10-718 Data Analysis (Spring 2019)
Assignment 3: Taxi Travel Time Prediction
Version 1.2

Beat Kaggle baseline by 5pm on April 16, 2019.
Submit hard copy of report between 3pm and 5pm
on April 22, 2019 to GHC 8123.
Submit code via email by report deadline
(email tianqil1@andrew.cmu.edu with subject
"10-718 HW3 Code Submission - [andrewID]").

1 Overview

Assignment 2 was focused on implementing a preliminary pipeline based on our Assignment 1 discussion. In Assignment 3, we will improve this pipeline.

We will do this in two ways:

1. You are expected to make modifications to the model you used for Assignment 2. This could be as little as spending more time fine-tuning your model or as much as choosing a different model class and changing most of your pipeline.

2. You will identify a source of outside information which improves your model's performance. Using this information requires additional adjustment of your pipeline.

For point 2, keep in mind that you may only add new features and not new data points (i.e., you can add more columns for the trips provided to you, but you may not add more rows of training data to your dataset). Additionally, data added for a validation or testing point may not use information gathered after the trip has begun. For example, weather conditions at the start of the trip is an appropriate feature, but weather conditions at the end of the trip is a feature which violates the predictive nature of this task. As a result of this restriction, you may not use the unlabelled validation/testing data as part of your pipeline (e.g., as unsupervised training data). You have access to it only for the purpose of labelling it once you have completed training.

Please ensure that your pipeline can be trained on a personal laptop.

2 Data

Just as in the previous assignment, we will be using data recorded by New York City taxicabs for the 2017 year. As before, the starting and ending locations are pre-processed by discretizing the region and reporting the index of the starting and ending sub-regions. For this assignment, we have performed the pre-processing steps used for Assignment 2.

3 Loss Function

We will use the same loss function as in the previous assignment: the root mean squared log error. More precisely, we will evaluate a set of predictions using the expression

$$RMSLE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\ln\frac{1+\hat{y}_i}{1+y_i}\right)^2}$$

where N is the number of trips evaluated, y_i is the actual travel time for trip i and $\hat{y}_i$ is the predicted travel time for the same trip. We use this loss because it has two properties that we consider desirable for our purposes: (1) it avoids over-penalizing outliers, and (2) it penalizes under-estimates more than over-estimates.

4 Kaggle Competition

As part of this assignment, you will be asked to beat a baseline model. You will not have access to the model, but you will have access to its performance (in terms of its $RMSLE$) on a set of validation data. The baseline is thus a submission in a Kaggle competition in which we expect you to participate. **You should create a Kaggle submission that beats the baseline's performance one week before the final report's deadline**.

The competition aspect of the assignment is meant to incentivize and motivate you. Your grade will not be negatively affected by your final ranking, but beating the baseline is part of the assignment grade. The competition, along with all the relevant data, can be found here.

5 Assignment

There are two components in this assignment.

1. **(6 points)** By 5pm on April 15, 2019, make a submission to Kaggle that beats the baseline. For this submission, you may choose to use external data, but doing so is not required. (The baseline was trained without any external data.) You may select up to two submissions for the public leaderboard; at least one of these two should beat the baseline. Note that you will only see performance on the validation set. You will be evaluated on a held-out test set, so do not over-fit to the validation set.

2. **(12 points)** In your report, describe and justify the pipeline used for your submission. Be sure to clearly address the following questions

   (a) Provide a clear, detailed description of your overall pipeline sufficient to reproduce your exact pipeline.

   (b) Describe the process you used to select your pipeline and improve it. This should include summaries of experiments you performed to evaluate potential improvements as well as a detailed description of your hyperparameter selection process.

   (c) Describe the additional data you used, how you gathered it, and what modifications you made to your pipeline to use it.

   (d) Perform a basic ablation analysis. At the very least, compare your model with "pipeline improvements and external data" to "pipeline improvements only," "external data only," and "no improvements and no external data."

   (e) Justify your choice of overall pipeline and external data so that a (non-ML expert) domain practitioner could understand why you made these choices. This justification could reference the outcomes of the experiments in point 'b' and ablation analysis in point 'd'.

   (f) Propose concrete and meaningful modifications or extensions to your solution.

6 Timeline

The timing for this assignment is outlined below (please read carefully):

1. April 8 - Assignment 3 is released and the Kaggle competition begins. Students can make submissions and see their performance on a validation set.
2. April 16 - Deadline to beat the proposed baseline. Students are notified via email as to whether either of their two selected submissions beat the baseline on the held-out test set. Note that the competition remains open to give students the opportunity to refine their pipelines.
3. April 22 - Deadline for the final report. Kaggle competition ends. Also, email your code to Roy (tianqil1@andrew.cmu.edu) with subject "10-718 HW3 Code Submission - [andrewID]".

You may use late days for the April 16 ("beat the baseline") or April 22 (final report) deadline. This will work the same way as using late days for any other assignment deadline. Using late days on one deadline does not shift the other deadline (e.g., if a student uses one late day for beating the baseline, they are still expected to submit the report on April 22 unless they plan to use an extra late day). Please notice that the Kaggle competition will end on April 22 whether or not a student uses late days for the final report deadline.