PLACEHOLDER
LOGO

Ride Hailing Supply and Demand Forecasting using Didi-Tech Dataset

{ Gopal Menon and Kimberly Williamson } University of Utah Data Mining Spring 2017

Introduction

Didi Chuxing is the leading ride hailing company in China and processes over 11 million trips, plans over 9 billion routes and collects over 50TB of data per day. They organized a worldwide algorithm challenge in the year 2016 [1] for forecasting ride supply and demand. We used the 2016 Didi algorithm competition dataset to try and forecast taxi trip supply and demand for any given date, time, and location using regression models covered in the Data Mining 2017 Spring semester at the School of Computing of the University of Utah. Below is a sample of some methods used:

1. Stochastic Gradient Descent
2. Polynomial Regression
3. Gaussian Kernel Regression
4. Hierarchial Clustering

Ride Hailing Data

In order to run the regression models, the categorical values in the Didi algorithm competition dataset needed to be converted into a regression friendly format. Depending on the type of categorical value, the new values are lists that consist of 0 values when the category is not present in an order and 1 or a count when the category is present in the order.

We have 1.3million rows in our training data consisting of ride hailing orders. When we took into consideration, only those orders where the start and end districts have data for traffic and places of interest, the number of training rows were reduced to 921,000.

Regression Results

Regression Type	Mean Squared Error
Linear using Stochastic Gradient Descent	245.50
Gradient Boosting using features based on top 10 eigen vectors	285.97
Ridge Regression using features based on top 10 eigen vectors	293.83
Lasso Regression using features based on top 10 eigen vectors	293.83
Polynomial degree 2 regression using features based on top 10 eigen vectors	2.34×10^{53}
Polynomial degree 3 regression using features based on top 10 eigen vectors	3.53×10^{73}
Polynomial degree 4 regression using features based on top 10 eigen vectors	9.31×10^{93}
Gaussian Kernel Regression	376.90

Table 1: Table caption

References

Patterns In Ride Hailing Data

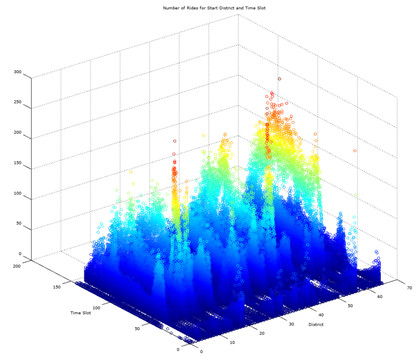

Figure 1: Number of rides scatter plot for start district and time slot.

We also looked at patterns that could be attributed to the passengers taking taxis to and from work. Using a Monday through Friday work week, we first found the peak timeslots. Then using the peak timeslots, we plotted the number of orders per start district and the number of orders per destination district for the first peak. While there does appear to be some pattern that could lead to deducing residential and commercial districts, the pattern is not strong enough to reduce the complexity of the dataset.

Since we were not getting good results, we decided to look for outliers. The outliers are usually computed using the inter-quartile distance between the first and third quartiles and using that value to include elements that are up to 1.5 times the inter-quartile distance on either side. Elements outside these limits are marked as outliers. However it is doubtful that these are outliers since it is actual data and the reported number of outliers is large.

We next looked at clustering methods to determine and filter out outliers. After experimenting with a couple of methods, there was no improvement in the mean squared error. We confirmed our previous theory that the identified outliers are most likely not true outliers, but result of the scattered nature of the dataset.

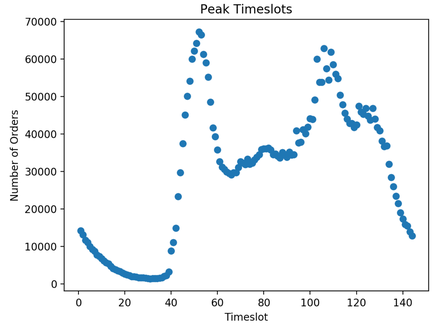

Figure 2: Number of rides scatter plot for a timeslot (Monday-Friday).

Conclusion

It looks like traditional regression methods do not work for complicated situations where we need to model human behavior. Deep learning techniques have been shown to give better results where traditional methods have failed or do not result in the desired level of accuracy. This is possibly a problem that is better suited to using deep learning.