Final Project

TA email: ml2012ta@csie.ntu.edu.tw

RELEASE DATE: 11/19/2012

COMPETITION END DATE: **01/06/2013 NOON ONLINE**

REPORT DUE DATE: **01/14/2013 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 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 search engine company that has been providing advertisements along with the searched results for years. In order to improve the service quality and make more profits, the company hopes to predict whether an ad will be clicked or not. The company analyzed the collected session logs, and applied some different machine learning algorithms to the logs. Nevertheless, the company is now facing a bottleneck in improving the performance. Therefore, the board of directors decided to hold a competition and make the problem 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 bipartite ranking problem, where the goal is to rank the relevant (clicked) examples higher than the irrelevant (unclicked) ones. You will be provided with examples of the form $(y, \mathbf{x})$, where $\mathbf{x}$ is the feature vector of some particular ad, and y is the label, with $+1$ indicating an relevant (clicked) example and -1 otherwise. There are $40,000$ training examples, and $50,000$ test ones.

Each component in $\mathbf{x}$ is either a predicted score of the ad or some characteristic that is related to the ad. The predicted score comes from some base machine learning algorithm that the company has applied before, with higher score indicating higher relevance.

There is another challenge with this data set, MISSING VALUE. In particular, some of the scores or characteristics may be missing for some of the ads. Those missing ones would be marked "**?**" in the data sets, in both training and testing.

Your goal is to give real-valued scores $g(\mathbf{x})$ to each test ad $\mathbf{x}$, and the score of a relevant ad should be ideally higher than the score of an irrelevant one. There are several possible approaches that come from the tools you have known. For instance, scoring test ads by a plain linear regression from $\mathbf{x}$ to y is a possibility; using $P(y|\mathbf{x})$ from logistic regression is another possibility; converting the goal of bipartite ranking to an equivalent goal of predicting $\text{sign}(y - y')$ from a pair of examples $(\mathbf{x}, \mathbf{x}')$, hence allowing you to reuse the binary classification tools that you have learned, is yet another possibility. The board welcomes you to be creative and propose solutions beyond those natural possibilities, of course.

As a final remark, the competition data set is derived from KDD Cup 2012 Track 2. *To maximize the level of fairness, you are not allowed to download the original data set from the website at any time.*

Evaluation Criterion

The evaluation criterion for this competition is the Area Under Curve (AUC). The essence of the criterion corresponds to the following error defined on two ads:

$$\mathsf{e}(g, \mathbf{x}, y, \mathbf{x}', y') = \Big[\!\Big[[\![y > y']\!] \neq [\![g(\mathbf{x}) > g(\mathbf{x}')]\!] \Big]\!\Big]$$

The details of the criterion can be found here:

`http://home.comcast.net/~tom.fawcett/public_html/papers/ROC101.pdf`

We will adopt Algorithm 3 in the paper above, and will release a script for computing the AUC on the competition website.

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 replicability. Thus, in addition to the outlines above, you should also describe how you pre-process your data; 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 choose) 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.

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 on 11/26/2012, a lucky day. Each team can freely submit the predictions on all $50,000$ test examples. But 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 AUC score evaluated on a randomly chosen (but fixed) $25,000$ out of the $50,000$ test examples. The board will secretly evaluates you on the other $25,000$.

The competition ends at noon on 1/6/2013. We'll have a mini-ceremony to honor the best team(s) on 1/7/2013. 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 1/14/2013.
Medal: Medals cannot be used on the final project.
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 package 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 2/28/2013 for the graders' possible inspections.

Grade: The final project is worth 600 points. That is, it is equivalent to three 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 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.