Project Proposal:

Taxi Riding Data Analysis and Demand Prediction System

Wenqi Xu & Zhenwei He
CAS 772 Mobile Data Analytics 2017 Fall
Instructor: Prof. Rong Zheng

Scope:

Main Objective:

1. Predict taxi riding demands, average fare and tip on specific location at specific time
2. Analyse taxi rush hour and hotspot
3. Implementing a cloud system as backend and a web application as frontend
4. Visualization of some data analysis results and comparison with Uber riding demands

Optional Objective:

1. Add travel plan and cost prediction function

Motivation:

Taxi has always been a significant way to travel in the city. There are over 200 million taxi rides in NYC each year. Data about these taxi rides have been available. How to use these data to make transportation more efficient is a challenge. It would be useful if taxi drivers can know where he can pick up more passengers and get more tip on a specific time. Also, the analysis of taxi rush hour and hotspots can help taxi company better distribute the taxi resources. Recommending an optimal travel plan will also be helpful for passengers. We are eager to see a big data solution to this problem.

Background and Related Work:

Analysis of taxi riding data is an urban computing problem, which connects sensing technologies, novel visualization methods, advanced data management and analytics models to create solutions that improve the urban environment (Schneider, 2015 Nov. 17). Todd W. Schneider (Zhang, Cheng & Xiong, 2015 May 15) analyzed New York taxi riding pattern. The result shows some compelling pattern for different Taxi (Yellow taxi, Uber and Lyft) and for different areas as well as pick up time. Zhang (Zheng et al., 2014) analyze NYC taxi data on the topic of "Understanding Taxi Economics." The question of how the revenue varying over different time and neighborhood is investigated. It is implemented in Map-Reduce Framework with Hadoop Streaming API and Python. Our project is different from the above in the following aspects: 1) Instead of analyzing taxi riding pattern, we focus on using machine learning models predict taxi riding demand, average fare and tip and so forth, which provide direct economic benefits. 2) We implement a cloud web system for users based on novel technologies such as Spark and AWS.

Proposed Solution:

Dataset Collection:

1. Taxi dataset from the New York City Taxi and Limousine Commission:
   a. Feature: Pickup & Drop Off location, time stamp, fare, tips
   b. Sample data from some districts
   c. Dataset Size: Million to Billion records
2. Uber trip dataset in New York City:
   a. Feature: Pickup location and time stamp
   b. Sample data from some regions
   c. Dataset Size: Million records
3. Historical weather data: http://weathersource.com/
   a. Feature: weather type, temperature and so forth

Main Function Steps:

1. **Data Pre-processing:**
   Fetch datasets above and preprocess them (feature selection and clean);
   Then divided them into parameter tuning dataset, training dataset, and test dataset
2. **Backend:**
   a. Use Random Forest (Primary), Linear Regression (Optional) to build up demands prediction model
   b. Use Random Forest(Primary) and K-Nearest Neighbor (Optional) model to build up rush hour and hotspot prediction model
   c. Integrate two models and ensemble different classifiers
   d. Tune parameters (tree_num, tree_depth, feature_num, feature selection method, K, Item_num and so on)
   e. Use the best parameters to train and test a complete model on Spark/AWS
3. **Frontend:**
   Use AWS S3, Google Map API, and AWS Lambda and so on to build a Web Application for the prediction function
4. **Visualization:**
   a. Use Python to show the riding demands comparison result of taxi and Uber, different locations, different time and different weather
   b. Add this Visualization to Web front-end (optional)
5. **Other data analysis (optional)**

Optional Function Steps:

1. Use Google Map API or path plan algorithms like Dijkstra to plan a travel path, which includes an option to avoid rush hour and possibly congested locations
2. Use the prediction of fare and tips to predict the cost of a travel path on specific time
3. Add this function to the Web application

Reference

[1] Schneider, T. (Ed.). (2015, November 17). Analyzing 1.1 Billion NYC Taxi and Uber Trips, with a Vengeance. Retrieved October 06, 2017, from http://toddwschneider.com/posts/analyzing-1-1-

billion-nyc-taxi-and-uber-trips-with-a-vengeance/

[2] Zhang, S., Cheng, H., & Xiong, G. (2015, May 15). NYC-Taxi-Data-Analysis. Retrieved October 06, 2017, from https://github.com/marcogx/taxi-analysis

[3] Zheng, Y., Capra, L., Wolfson, O., & Yang, H. (2014). Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *5*(3), 38.