CIS 732

Machine Learning and Pattern Recognition
Fall, 2001

Homework Assignment 2 (Machine Problem)

Sunday, 30 September 2001
Due: Thursday, 18 October 2001
(before midnight Friday 19 October 2001)

Refer to the course intro handout for guidelines on working with other students.

Note: Remember to *submit your solutions in electronic form using* ***hwsubmit*** and *produce them only from your personal source code, scripts, and documents from the machine learning applications used in this MP* (not common work or sources other than the textbook or *properly cited* references).

You have about 3 weeks to complete 3 parts to this machine problem (MP), so please start early and finish about one part per week. The point value of each part is an approximate indicator of difficulty (your personal assessment can and should vary). Problem 3 is considerably harder because you are being asked to write you own code.

Problems

First, log into your course accounts on the KDD Core (Ringil, Fingolfin, Yavanna, Nienna, Frodo, Samwise, Merry, Pippin) and make sure your home directory is in order. Notify admin@www.kddresearch.org (and cc: cis732ta@www.kddresearch.org) if you have any problems at this stage.

1. (20 points total) **Running ID3 in MLC++.**
   In your web browser, open the URL
   http://www.kddresearch.org/Courses/Fall-2001/CIS732/Homework/Problems/MP2/
   and download the file
   MLC++-2.01.tar.gz
   to your local system (this can be a Windows, Unix, Mac, or other system, but the binaries are precompiled for ix86 Linux. Follow the instructions in the *MLC*++ manual (Utilities 2.0, in your first notes packet and at http://www.sgi.com/tech/mlc) for installing it MLC++ your home directory.

   a) (8 points) *Your solution to this problem must be in MS Excel, PostScript, or PDF format, and you must use a spreadsheet (I recommend GNUmeric or Excel 2000/XP) to record your solution.* Follow the instructions in the *MLC++ Utilities 2.0 User Guide* (also in your first notes packet) to run the *ID3* inducer on the following data sets from the UCI Machine Learning Database Repository: *Credit (CRX), Monk1, Mushroom, Vote*. Use the .test files for testing. Turn in the ASCII file containing the decision tree and another file (.xls, .ps, or .pdf) containing a table of test set accuracy values for each data set. (For the next machine problem, you will compare the ID3 results – accuracy, overfitting, example learning curves – with Simple Bayes and C4.5.)

b) (12 points) Repeat this process with the Feature Subset Selection (FSS) inducer, which you can read about in the MLC++ user guide. The wrapped inducer should be ID3. Report *both test and training* accuracy. Think carefully about how to generate training set accuracy.

2. (32 points total) **Running Feedforward ANNs in NeuroSolutions.** Download the *NeuroSolutions 4* demo from http://www.nd.com and install it on a Windows 98/Me/XP Home or NT4/2000/XP Pro machine. NS4 is installed on the "hobbits" (4 Pentium Pro workstations dual-booting Windows 2000 Professional and Red Hat Linux 6.2, located in 227 Nichols Hall), and you may log in with your CIS login to use them.
   a) (10 points) Use the NeuralBuilder wizard (which is fully documented in the online help for *NeuroSolutions 4*) to build a multilayer perceptron for learning the sleep stage data provided in the example data directory. Your training data file should be *Sleep1.asc* and your desired response file should be *Sleep1t.asc*. Use a 15% holdout data set for cross validation. **Report both training and cross validation performance (mean-squared error)** by selecting the appropriate probes in the wizard or stamping them from the tool palettes, and recording the final value after training (for 2000 epochs, twice the default). Replace the sigmoidal activation units with linear approximators to the sigmoid transfer function. Finally, double the number of hidden layer units. Turn in a screenshot showing the revised network, the progress bar, and the MSE values after training.
   b) (8 points) Train a Jordan-Elman network for the same task and report the results. Use the default settings and the input recurrent network (the upper left entry among the 4 choices). Take a screen shot of your artificial neural network after training (in Windows, hit Print-Screen and paste the Clipboard into your word processor).
   c) (8 points) Train a time-delay neural network for the same task and report the results.
   d) (6 points) Train a Gamma memory for the same task and report the results.

3. (48 points) **Implementing Simple Bayes.** To be posted; see the MP2-4 version from Fall, 1999. The specification for this problem shall match exactly. *There will be a follow-up using your code in later MPs, so it is a good idea not to skip this one.*

Extra credit

a) (5 points) **Class Participation.** Post your turn-to-a-partner exercise from class on Thu 04 Oct 2001 in the class web board.

b) (5 points) Try the *MATLAB Neural Network* toolkit on *Sleep1* and report the same results for a feedforward ANN (specifically, a multi-layer perceptron) trained with backprop. This package can be found on the KDD Core systems, including a Windows version installed on the Hobbits.