Final Project: Regular Track

TA email: `html_ta@csie.ntu.edu.tw`

RELEASE DATE: 10/25/2023

BLUE CORRECTION: 11/02/2023 05:30

REPORT DUE DATE: **12/27/2023 13:00**

Unless granted by the instructor in advance, no late submissions will be allowed. That is, you will not be allowed to submit your report after the deadline and will get zero point for the final project. The gold medals cannot be used for the final project.

Any form of cheating, lying, or plagiarism will not be tolerated. Students can get zero scores and/or fail the class and/or be kicked out of school and/or receive other punishments for those kinds of misconducts.

You need to write your report in English (strongly encouraged) or Traditional Chinese with the common math notations introduced in class or in the problems. We do not accept solutions written in any other languages.

Introduction

In this final project, you have the opportunity to participate in an exciting machine learning competition. The You(-know-what) bike company tries to predict the number of bikes in stations around NTU to improve their bike dispatching system. In particular, the company has to predict the number of bikes in 112 stations around NTU for every 20 minutes in the specific time windows, where the number of bikes is a non-negative integer.

Now, having collected some data from the bike service, the company has decided to ask you, a rising star in the data science team, to help with the prediction task. You need to prepare data, study different approaches for the prediction task, and then recommend the best choice to the company. The best choice can be referred to as your competitiveness on the Kaggle scoreboard or other appealing properties of your approaches, such as efficiency and interpretability. Then, you need to submit a comprehensive report that describes not only your recommended approach to the company but also the reasoning behind your recommendation. Well, let's get started!

Data Set

The data is available at the following link and is collected from the *YouBike2.0 Taipei City public bicycle real-time information* on *Data.Taipei*[1] **To maximize the level of fairness, you are NOT allowed to fetch additional data from the source on *Data.Taipei* any time.**

`https://github.com/hyusterr/html.2023.final.data/tree/release`.

The problem is formalized as a regression problem, where the goal is to predict the "ground-truth" number (a non-negative integer) of bikes in each station every 20 minutes accurately. Similar to any real-world data, some of the examples that you get may contain missing values.

The data is divided into the training set and the test sets. The training set contains information on over 1000 bike stations in Taipei City. The detailed meanings of every field of the data are shown on the `README` file on the GitHub repository. For every bike station, there is information about its address, its coordinate, and its number of bikes every minute, with a time span from 10/2 to 10/11 and another time span from 10/15 to 10/20.

Starting from 10/26, we will release the collected training data of the previous day before 13:00 at the same link, that is, the data from 10/25 will be released before 10/26 13:00, the data from 10/26 will be released before 10/27 13:00, and so on. As all predictions will need to be submitted before some

[1] `https://data.taipei/dataset/detail?id=c6bc8aed-557d-41d5-bfb1-8da24f78f2fb`.

deadlines (see below), it is learners' responsibility to design a reasonable process to include the latest data.

Evaluation

You have to predict the number of bikes in the 112 stations around NTU for every 20 minutes for the following time windows.

- Public test set: From 10/21/2023 00:00 to 10/24/2023 23:40. (i.e. 00:00, 00:20, 00:40, ... 23:00, 23:20, 23:40). This result is going to be shown on the scoreboard of the competition website.
- Private test set I: From 12/4/2023 00:00 to 12/10/2023 23:40. This needs to be submitted **before** 12/3/2023 23:59.
- Private test set II: From 12/11/2023 00:00 to 12/17/2023 23:40. This needs to be submitted **before** 12/10/2023 23:59.
- Private test set III: From 12/18/2023 00:00 to 12/24/2023 23:40. This needs to be submitted **before** 12/17/2023 23:59.

All times above are in the **UTC+8** time zone. If you miss a deadline for a particular private set, you do not get any score from that set. We will choose the best two out of three private scores as your out-of-sample evaluation. So if you do well in private sets I and II, then participating in private set III is optional.

Let $b_{i,t} \in \mathbb{N} \cup \{0\}$ be the ground-truth number of bikes at time t and station i and there are in total s_i stops in the station i, and $\hat{b}_{i,t} \in \mathbb{R}$ be your predicted number. The evaluation metric is defined by

$$\text{err}(\hat{b}_{i,t}, b_{i,t}, s_i) = 3 \left| \frac{b_{i,t} - \hat{b}_{i,t}}{s_i} \right| \times \left(\left| \frac{b_{i,t}}{s_i} - \frac{1}{3} \right| + \left| \frac{b_{i,t}}{s_i} - \frac{2}{3} \right| \right)$$

That is, it is the absolute error rate weighted by some scaling of how far the ratio of parked bikes is far from $[\frac{1}{3}, \frac{2}{3}]$. That is, a nearly-empty or a nearly-full site is more important.

As in your training data, some ground-truth numbers might be missing. If $b_{i,t}$ is missing from some (i, t), we fetch the first non-missing value from the sequence of $(b_{i,t+1}, b_{i,t+2}, \ldots)$ to impute $b_{i,t}$ during evaluation.

Survey Report

You are asked by the company to study at least THREE machine learning approaches using the data. Then, you should make a comparison of those approaches according to some different perspectives, such as (but not limited to) efficiency, scalability, and interpretability. Then, you need to recommend THE BEST ONE of those approaches as your final recommendation and provide the "cons and pros" of the choice. The report from the study is considered THE most important and incredibly exciting part of our final project.

The survey report should be no more than SIX A4 pages with readable font sizes. The most important criterion for evaluating your report is reproducibility. Thus, in addition to the outlines above, you should also describe how you pre-process your data, such as the features you build; introduce the approaches you tried and provide specific references, especially for those approaches that we didn't cover in class; list your experimental settings and the parameters you used (or chose) clearly. Other criteria for evaluating your survey report would include, but are not limited to, clarity, strength of your reasoning, "correctness" in using machine learning techniques, the work loads of team members, and properness of citations.

Our sincere suggestion: *Think of your TAs as your boss who wants to be convinced by your report.*

For grading purposes, a minor but required part in your survey report for a two- or three-people team (see the rules below) is how you balance your work loads.

Competition

The submission site will be based on Kaggle and will be announced later. Please simply form a team on the site and then participate in the competition. You should form a *single* team and NOT use multiple team accounts. Otherwise your team will be considered as violating the course policy, which results in severe consequences. Use your submissions wisely—*you do not want to leave the TAs with a bad impression that you just want to "query" or "overfit" the test examples.* After submitting, there will be a scoreboard showing the test error of the public test set. The "hidden" test error on the private test sets will eventually be used to evaluate your performance. As discussed above, we will open 3 Kaggle competitions for 3 private test set with the submission deadline at 12/3/2023 23:59, 12/10/2023 23:59, and 12/17/2023 23:59, respectively. After each deadline, the competition site will continue to be open for you to test your ideas and finish your reports. However, any late submission will not be taken into account in grading the competition performance.

Misc Rules

Report: Please upload one report per team electronically on Gradescope. You do not need to submit a hard-copy. You should align your *team name* on your report and the Kaggle website. The report is due at 13:00 on 12/27/2023.

Teams: By default, you are asked to work as a team of size THREE. A one-person or two-people team is allowed only if you are willing to be as good as a three-people team. It is expected that all team members share balanced work loads. Any form of unfairness, such as the intention to cover other members' work, is considered a violation of the honesty policy and will cause some or all members to receive zero or negative score.

Bonus Track: As mentioned in class, we will have a bonus track surrounding LLMs. It is up to each team to decide whether to participate in the Bonus Track. Stay tuned for its rules!

Algorithms: You can use any algorithms, regardless of whether they were taught in class.

Packages: You can use any software packages for the purpose of experiments, but please provide proper references in your report for reproducibility.

Source Code: You do not need to upload your source code for the final project. Nevertheless, please keep your source code until 01/31/2024 for the graders' possible inspections.

Grade: The final project is worth 600 points. That is, it is equivalent to 2.5 usual homework sets. At least 540 of them would be reserved for the report. The other 60 may depend on some minor criteria such as your competition results, your work loads, etc.

Collaboration: The general collaboration policy applies. In addition to the competitions, we still encourage collaborations and discussions between different teams.

Data Usage: As mentioned, you **cannot** query data from the source on Data.Taipei. But you can generally use other external data sets for your experiments. Whenever you use external data, be sure to provide the data source and clearly illustrate how to reproducibly use them, such as how to design features and how to avoid data snooping, in the report.