Final Project

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

RELEASE DATE: 11/26/2021 (GOOD LUCK!)

REPORT DUE DATE: **01/20/2022 13:00 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 telephone company wants to predict whether the customers would stop using its services and why the customers stop using its services. This is quite important since if we know whether/why customers stop using the services beforehand, the company can try to keep the customers (and retain the revenue) by taking some proper marketing actions.

Now, having collected some data from telephone company, the CTO wants to challenge you, a new coming data scientist in the company, to help with the task. You need to fight for the most accurate prediction on the score board. Then, you need to submit a comprehensive report that describe 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 IBM telco customer churn data. To maximize the level of fairness, you are not allowed to download or check the original data at any time.

The problem is a multi-class classification problem, where the goal is to predict the customers' category:

{`No Churn, Competitor, Dissatisfaction, Attitude, Price, Other`}

`No Churn` means that the customer will keep using the service, while the other five categories indicate the reason that the customer leaves the service.

The customers, represented by unique IDs in the data, will be divided to training IDs and testing IDs. There are few files that contain the information of customers, such as gender, age, etc. The label file, `status.csv`, shows the category of the training IDs.

For more detailed description, and for downloading the data, please go to

`https://www.kaggle.com/c/html2021final/data`

Note: there may be some missing values in the data, including information of customers and labels of the training IDs, so you should deal with them carefully.

Evaluation

For the evaluation, we calculate F1-scores with respect to each category and then take average on the six F1-scores. For the introduction and definition of the F1-score, please refer to

`https://en.wikipedia.org/wiki/F-score`

Survey Report

You are asked by the CTO 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 (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 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

We use Kaggle as our competition site:

`https://www.kaggle.com/c/html2021final/`

Please simply form a team on the site and then participate in the competition. 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 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 13:00 on 01/20/2022.

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 03/31/2022 for the graders' possible inspections.

Grade: The final project is worth 1000 points. That is, it is equivalent to 2.5 usual homework sets. At least 900 of them would be reserved for the report. The other 100 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.