10-718 Data Analysis (Spring 2019)
Taxi Travel Time Prediction
Version 1.0

Submit hard copy between 2pm and 5pm
on February 11, 2019 to GHC 8004

1 Background

Taxi services and ride-sharing services (Lyft, Uber, etc.) regularly collect information on trips. These travel traces contain a complex mixture of temporal and spatial data, and are usually augmented with quantities such as fare amount and payment type. As such, this raw data contains information about both users' travel patterns and the development of an ever-fluctuating industry.

Analyzing this type of data has recently proven useful for a variety of purposes. For example, travel traces can be used to identify interactions among geographical regions of a city as well as to identify predominant travel flows [2]. Thus, they have been used for city planning (e.g. street planning, policy making), usually with the goal of reducing congestion in urban areas. Furthermore, intelligently utilizing taxi travel traces has the potential to improve current electronic taxi dispatching systems.

Concretely, if we are able to predict how long a driver will have his taxi occupied, dispatchers can better identify which driver to assign to new incoming requests [1]. Unfortunately, it is not easy to accurately predict taxi travel time because traffic is affected by many external factors such as weather, holidays, and accidents. In addition, historical data may not be enough to make predictions, as user and city patterns evolve over time.

In this course, we will focus on this problem of estimating travel time given known start location, end location, and start time. There are a few existing approaches for solving this type of problem. We encourage you all to research the approaches used and extend/adapt them using your knowledge of machine learning.

2 Data

We will be using data recorded by New York City taxicabs for the 2017 year. Note that starting and ending locations are pre-processed by discretizing the region and reporting the index of the starting and ending sub-regions. Details about the discretization scheme can be found here. When describing your proposed pre-processing, you should address how you will handle the location data. The data can be accessed here.

3 Assignment 1

For this assignment, perform the following four steps. Write a report with one section per step.

1. Familiarize yourself with the data; identify potential difficulties and required cleaning/pre-processing (3 points).
    - You will be expected to demonstrate to us that you interacted with the data to a sufficient degree. This can be done using data visualization, summary statistics, etc. You should convey enough information to support your choice of pipeline.
2. Formulate a machine learning problem that will help the domain expert achieve his/her goal of being able to accurately predict taxi travel time of future trips (6 points).
    - Some questions to consider include: (1) How can we deal with the size of the dataset? (2) How will we compare different approaches (what loss function(s) should we use)? (3) How can we address the influence of external factors and the dynamic nature of the data?
3. Propose a detailed analytical pipeline to solve the machine learning problem (3 points).
    - You do not have to train any models for this assignment, but preliminary analysis is encouraged. The main purpose is for you to understand the problem and formulate a road-map for the analysis you want to conduct in the subsequent assignments.
4. Design an experiment to evaluate the effectiveness of your approach (6 points).
    - You will be expected to compare to at least one existing approach. It may help to perform some literature search to see what approaches others have taken to solve the problem. A personal laptop should be sufficient for performing your entire proposed pipeline.

References

[1] ECML/PKDD 15: Taxi Trip Time Prediction. https://www.kaggle.com/c/pkdd-15-taxi-trip-time-prediction-ii. [Online; accessed 20-Jan-2019].

[2] Agostino Nuzzolo, Antonio Comi, Enrica Papa, and Antonio Polimeni. Understanding taxi travel demand patterns through floating car data. 2018.