Final Project

TA email: ml2011ta@csie.ntu.edu.tw

RELEASE DATE: 12/5/2011

DUE DATE: **EXTENDED TO 01/13/2012 4PM 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

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 43907. The board has decided to reserve 23907 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/ml11fall/data/proj_test.zip`

The other 20000 is taken as the training set that you can freely use.

`http://www.csie.ntu.edu.tw/~htlin/course/ml11fall/data/proj_train.zip`

The data sets above are processed from the original "mediamill" data set that you can get on public websites. Each line of the training set (after unzipping) represents an example that takes the form:

`feature 1, feature 2, feature 3, ⋯, feature 120, LABELS`

`LABELS`? You are asking. Yes, the data set is a *multi-label classification* one.

`http://mlkd.csd.auth.gr/multilabel.html`

That is, unlike traditional single-label classification tasks, each example can now be associated with zero, one, two, or multiple labels. For instance, an article can belong to both the `POLITICS` and `ECONOMICS` category, if it happens to be talking about ECFA. When the labels are $\{1, 2, 5\}$, the `LABELS` part would take the form $(1, 1, 0, 0, 1, 0, \cdots, 0)$, which contains 1 only in the $\{$`1,2,5`$\}$-th positions and 0 otherwise.

Each line of the training set (after unzipping) contains 120 features and a length-101 binary vector for the multiple labels. Each line of the test set contains 120 features.

Perhaps the simplest way of doing multi-label classification is to convert it into multiple binary classification problems. That is, for each category k, build a binary classifier to decide whether an example $\mathbf{x}$ belongs to k. Such a simple approach can always be your starting point, but feel free to brainstorm for more, or climb on the giant's shoulder on the website above.

There is also another challenge with this data set, MISSING VALUES! Due to the company's poor data management, a random 10% of the training features are missing (while the test features are intact). Missing values are indicated by '?' in the data set. You are encouraged to develop creative methods to deal with the missing values.

One final remark: To maximize the level of fairness, you are not allowed to download the original data set from the website at any time.

Evaluation Criteria

We will consider two evaluation criteria for the multi-label classification tasks. The first one is called the Hamming loss. For the answer multi-label binary vector $\mathbf{y}$ and the prediction binary vector $\tilde{\mathbf{y}}$, the Hamming loss simply computes the number of different positions betwen $\mathbf{y}$ and $\tilde{\mathbf{y}}$. For instance, if $\{1, 2, 3\}$ are the actual labels ($\mathbf{y} = (1, 1, 1, 0, \cdots, 0)$) and your algorithm produces $\{2, 4\}$ ($\tilde{\mathbf{y}} = (0, 1, 0, 1, 0, \cdots, 0)$), the Hamming loss of the prediction is 3.

A more complicated criterion is the F1-score. You can check

`http://en.wikipedia.org/wiki/Precision_and_recall`

for details. We will provide a script for computing the F1-score on the 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 interpret-ability. In addition, for each of the two criteria, you need to recommend THE BEST ONE of those approaches as your final recommendation for such a criterion, and provide the "cons and pros" of the choice.

The survey report should be less than or equal to SIX A4 pages with readable font sizes. The one 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. The comptition contains two tracks, each corresponding to one of the evaluation criterion.

The submission site would be announced on 12/12/2011. Each team can freely submit the predictions on all 23907 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 23907 test examples. The board will secretly evaluates you on the other 18907.

The competition ends at noon on 1/8/2012. We'll have a mini-ceremony to honor the best team(s).

Misc Rules

Report: Please upload your report electronically on CEIBA. You do not need to submit a hard-copy.
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 2/28/2012 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.