**INFO 7390- Advances In Data Sciences and Architecture**

Analysis & Predict taxi trip in New York based on climate data & geospatial data

Ziwei Fan 001855517

Jan.29th 2017

1. **Abstract:**

The status of taxi trip duration and number of request will greatly depend on the time slot and the area of one city. In peak hour, the number of request will go up and meanwhile, the trip duration will increase as well in some area.

Additionally, the weather condition in different month will appear with different temperature, precipitation or snow-fall. That may cause the different consequence on the traffic flow.

Thus first up, in my thesis, my aim is to **visualize the trip duration of taxi by geospatial tool to animate and predict it by model**. Apart from analytics, **analysis how the weather affects the traffic in New York.** Above all, my aim will focus on the several perspectives as following:

|  |  |  |  |
| --- | --- | --- | --- |
| **Goal** | | | **Realistic scenario** |
| **Application** | **Phases** | **Phase I:**  **Analysis** | **1.Trip Duration Analysis** |
| 1. Data Preprocessing 2. Trip duration overview 3. Feature Extraction 4. Geo information distribution |
| **2.Weather-Trip duration Analysis** |
| 1. Pattern Analytics(Historical dataset) 2. Principle Component Analysis(PCA) |
| **Phase II: Visualization** | Geospatial Request Animation with Flare |
| **Phase III:**  **Prediction** | **Trip Duration Prediction** |
| Mini-Batches K means |
| **Measures of validation** | **Score** | |  |

1. **Introduction:**

In the Trip duration analysis,the geospatial data in **latitude and longitude will be visualized with the changing of time slot, and in further step, the time cost in trip will be forecasted.** In this section, diverse practical tools to draw geospatial image could be selected to animate the interactive image. In this project, Flare is more recommended to build the animation. Next step, the further prediction of taxi trip duration will be conducted by mini-batches K means, since the items of record are more than 10,000. What we should do is to evaluate and score which model & algorithm could fit better in this context via K-Fold-Crossval Mean Score or Accuracy Score—that is the best model.

Additionally, referenced by the ‘Road Weather Management program’ [1], vast range of the conditions will influence the roadway, consequently, the traffic flow will be impacted. For example, in term of the precipitation, the visibility distance and pavement friction etc. will go down with the continuous rain. Consequently, driver capabilities and vehicle performance will be influenced due to these aspects. Totally, the crash risk and traffic flow will be affected. Apart from the conditions analytics, it is recorded by the specific ‘weather-related crash statistics’ here. It is reported that in the snow/sleet day, 210,341 crashed in annual average. 4% of vehicle crashed and 739 person killed.

Based on the potential risk from weather condition, it is crucial to **uncover the correlation between traffic flow and weather condition: explore the association between traffic flow & weather condition**; First up, I preprocess the raw data collected from diverse departments, such as National Weather Service and traffic department. Then, it is crucial to dig out the covered relationship by data merged through the data visualization. With analytics based on the correlations found, the predictions based on several algorithms are play significant role to forecast what will occur in the future and remind the service department to take actions.

1. **Code with Documentation:**
2. **Data Resources:**

* Taxi Trip Duration 2016 in NY：<https://www.kaggle.com/c/nyc-taxi-trip-duration/data>
* New York City Taxi with OSRM：<https://www.kaggle.com/oscarleo/new-york-city-taxi-with-osrm/data>
* Weather data in New York City – 2016：<https://www.kaggle.com/mathijs/weather-data-in-new-york-city-2016/data>

Rather than merely 1 data source, the 3 data source will be collected and merged by different attributes according to different requirement.

Besides, the time-series analytics is necessary for further mining.

Description about the data sources [2] [3] [4]:

* The data is based on individual trip attributes: pickup\_datetime, dropoff\_datetime, pickup\_lat-long etc.
* The data extracted by OSRM contains fastest routes information for each data point: Starting\_Street, End\_Street, total\_distance, total\_TravelTime etc.
* The weather data contains the weather data collected by the weather station in Central Park between Jan.-Jun. 2016: Min\_Temperature, Max\_Temperature, Average\_Temperature, Precipitation, Snow\_fall, snow\_depth.

1. **Algorithms:**

The following algorithms can be implemented in the topic:

* 1. **Analytics:**
* PCA (Principle component analysis):

PCA is a technique used to emphasize variation and bring out strong patterns in a dataset. It's often used to make data easy to explore and visualize. It is useful for eliminating dimensions.

* Flare:

[10] Flare is an ActionScript library for creating visualizations that run in the Adobe Flash Player. From basic charts and graphs to complex interactive graphics, the toolkit supports data management, visual encoding, animation, and interaction techniques. Even better, flare features a modular design that lets developers create customized visualization techniques without having to reinvent the wheel.

* Pearson Correlation Coefficient:

Analysis the degree of correlation between some dimensions of weather & traffic data.

[5] Pearson Product-Moment Correlation <https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php>

‘Pearson Correlation Coefficient is a measure of strength of a linear association between 2 variables and is denoted by r ranged from 0-1: 0 indicates there is on association between the 2 variables. A value greater than 0 indicates a positive association. A value less than 0 indicates a negative association. Basically, it attempts to draw a line of best fit through the data of 2 variables.’

* 1. **Predict:**
* Mini-batches K means:

[11] The [MiniBatchKMeans](http://scikit-learn.org/stable/modules/generated/sklearn.cluster.MiniBatchKMeans.html#sklearn.cluster.MiniBatchKMeans) is a variant of the [KMeans](http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html#sklearn.cluster.KMeans) algorithm which uses mini-batches to reduce the computation time, while still attempting to optimise the same objective function. Mini-batches are subsets of the input data, randomly sampled in each training iteration. These mini-batches drastically reduce the amount of computation required to converge to a local solution. In contrast to other algorithms that reduce the convergence time of k-means, mini-batch k-means produces results that are generally only slightly worse than the standard algorithm.

* 1. Score:
* Kfoldcrossval Mean Score:

[8] What Is K-Fold Cross Validation? <https://magoosh.com/data-science/2017/12/08/k-fold-cross-validation/>

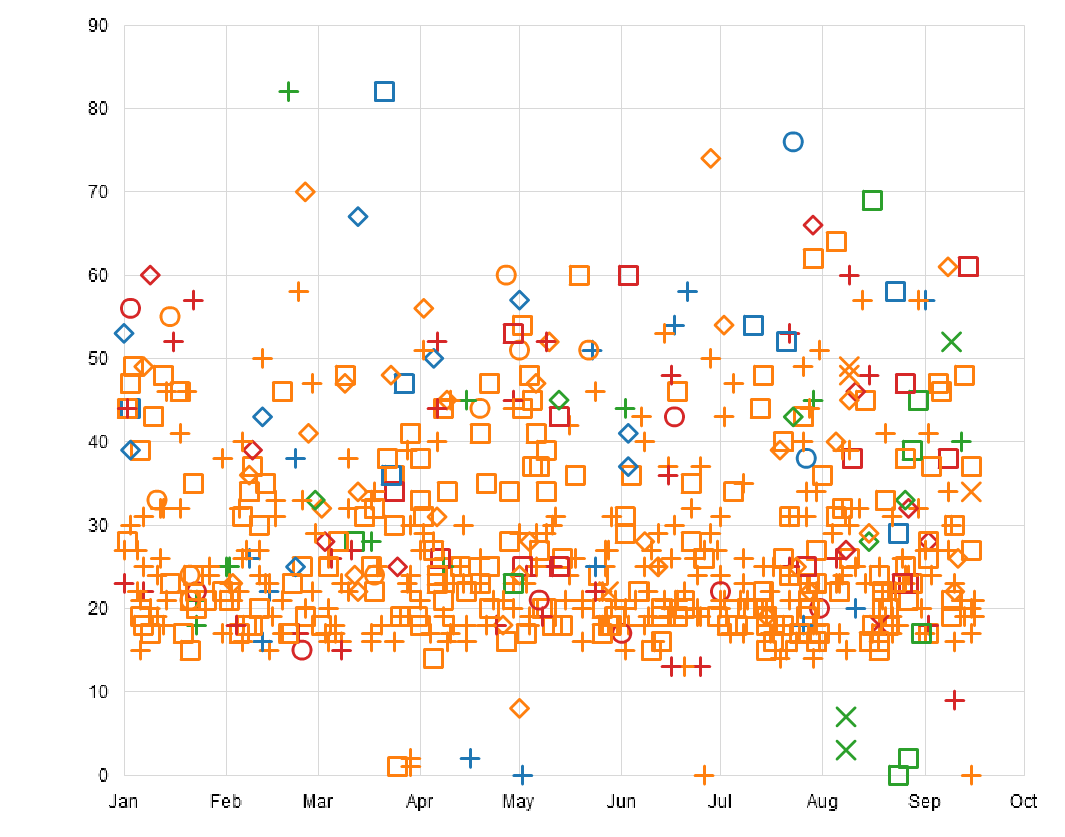
‘In K-Fold Cross Validation, all the entries in the original training data set are used for both training as well as validation. Also, each entry is used for validation just once.’

* Or: Accuracy Score

[9] 3.3. Model evaluation: quantifying the quality of predictions <http://scikit-learn.org/stable/modules/model_evaluation.html#accuracy-score>

‘Accuracy Score computes the accuracy, either the fraction (default) or the count (normalize=false) of correct predictions. ’

1. **Result & Analytics:**
2. **Data Analytics:**
   1. **Trip Duration Analysis:**
   2. Data Preprocessing
   3. Trip duration overview and distribution:
   4. Geo information distribution plot:
   5. Geospatial visualization based on lat-long:
   6. Animation based on geospatial data:
   7. **Weather-Trip duration Analysis:**
   8. Feature extraction:
   9. Correlation analysis-Flare:



1. **Prediction on Trip duration:**
2. **Validation Analytics:**

3.1 Score:

3.2 Speed:

1. **Discussion:**

Q1. How to merge the different cleaned datasets into one dataset without Re-output & Re-read?

Q2. How to choose the best model to train the dataset?

1. **Reference:**

[1] How Do Weather Events Impact Roads? <https://ops.fhwa.dot.gov/weather/q1_roadimpact.htm>

[2] New York City Taxi Trip Duration: <https://www.kaggle.com/c/nyc-taxi-trip-duration/data>

[3] New York City Taxi with OSRM: <https://www.kaggle.com/oscarleo/new-york-city-taxi-with-osrm>

[4] Weather data in New York City - 2016 <https://www.kaggle.com/mathijs/weather-data-in-new-york-city-2016>

[5] Pearson Product-Moment Correlation <https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php>

[6] Logistic Regression: <https://en.wikipedia.org/wiki/Logistic_regression>

[7] Decision Tree: <https://en.wikipedia.org/wiki/Decision_tree>

[8] What Is K-Fold Cross Validation? <https://magoosh.com/data-science/2017/12/08/k-fold-cross-validation/>

[9] 3.3. Model evaluation: quantifying the quality of predictions <http://scikitlearn.org/stable/modules/model_evaluation.html#accuracy-score>

[10] Flare data visualization for web:

<http://flare.prefuse.org/>

[11] Clustering-mini-batch-kmeans:

<http://scikit-learn.org/stable/modules/clustering.html#mini-batch-kmeans>