CIS 732/830 Advance in AI

Spring 2007

Homework 2 of 8: Machine Problem (MP2)

Assigned Fri 09 Feb 2007
Due: before midnight Fri 23 Feb 2007

The Badges Game (data courtesy of Haym Hirsh, Rutgers; problem statement adapted from Dan Roth's machine learning course, CS446, at the University of Illinois).

Pre-registered attendees of the 1999 International Conference on Machine Learning (ICML-1994) and 1994 Computational Learning Theory Conference (COLT-1994) each received a badge labeled with a "+" or "-". The labeling was due to some function known only to the badge generator (Haym Hirsh), and it depended /only/ on the attendee's name. The goal for conference attendees was to identify the unknown function used to generate the +/- labeling.

The list of attendees along with their labels, is at:
http://l2r.cs.uiuc.edu/~danr/Teaching/CS446-06/badges.org
The data is presented in the form tuples: a +/- label followed by the person's name. There are 294 examples, 210 positive and 84 negative.

Your task is to do the same: identify the unknown function used to generate the +/- labeling.

In doing so, think about how would you formalize it as a learning problem and what are the difficulties that arise in doing it. You may use inducers in the Waikato Environment for Knowledge Analysis (WEKA). Start by thinking about how will you go about doing it and what the difficulties may be.

1. (25%) Prepare a training data set for WEKA 3.5.x by writing a program to convert the badges data to Attribute-Relational File Format (ARFF):

   http://www.cs.waikato.ac.nz/~ml/weka/arff.html
   http://weka.sourceforge.net/wekadoc/index.php/en:ARFF_%283.5.1%29

   Turn in your code and the file badges.arff with a header containing brief documentation as shown in the ARFF docs and WekaDoc Wiki.

2. (25%) Run the data through the following inducers first:

   J48

Multilayer Perceptron (MLP)
Naive Bayes

and collect SUMMARY information describing your results. For example, report the number of nodes in a decision tree and the confusion matrix (frequency for desired vs. actual, positive vs. negative class labels), but keep the actual tree in a separate text file, which you should also turn in. Look up PRECISION and RECALL and report these values.

Discuss how the decision tree and MLP output succeed or fail in expressing a human-comprehensible concept for the Badge Problem. Refer to descriptive statistics for the actual hypotheses learned.

3. (25%) Prepare a similar data set for NeuroSolutions (http://www.nd.com). Turn in any scripts or programs you used to process the raw data along with the .asc file.

4. (25%) Run the data through a Multilayer Perceptron, trying Step and Momentum training rules and several different numbers of hidden units. Report the confusion matrices in the same style as WEKA, and indicate what configuration (ANN parameters) gave you the best performance.

Extra credit (5% each). Collect the following information by preparing the same data set for MLC++:

a) Output data from the ID3 inducer in MLC++ and compare this to J48 in WEKA 3.5.x.

b) Plot an example learning curve using the above data in increments of 10.

c) Run the Feature Subset Selection (FSS) wrapper with ID3 as the base inducer. Compare this against the feature selection wrapper inducer in WEKA with J48 as the base inducer.

d) How would you discover the Badge concept from scratch? (Hint: think about hybrid approaches and inductive approaches such as genetic programming that are purported to learn the functional form of the concept.)