Demand Prediction in Bike Share Systems

Zhaonan Qu
zhaonanq@stanford.edu

Objectives

Bike Share systems are becoming increasingly popular in urban areas. A major challenge in their operations is the unbalanced demand and supply at bike stations as a function of time. A quantitative, predictive would help operators dispatch bike transports more efficiently. This project aims to build a such a model for bike arrivals at and departures from stations, given information on:

- day of the week
- time of day
- temperature
- precipitation

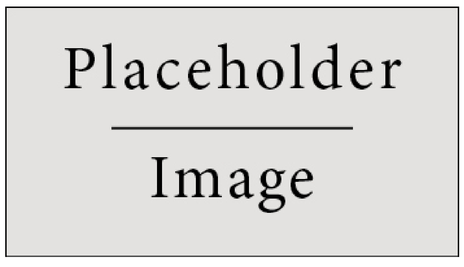

Figure 1: January 2017 Citi Bike Usage in time intervals of 30 minutes.

Data

Data provided by Citi Bike in New York City. It contains trip duration, date and time, as well as coordinates of stations. It also contains user-specific information such as age and gender. Because we would like to predict demand and supply of bikes, we bucketize time and aggregate the number of trips within 30 minute intervals. There are around 70000 data points in the training set and 12000 data points in the test set.

Features

The raw features provided by the dataset that are used include locations of stations and date and time of trip. We combined historical weather data from NOAA's database, which includes maximum and minimum daily temperature and precipitation. Our code also determines if a given date is weekend or a federal holiday.

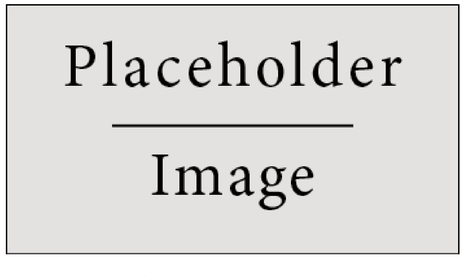

Figure 2: A section of processed data. We bucketized time and aggregated number of bikes departing from each station.

Neural Network

We apply a simple neural network to train the model:

$$o = \sum_{j=1}^{n_1} w_j^{[2]} h_j + w_0^{[2]}$$
$$h_j = g(\sum_{i=1}^{n_0} x_i w_{i,j}^{[1]} + w_{0,j}^{[1]})$$

where n_0 is the number of variables and n_1 is the number of neurons in the hidden layer. Here g is a nonlinear function such as ReLU or logistic. In practice we implemented the DNNRegressor provided through TensorFlow, with two hidden layers each with 100 neurons.

Baseline Model: Linear Regression

We focus on the problem of predicting bike demand during morning peak hours on work days. We performed linear regression on variables: time bucket, coordinates of station, max and min temperature of that day, and precipitation. We used mean squared distance as loss function, and trained the model with L2 regularization:

$$\ell(\theta) = \frac{1}{m} \sum_{i=1}^{m} (\theta x^{(i)} - y^{(i)})^2 + \alpha \|\theta\|^2$$

We implemented the linear regression in TensorFlow with Adam optimization algorithm.

Results

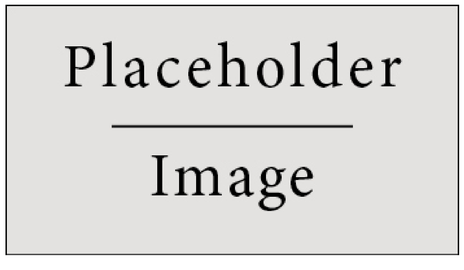

Figure 3: Training loss as a function of epochs. To be added later.

Here are the training errors for our two models.

Treatments	Linear Regression	Neural Network
Training Error	0.02	0.0002
Test Error	0.05	0.008

Table 1: Mean squared error of training models.

Conclusion

We see that weather is the most predictive parameter for demand of bikes. Moreover, neural network performs better than linear regression, as expected. Our next step is to investigate whether other parameters could improve the predictive power of the model.

References

[1] J. M. Smith and A. B. Jones.
Book Title.
Publisher, 7th edition, 2012.

[2] A. B. Jones and J. M. Smith.
Article Title.
Journal title, 13(52):123–456, March 2013.