Homework 3
Problem Set
Assigned Sun 25 Feb 2007
Due Fri 09 Mar 2007

This written assignment is designed to get you started thinking about your term project, and to give you some practice with the basics of statistical evaluation of hypotheses and computational learning theory.

For problems 1-2, refer to the Santa Fe Time Series data sets archived at Andreas Weigend's site:
http://www-psych.stanford.edu/~andreas/Time-Series/SantaFe.html

and to NeuroSolutions:
http://www.nd.com

You may download an evaluation copy of NeuroSolutions 5 from NeuroDimension, or use the KDD Lab's single developer copy of NeuroSolutions 4.

1. (20%) Time series data, representation, and learning problem definition. Download all six data sets and put them into .asc format for NeuroSolutions. Discuss the limitations of the same representation if used to produce data in WEKA's Attribute-Relational File Format (ARFF). State a learning problem specification (input, target output, and evaluation criterion).

2. (20%) Backpropagation through time. Run the following networks on the first data set to PREDICT A CONTINUATION of the data.
a) Input recurrent (part of the "simple recurrent network" group; to be discussed in lecture)
b) Elman recurrent network
c) Time-delay neural network
d) Gamma memory
Turn in a saved .nsb file containing the breadboard for each artificial neural network and a screenshot of the training curve.

For problem 3, refer to Chapter 5 of Mitchell's _Machine Learning_ and the following online references:
http://en.wikipedia.org/wiki/P-value
http://en.wikipedia.org/wiki/Statistical_significance
http://en.wikipedia.org/wiki/Student%27s_t-distribution
http://en.wikipedia.org/wiki/T-test

In PS5, we will continue this exercise and also calculate confidence intervals.

http://en.wikipedia.org/wiki/Confidence_intervals

3. (20%) Statistical evaluation of hypotheses.

a) (5%) For the above networks, calculate the mean and variance of cross-validation error with 5 CV folds, over 10 runs. You may use the MacroWizard (and export Visual Basic for Applications code) to automate your data collection, or simply use a DataWriter and process the results using a spreadsheet.

b) (5%) Use Gnuplot, OpenOffice, or Microsoft Excel to plot the results with error bars (mean, first and third quaartiles).

c) (10%) Calculate the p value of the following hypotheses:

- Input recurrent networks outperform Elman recurrent networks
- Input recurrent networks outperform time-delay neural networks
- Gamma memories outperform time-delay neural networks
- Gamma memories outperform input recurrent networks

This should be done using a Student t-test with the data from PS3-3a.

4. (20%) Computational learning theory. In your own words, state and prove the VC dimension of:

a) perceptrons (linear threshold gates) on two inputs

b) axis-aligned rectangles in R^2

c) half-intervals in R

d) full intervals in R

5. (20%) Term project planning. Select your term project from among the 3 areas and 6 data sets specified in class the week of Mon 26 Feb 2007. Write a short (1 page) statement defining YOUR

a) input data (how it is generated and prepared from the raw data)

b) training target

c) objective criteria, loss function(s), or error measure(s)

d) mode of machine learning (supervised, semi-supervised, unsupervised, reinforcement) and (potential) methods to be applied

In Machine Problem 4, you will practice with Evolutionary Computation in Java (ECJ) and a system for Inductive Logic Programming (ILP).