The University of Adelaide, School of Computer Science

Introduction to Machine Learning

Semester 2, 2015 Assignment 2: Implementation of AdaBoost

DUE: 6 Oct. 2015, Tuesday 11:55 PM

Submission

Instructions and submission guidelines:

- You must sign an assessment declaration coversheet to submit with your assignment. The assessment declaration coversheet is included in this package “coversheet.pdf”.

- Submit your assignment via the Course (Moodle) Forum.

Reading

With this assignment, you will see how Adaboost works on a classification task. The AdaBoost algorithm is described in the class and more informtion on AdaBoost can be found on the web pages: https://en.wikipedia.org/wiki/AdaBoost

Please read “A Short Introduction to Boosting” by Yoav Freund and Robert E. Schapire, which can be found here: http://www.cs.princeton.edu/~schapire/papers/FreundSc99.ps.gz

and,

http://rob.schapire.net/papers/explaining-adaboost.pdf

If you find difficulties to understand this paper, you may read other tutorial/survey papers on the same webpage. You are also encouraged to read my recent paper which can be downloaded here: http://arxiv.org/abs/0901.3590

Coding

You are provided with the training data (xi; yi); i = 1...., belonging to two classes, with label yi is

{+1, -1}. You should use these training data to train an Adaboost classifier.

The provided dataset is: “breast-cancer.mat”. You can load the data in Matlab use the command “load”. After you load the data in Matlab, data.X contains the training data. Each row is a data point in data.X and data.y is the corresponding label.

Please implement the AdaBoost algorithm as given on page 3 of the Freund and Schapire paper. The algorithm requires that you train a weak learner on data sampled from the training set. While I expect you to design your AdaBoost program in such a way that you can plug in any weak learner, I would like you to use Decision Stumps for this assignment. Decision Stumps are simply one-level decision trees. That is, the learner selects an attribute for the root of the tree and immediately classifies examples based on their values for that attribute.

To simplify the task, I have also provided a Matlab implementation of Decision Stump (“build_stump.m”). This is for reference only. Please be aware that you may need to modify the decision stump code for your own needs.

Please start early. This might be a tough algorithm to implement and debug. You can choose either Matlab or C/C++ to implement AdaBoost. I would personally suggest Matlab. If you use C/C++, you can use Matlab to convert the “breast-cancer.mat" data into some other format that is easier to read in C/C++.

Your code should not rely on any 3rd-party toolbox. Only Matlab's built-in API's or C/C++'s standard libraries are allowed. When you submit your code, please report your algorithm's training error on the given dataset.

You are also required to submit a report (2 to 4 pages in PDF format), which should have the following sections:

- An algorithmic description of the AdaBoost method.
- Your understanding of AdaBoost (anything that you think is relevant to this algorithm)
- Some analysis of your implementation. You should include the training error curve against the number of iterations on the provided “breast- cancer” data set in this part.
- You may train a SVM and compare the results of SVM with AdaBoost. What do you observe?

In summary, you need to submit

(1) the code that implements AdaBoost and

(2) a brief report (2~4 pages in PDF).