The Butterfly: Relating NYC taxi trips with Dow Jones Industrial Average (DJIA)

Julian Gao (julianyg) Qiujiang Jin (qiujiang)

October 28, 2016

1 Introduction

New York City, the most populous metropolis on U.S. East Coast, has one of the world's busiest public transportation system. The symbolic yellow cab plays a significant role in this complex. Millions of taxi trips per day are records of peoples' behavior, containing information of the population flow, where numerous tasks are executed, money transacted. From this massive cornucopia of data, we want to find out the correlation between taxi trips and market index, and build up a predictor based on this correlation.

2 Task Definition

Our task is defined as, given a time series of taxi data and DJIA, train a predictor that can take taxi trips data from a single day, to predict the rise/drop of DJIA on the next day. The comparison is taken between current day's closing DJIA and the next day's opening DJIA, where rise and drop are indicated by +1 and -1. For example, we first train the predictor with taxi trip record between Apr 2015 - Apr 2016, as well as the DJIA of that time period. Next we give the predictor a record from Sept 5^{th}, 2016, and it will predict rise/drop of DJIA on Sept 6^{th}.

3 Potential Problems & Approaches

The task can be decomposed into following aspects:

Extract useful features from data. The taxi trip records are extremely information-rich, especially accumulated by millions of individuals in a metropolis. The routes may indicate events with magnitude from some region-wide events to personal daily routines. The dataset only provides a few main features: time, duration, trip fare, and location. The GPS coordinates will be gridified to discretize region indicators. The original data contain 18 features, and we need to cut them down while preserving useful information.

Deal with high-dimensional vectors. Since we are taking taxi data from a single day to predict next day's DJIA trend, we have to combine all data points within one single day to a single data point. This is the greatest challenge in training, which may create a huge vector space, if we are to incorporate the location (e.g., grid $g_i \to g_j$) information into training. The high-dimensional data points will definitely cause computational complexity issues, and training with more dimensions than data points is a danger zone for any type of machine learning. To eliminate this problem, one idea is to calculate the average daily pick-up and drop-off rate of passengers for each grid, but this may also generalize the data too much and lead to information loss. Another possible solution is to first performed K-Means on the full data set, generate cluster centroids, then calculate meta-data of each day on top of the clustered centroids.

Train method selection. One simple and intuitive approach is to incorporate both discrete indicators and continuous feature values into one single vector for linear regression. Kernelized SVM methods are also applicable, but may get restricted by the huge vector size. Other machine learning algorithms, such as random forest, are also potential solvers for this problem.

4 Baseline & Oracle

We implemented the baseline prediction with a naïve approach. We take the testing data (records strictly within one single day) and calculate the average fare rate ($/sec). A threshold is given, and the predictor reports a rise in the next day's DJIA if average fare above the threshold, else reports drop. The oracle is a hardcoded case, which stores a subset of facts, and returns the fact when tested.

5 Dataset

We adapt the yellow taxi trip and fare data from NYC government website[1]. The entire dataset is humongous, containing time series from 2009, and is split into months. For convenience and due to time/resource limitations, we use a small portion from Mar 2016, which is 1.8GB in size. The data is stored in csv files, organized as follows:

1, 2016-03-01 00:00:00, 2016-03-01 00:07:55, 1, 2.50, -73.97674560546875, 40.765151977539062, 1, N, -74.004264831542969, 40.746128082275391, 1, 9, 0.5, 0.5, 2.05, 0, 0.3, 12.35, where each column corresponds to

VendorID, tpep pickup datetime, tpep dropoff datetime, passenger count, trip distance, pickup longitude, pickup latitude, RatecodeID, store and fwd flag, dropoff longitude, dropoff latitude, payment type, fare amount, extra, mta tax, tip amount, tolls amount, improvement surcharge, total amount.

The DJIA is gathered from Yahoo finance[2], and is processed to convert numbers into date: +1/-1 label mapping.

6 Related Work

Harvard NYC taxi data prediction project.[3]
Previous CS 221 final project.[4]

References

[1] "Nyc taxi data." http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml.

[2] "Djia history data." https://finance.yahoo.com/quote/%5EDJI/history?p=%5EDJI.

[3] "Harvard taxi data." http://sdaulton.github.io/TaxiPrediction/.

[4] "Predicting taxi pickup." https://web.stanford.edu/class/cs221/restricted/projects/vhchoksi/final.pdf.