Final Project

TA email: `ml2013ta@csie.ntu.edu.tw`

RELEASE DATE: 11/29/2013

COMPETITION END DATE: **01/05/2013 NOON ONLINE**

REPORT DUE DATE: **01/16/2014 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 startup company that features a coming product on the mobile phone. The core of the product is a robust character recognition system. The company has collected some written characters in order to build a prototype of the product, and wants to demo the robustness of the prototype. The board of directors of the company decided to hold a competition and make the problem of building the prototype 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 multiclass classification problem, where the goal is to classify all the characters accurately. You will be provided with examples of the form $(y, \mathbf{x})$, where $\mathbf{x}$ is the pixels of some character, and y is the label. The prototype uses the Chinese zodiacs as the possible characters, and hence there are 12 possible labels.
The characters gathered by your company contains a big data set of size 12288. The board has decided to reserve 6144 of them as test examples, and *you are not allowed to peep/label the true answers of those.* The following file contains half of the test examples.

`http://www.csie.ntu.edu.tw/~htlin/course/ml13fall/project/proj_test1.zip`

The other half of the test examples will be released a few days before the competition deadline.
The other 6144 is taken as the training set that you can freely use.

`http://www.csie.ntu.edu.tw/~htlin/course/ml13fall/project/proj_train.zip`

To maximize the level of fairness, you are not allowed to manually label the test examples or write (and add) any additional characters at any time.

Evaluation Criterion

The evaluation criterion for this competition is simply the classification accuracy, which is $1 - E_{0/1}(g)$.

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 before 12/9/2013, a lucky day. Each team can freely submit the predictions on the first batch of 3072 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 classification accuracy. A few days before the competition deadline, another 3072 test examples will be released, and the board will evaluate you on those.

The competition ends at noon on 1/5/2014. We'll have a mini-ceremony to honor the best team(s) on 1/6/2014. 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 1/16/2014.
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 2/28/2014 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.