CSL407 Machine Learning
Homework 5

Due on 10/27/2014, 11.55pm

Instructions: Upload to your moodle account one zip file containing the following. Please do not submit hardcopy of your solutions. In case moodle is not accessible email the zip file to the instructor at ckn@iitrpr.ac.in. Late submission is not allowed without prior approval of the instructor. You are expected to follow the honor code of the course while doing this homework.

1. **You are allowed to work in teams of size at most 2 for this homework. The submission of the zip file can be made by one of the team members.**
2. A neatly formatted PDF document with your answers for each of the questions in the homework. You can use latex, MS word or any other software to create the PDF. Include the names of the team members and the roll numbers in the pdf document.
3. Include a separate folder named as 'code' containing the scripts for the homework along with the necessary data files. Name the scripts using the problem number.
4. Include a README file explaining how to execute the scripts.
5. Name the ZIP file using the following convention **rollnumber1_rollnumber2_hwnumber.zip**

In this homework you will be experimenting with Ensemble Learning, in particular AdaBoost and its variant called TrdaBoost that is used in the Transfer Learning setting. You will use the linear SVM implementation from the previous homework.

1. We will experiment with AdaBoost algorithm using dataset1.mat included in the zip file.
   - Implement the code to perform a 10- fold stratified cross validation on the dataset.
   - Implement the AdaBoost algorithm described in Algorithm 1 . You can use a linear support vector machine as your base learner (weak classifier). You will have to devise a way of sampling from the distribution D_t to learn the weak hypothesis for the t^{th} iteration. Set the maximum number of iterations to be 500. For every iteration of the AdaBoost algorithm store the training accuracy of the combined hypothesis learned till that iteration. Average the training accuracy for every iteration across the multiple folds and plot this data. Also report the test accuracy for each fold and the average accuracy across all the 10 folds.
   - Compute the confusion matrix for each fold and the average across all folds.
   - Compute the precision and recall values for both classes for every fold and average it across all folds.
2. We will now experiment with a variant of `AdaBoost` called `TrAdaBoost` that is used in Transfer Learning setting. The paper 'daiicml2007.pdf' included in the zip folder discusses the algorithm and the the transfer learning setting in which it is employed.
   The goal of TrAdaBoost is to improve the performance of a learning task in a target domain using information from a source domain that is different yet related to the target domain. It requires availability of labeled data from the source and target domains. The algorithm assumes difference

Algorithm 1 Adaptive Boosting Algorithm

Given $(x_1, y_1), \ldots, (x_N, y_N)$, where $x_i \in X$ is a data point and $y_i \in Y = \{-1, +1\}$ is the class label associated with the data point x_i, a base learning algorithm Weak Learner, the maximum number of iterations T.

Initialize Initialize the weights $D_1 = \{D_1(1), D_1(2), \ldots, D_1(N)\}$. If the user does not specify an initial weight vector, assign $D_1(i) = \frac{1}{N} \forall i$.

For $t = 1, \ldots, T$

(a) Find the classifier $h_t : X \to \{-1, +1\}$ that minimizes the error ϵ_t with respect to the distribution D_t.

$$\epsilon_t = \sum_{i=1}^{N} D_t(i)[y_i \neq h_t(x_i)] \tag{1}$$

where

$$[y_i \neq h_t(x_i)] = \begin{cases} 0 \text{ if } y_i = h_t(x_i) \\ 1 \text{ if } y_i \neq h_t(x_i) \end{cases} \tag{2}$$

(b) if $\epsilon_t \geq 0.5$; stop.

(c) Set $\alpha_t = \frac{1}{2} \ln \frac{1-\epsilon_t}{\epsilon_t}$

(d) Update the weights for all the data points $i = \{1, \ldots, N\}$

$$D_{t+1}(i) = \frac{D_t(i) \exp\left(-\alpha_t y_i h_t(x_i)\right)}{Z_t} \tag{3}$$

where Z_t is the normalization factor

$$Z_t = \sum_{i=1}^{N} D_t(i) \exp\left(-\alpha_t y_i h_t(x_i)\right) \tag{4}$$

Output The final hypothesis

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right) \tag{5}$$

---

in the data distribution between source and target domain. It adapts the boosting algorithm to increase the weights of source domain examples that are useful for learning the target task, while at the same time reducing by a constant factor the weights of source examples are very different from the target domain data. The pseudocode in the paper assumes a binary classification task. Detailed understanding of the theoretical underpinnings of the `TrAdaBoost` is not required for implementation of the algorithm.

We will use newsgroup data (also included in the zip folder as `recvstalkmini.mat`) for learning the source and target domain tasks. This data has been sampled from the larger 20-Newsgroups dataset[1]. The 20 Newsgroups dataset is a collection of approximately 20,000 newsgroup documents, partitioned across different newsgroups. It consists of seven top level categories, with each category consisting of multiple sub-categories. For this homework, we describe the task to be that of classifying the top level category, with the source and target domain data being drawn from different subcategories. Thus the two datasets consists of different subcategories resulting in a difference in their distribution. Due to computational constraints, we randomly pick a smaller number of samples for the source and target domain dataset for the reduced dataset.

- Experiment with different amount (1%, 5%, 10% and 20%) of label data from the target domain. Use accuracy on the remaining target data as the performance measure. Average the accuracy across 10 random folds. Tabulate the results from these experiments (average accuracy and standard deviation).

[1] http://qwone.com/ jason/20Newsgroups/