

Final Project

TA in charge: Hanhsing Tu and Ming-Feng Tsai

RELEASE DATE: 12/18/2008

DUE DATE: 01/15/2009, 4:00 pm IN ROOM 536

TA SESSION: by appointment

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

You should write your report in English with the common math notations introduced in class. We do not accept reports written in any other languages.

Introduction

The main theme of the final project is a machine learning competition. Imagine that you are a manager who leads a research team in a data analysis company. Recently, your company receives an important case that is worth millions of dollars. Then, the board of directors asks each research team to study some machine learning approaches for dealing with the case in order to provide concrete recommendations. To get more year-end bonus and future research funds, you have to offer a comprehensive report based on your professional expertise. The report will be evaluated not only by the prediction performance of the recommended approaches, but also by the reasoning behind your recommendations.

Data Sets

The case received by your company contains a big data set of size 50000. The board has decided to reserve 10000 of them as test examples, and *you are not allowed to peep the true answers of those*. The following file contains the examples without the answers (labels).

http://www.csie.ntu.edu.tw/~htlin/course/ml08fall/data/proj_test.dat

For the other 40000, the board separates it to the following three training sets:

- LARGE: containing all those 40000 examples

http://www.csie.ntu.edu.tw/~htlin/course/ml08fall/data/proj_large_train.dat

- MEDIUM: containing 4000 out of LARGE

http://www.csie.ntu.edu.tw/~htlin/course/ml08fall/data/proj_medium_train.dat

- SMALL: containing 400 out of MEDIUM

http://www.csie.ntu.edu.tw/~htlin/course/ml08fall/data/proj_small_train.dat

The data sets above are processed from and take the same formats (including the order of features) as the original “Quantum Physics” data set here:

<http://kodiak.cs.cornell.edu/kddcup/datasets.html>

You can browse through the site to know more about the task. Nevertheless, to maximize the level of fairness, *you are not allowed to download the original data set from the website at any time.*

Survey Report

You are asked by the board to study at least THREE machine learning approaches using any (or all) of the three training sets above. Then, you should make a comparison of those approaches according to some different perspectives, such as accuracy, efficiency, scalability, popularity, and interpret-ability.

Based on the results of your comparison, you are asked to choose THE BEST ONE of those approaches as your final recommendation, and provide the “cons and pros” of the choice.

The survey report should be less than or equal to 4 pages. The single 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; 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.

For grading purposes, a minor but required part in your survey report for a two-people team (see the rules below) is how you balance your work loads.

Competition

To heat things up, the board has decided to set up an in-company competition. In the competition, the goal is simply to beat other teams’ approaches in terms of *accuracy*. There are three tracks of competitions, one corresponding to each training set above.

The submission site for each track would be announced on 12/26/2008. Each team can freely submit the predictions on all 10000 test examples as an entry for each track. 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 for each track showing the accuracy evaluated on a randomly chosen (but fixed) 5000 out of the 10000 test examples. The board will secretly evaluates you on the other 5000.

The competition ends at 11:59 pm on 1/7/2009. There will be a mini-ceremony for honoring the best team(s).

Misc Rules

Teams: By default, you are asked to work as a team of size TWO. A one-person team is allowed only if you are willing to be as good as a two-people team. It is expected that both team members share balanced work loads. Any form of unfairness in a two-people team, such as the intention to cover the other member’s work, is considered a violation of the honesty policy and will cause both members to receive zero 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 reproducibility.

Source Code: You do not need to upload your source code for the final project. Nevertheless, please keep your source code until 2/28/2009 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 (such as the original one on the website) 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 strictly prohibited, too.