Final Project

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

RELEASE DATE: 11/29/2020

COMPETITION END DATE: **01/11/2021 NOON ONLINE**

REPORT DUE DATE: **01/19/2021 NOON ONLINE**

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 also 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 should write your solutions in English 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 are going to be part of an exciting machine learning competition. A hotel booking company tries to predict the daily revenue of the company from the data of reservation. In particular, after a room reservation request is fulfilled (i.e. no cancellation), on the arrival date, the revenue of the request is the rate of the room (called ADR) multiplied by the number of days that the customer is going to stay in the room, and the daily revenue is the sum of all those fulfilled requests on the same day. The goal of the prediction is to accurately infer the future daily revenue of the company, where the daily revenue is quantized to 10 scales.

Now, having collected some data from the hotel booking service, the CEO of the company decides to ask you, a rising star in the data science team, to help with the prediction task. You need to fight for the most accurate prediction on the score board. Then, you need to submit a comprehensive report that describes not only the recommended approaches, but also the reasoning behind your recommendations. Well, let's get started!

Data Set

The data sets are processed from the Kaggle hotel booking demand data.

`https://www.kaggle.com/jessemostipak/hotel-booking-demand`

To maximize the level of fairness, you are not allowed to download the original data at any time. But you are welcomed to go check the descriptions of the data and the discussions on Kaggle.

The problem is formalized as an ordinal ranking (i.e. ordinal regression or ordinal classification problem, where the goal is to predict the revenue "truth" (a discrete value) accurately using the error function

$$\mathrm{err}(y, \tilde{y}) = |y - \tilde{y}|.$$

Both the predicted value $\tilde{y}$ and the truth value y are assumed to be within $\{0, 1, 2, \ldots, 9\}$.

The data will be divided to the training sets and the test sets. The training data is composed of two files. `train.csv` contains room reservation requests. The meanings of each field can be found in the original Kaggle site. `train_label.csv` contains the revenue "truth" for each day of the training period.

The test data is essentially the same as the training one. `test.csv` is the counterpart of `train.csv`, with some "future" fields hidden. `test_nolabel.csv` is the counterpart of `train_label.csv`, except that the "truth" is hidden. *You are not allowed to peep/label the true answers of the test sets.*

Survey Report

You are asked by the board to study at least THREE machine learning approaches using the training set above. Then, you should make a comparison of those approaches according to some different perspectives,

such as efficiency, scalability, popularity, and interpretability. In addition, you need to recommend THE BEST ONE of those approaches as your final recommendation and provide the "cons and pros" of the choice.

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 would be announced later. 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 score board showing the test error on a random half of the data set. The "hidden" test error on the other half will eventually be used to evaluate your performance.

The competition ends at noon on 01/11/2021. We'll have a mini-ceremony to honor the best team(s) on 01/12/2021. The competition site will continue to be open until the due day of the report.

Misc Rules

Report: Please upload one report per team electronically on Gradescope. You do not need to submit a hard-copy. The report is due at noon on 01/19/2021.
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.

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 03/31/2021 for the graders' possible inspections.

Grade: The final project is worth 800 points. That is, it is equivalent to two usual homework sets. At least 720 of them would be reserved for the report. The other 80 may depend on some minor criteria such as your competition results, your discussions on the boards, 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: You can use only the data sets provided in class for your experiments, and you should use the data sets properly. Getting other forms of the data sets is strictly prohibited and is considered a serious violation of the honesty policy. Using any tricks to query the labels of the test set is also strictly prohibited.