Homework # 1

DUE TUESDAY, OCTOBER 10, 2000, AT 2:30 PM

Collaboration in the sense of discussions is allowed, but you should write the final solutions alone and understand them fully. Do not read class notes or homework solutions from previous years at any time. Other books and notes can be consulted, but not copied from. You should justify your answers, at least briefly. Definitions and notation follow the lectures.

The handouts and data for the homeworks can be found at:

http://work.caltech.edu/cs156/00/homeworks.htm

1. Bins and Marbles

Consider a sample of 10 marbles drawn from a bin that has red and green marbles. The probability that any marble we draw is red is π (independently). For $\pi = 0.05$, $\pi = 0.5$, and $\pi = 0.8$, we address the probability of getting no red marbles ($\nu = 0$), in the following cases.

(i) We draw only one such sample. Compute the probability that $\nu = 0$.

(ii) We draw 1,000 independent samples. Compute the probability that (at least) one of the samples has $\nu = 0$.

(iii) Repeat (ii) for 1,000,000 independent samples.

2. Gradient Descent

Consider the nonlinear function $E(u, v) = (ue^v - 2ve^{-u})^2$. We start at the point $u = 1$ and $v = 1$, and take a step of length 0.1 in the u, v space with a view to minimizing E.

(i) Compute the values of $\frac{\partial E}{\partial u}$ and $\frac{\partial E}{\partial v}$ at the starting point.

(ii) If we descend along the gradient, compute the value of $\Delta E = E(u + \Delta u, v + \Delta v) - E(u, v)$ using the first-order approximation.

(iii) Repeat (ii), but without approximation.

(iv) If we descend along the u-coordinate instead, compute the value of ΔE (without approximation). How does it compare to (iii)?

(v) Repeat (iv) for the v-coordinate.

(vi) Compute the values of $\frac{\partial E}{\partial u}$ and $\frac{\partial E}{\partial v}$ at the new point $(u + \Delta u, v + \Delta v)$ in (ii). Compare these values to the values in (i). How is the difference related to the difference between (ii) and (iii)?

3. Backpropagation

Following the class notes, implement the backpropagation algorithm that takes as input a network architecture $(d^0 = d, d^1, d^2, ..., d^L = 1)$ and a set of examples $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)$ where $\mathbf{x}_n \in \mathcal{R}^d$ and $y \in \mathcal{R}$, and produces as output the network weights. The algorithm should perform gradient descent on one example at a time, but should also keep track of the average error for all the examples in each epoch. Try your algorithm on the data set in

http://www.work.caltech.edu/cs156/00/hw/hw1/train.dat

(the first two columns are the input and the third column is the output). Test the convergence behavior for architectures with one hidden layer ($L = 2$) and 1 to 5 neurons ($d^1 = 1, 2, 3, 4, 5$), with combinations of the following parameters:

(i) The initial weight values chosen independently and randomly from the range (-0.02,0.02), the range (-0.2,0.2), or the range (-2,2).

(ii) The learning rate η fixed at 0.01, 0.1 or 1.

(iii) Sufficient number of epochs to get the training error to converge (within reason).

Turn in your code and a single parameter combination that resulted in good convergence for the above architectures.

4. Generalization

Using your backpropagation program and data from problem 3, train different neural networks with $L = 2$ (an input layer, one 'hidden' layer, and an output layer) where the number of neurons in the hidden layer is 1, 2, 3, 4, or 5. Use the following out-of-sample data to test your networks:

http://www.work.caltech.edu/cs156/00/hw/hw1/test.dat

Plot the training and test errors for each network as a function of the epoch number (hence the 'intermediate' networks are evaluated using the test data, but the test data is not used in the backpropagation). Repeat the experiment by reversing the roles of the training and test sets (you may need to readjust the parameter combination from problem 3), and plot the training and test errors again. Briefly analyze the results you get.