Introduction to Machine Learning - Assignment 2

Instructor: Dan Lizotte

Due at the beginning of class on Monday, 7 May 2007

This assignment will familiarize you with WEKA which is a publicly available machine learning tool. It is marked out of 50 and is worth 15% of your final mark.
For this assignment, submit a hard copy of all of your answers, and e-mail your answer (i.e. labels) for the `hw2-br` dataset to `dlizotte@cs.ualberta.ca`.

---

1. [*50 points*] **Comparing Algorithms, Plotting Learning Curves** — WEKA
   In this assignment, you will use the WEKA system to compare the performance of different learning algorithms on two datasets, hw2-1 (2 classes) and hw2-2 (multiple classes), when using different amounts of training data. We will also have a contest for who can learn the best classifier for hw2-br (2 classes).

   The JAVA code for WEKA is available from

   `http://www.cs.waikato.ac.nz/ml/weka/`

   The data sets are available at

   `http://www.cs.ualberta.ca/~dlizotte/teaching/assignments/hw2_data.tar.gz`

   You will run each the following specific learning algorithms:

   - **Decision Stump** select "DecisionStump" under "trees"
   - **Decision Tree, No Pruning:** select "J48" under "trees" - you will have to change one setting
   - **Decision Tree, With Pruning:** select "J48" under "trees" - use the default settings
   - **Support Vector Machine:** select "SMO" under "functions"
   - **Naive Bayes:** select "NaiveBayesSimple" under "bayes"

   You will need to refer to the following website for detailed instructions on how to use WEKA:

   `http://www.cs.ualberta.ca/~dlizotte/teaching/assignments/WEKA.html`

   This webpage describes exactly how to choose a training set, a test set, and an algorithm so that you can generate the output you need for this assignment. You should apply each of the algorithms to each of the data sets `hw2-1`, and `hw2-2`, and report the test error for each combination of data set, algorithm, and number of training examples.. Each of these data sets has several training data files and one test data file, each of the form `foo-train#.arff` for training and `foo-test#.arff` for testing. The number in the file name specifies the number of instances in that dataset.

The complete list of data files is:

```
hw2-1 data files
      hw2-1-train10.arff        10 training examples
      hw2-1-train20.arff        20 training examples
      hw2-1-train50.arff        50 training examples
      hw2-1-train100.arff       100 training examples
      hw2-1-train200.arff       200 training examples
      hw2-1-test100.arff        test data file

hw2-2 data files
      hw2-2-train25.arff        25 training examples
      hw2-2-train50.arff        50 training examples
      hw2-2-train100.arff       100 training examples
      hw2-2-train600.arff       600 training examples
      hw2-2-test200.arff        test data file

hw2-br data files:
      hw2-br-train300.arff        br training data file
      hw2-br-test300unl.arff      br test data file
```

(a) For each of `hw2-1` and `hw2-2`, you should turn in a table, whose rows are each labeled with the size of the training set, and whose columns are labeled with the learning algorithm. Each entry in the table should contain the test error rate achieved by that learning algorithm when trained on the given number of examples. (A test set of 100 examples is provided for each data set.)

`hw2-1:`

N	DecisionStump	DT-NoPruning	DT-WithPruning	SMO	NaiveBayes
10	err	err	err	err	err
20	err	err	err	err	err
50	err	err	err	err	err
100	err	err	err	err	err
200	err	err	err	err	err

`hw2-2:`

N	DecisionStump	DT-NoPruning	DT-WithPruning	SMO	NaiveBayes
25	err	err	err	err	err
50	err	err	err	err	err
100	err	err	err	err	err
600	err	err	err	err	err

(b) For the "bragging rights" dataset `hw2-br`, your goal is to produce the best possible classifier. Here, you may run any algorithm you wish on `hw2-br-train300.arff` — including any of the ones presented above (now with ANY settings), or any others. After you have decided on the appropriate classifier, you should then apply that classifier on the *unlabeled* `hw2-br-test300` data, to produce a list of labels, `br-labels.txt`. (This should be the raw outputs from WEKA – check the WEKA.html instruction page for details.) Write a short description of the classifier you chose to use. There will be *fabulous prizes* for the winners!