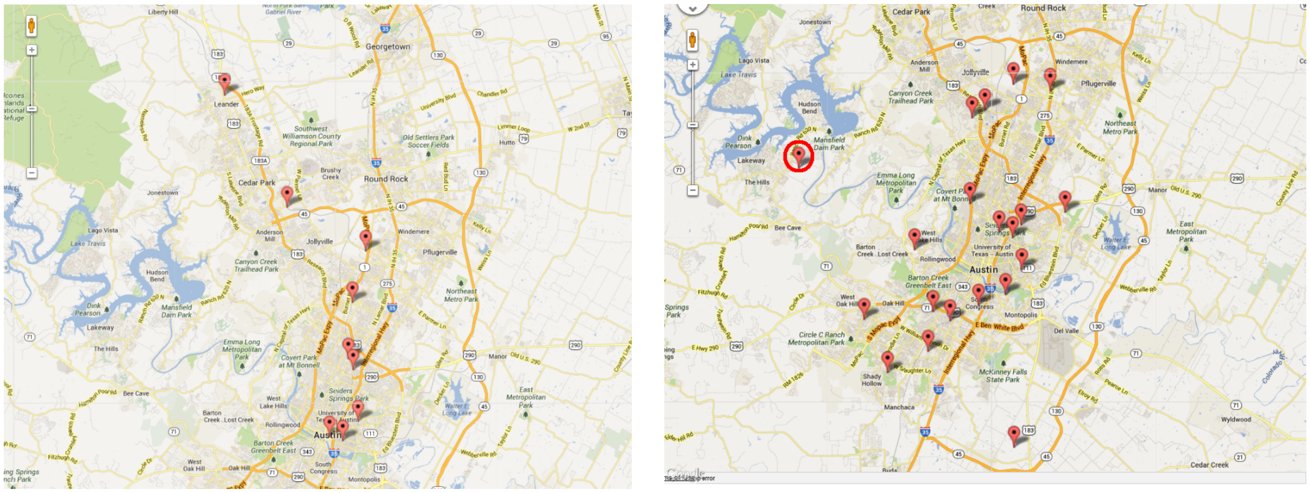
(a) Actual system in New York

(b) Our generated system in New York

Analysis of New York:

As seen above, our generated system creates stations in similar locations and hot spots to the current system, as expected. However, our stations are relatively clumped and jagged compared to the actual station locations, for two reasons - 1. we did not factor in a cost function to smooth the stations into distinct routes, like a real system would have; and 2. since our discretization was based on block groups, the locations to choose from, were not even in shape and not fine-grained enough.

Figure 2: Comparison of **Austin**. Approximately 0.34 rides/person/year



(a) Actual system in Austin

(b) Our generated system in Austin

Analysis of Austin:
Our model of Austin had several more stations than the current system. Most of them are along the heart of the city, and the outliers of the original system were done away with. Of particular interest is the station circled in red - note that it is next to several parks. This shows that in our model, parks correlates highly with ridership, which is fairly intuitive.

5 Assumptions and Restrictions

Due to the inherently fuzzy nature of the subject, we make a few assumptions to simplify our model.

- Transit culture - we assume that transit culture is the same across all the cities we train on (we assume that Angelinos are just as willing to take public transit as New Yorkers if the transit system is optimal).
- City boundaries - we assume that transit systems are self-enclosed within each city. In reality, they often cross boundaries.
- Cost - when we create our own transit system at the end, we do not take into account the cost, monetary or political, of erecting a station at a given point. This is because political battles are extremely hard to quantify, and probably the reason the station was not built at that point in the first place.

6 In Conclusion

We were pleasantly surprised at the outcome of our project - we did not at all expect such a good model, given the fuzziness of the problem and the simplicity of our model. Moreover, this project demonstrates the potential applicability of machine learning to the social sciences. In the future, we hope to further improve the model, and perhaps provide new insights into public transit systems.

7 Sources

- The American Public Transportation Association provides information on revenue and ridership per transportation district
- Google's public transit feed provides the location and type of each station
- Google Places provides the points of interest around each station
- The 2010 Census and American Community Survey provides up-to-date information on demographics, population, age, and income. Retrieved from National Historical Geographic Information System at the University of Minnesota.

Special thanks to Jeffrey Barrera, TA for Urban Design, and Peter Brownell, PhD., Reasearch Director at the Center on Policy Initiatives. And Dave, a super helpful TA.