Final Project

TA email: ml2016ta@csie.ntu.edu.tw

RELEASE DATE: 05/08/2017

COMPETITION END DATE: **06/19/2017 NOON ONLINE**

REPORT DUE DATE: **06/27/2016 NOON ONLINE**

Unless granted by the instructor in advance, no late submissions will be allowed.

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. Consider a company that runs intelligent advertisement service. The key to a successful service is to predict whether a user would click some advertisement. The prediction should be based on some known profile about the user. Now, having collected some data from the service, the board of directors of the company decides to hold a competition and make the problem of click prediction open to experts like you. To win the prize, you need to fight for the leading positions 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 Sets

The problem is formalized as a binary classification problem, where the goal is to predict the click "truth" of each (user, ad) pair accurately. We will have two tracks of competition. The details of the tracks, which would differ by evaluation criteria (i.e. error functions), will be announced later. The data will be divided to the training set and the test set. For the test set, the click "truth" will be hidden.

The training data contain 8 full days (day 0 to day 7) of data, and 7 half days for the mornings of day 8 to day 14. The test data contain the other 7 half days for the afternoons of day 8 to day 14. Each row of the training data starts with a timestamp, followed by an advertisement id, the click "truth" (0 for no click and 1 for click), and the associated user features. The user features are binary, and only the indices of the value-1 features are listed in the row. The test data are of the same format as the training data, except that the click "truth" is hidden.

In practice, you won't be able to peep the day-9 morning data when predicting in the afternoon of day 8. For this competition, we decide not to have such restrictions. But it is **strongly encouraged** for you to check how much your model benefits from "peeping" future data, and recommend the best approach based on **realistic (no-peeping) scenario**.

The data sets are processed from the Yahoo! R6B data, which aims for predicting whether a user would be interested in some news article. To maximize the level of fairness, you are not allowed to download the original Yahoo! data at any time. But you are welcomed to go check the descriptions of the data.

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 **for each track** 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 replicability. 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 want 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 board 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 06/19/2017. We'll have a mini-ceremony to honor the best team(s) on 06/20/2017. 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 CEIBA. You do not need to submit a hard-copy. The report is due at noon on 06/27/2017.
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 replicability.

Source Code: You do not need to upload your source code for the final project. Nevertheless, please keep your source code until 08/01/2017 for the graders' possible inspections.

Grade: The final project is worth 400 points. That is, it is equivalent to two usual homework sets. At least 360 of them would be reserved for the report. The other 40 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. Using any tricks to query the labels of the test set is strictly prohibited.