CS 470/670 – Intro to Artificial Intelligence – Spring 2016

Instructor: Marc Pomplun

Assignment #5

Posted on April 21
Due by May 3, 5:30pm

Question 1: Classifying Iris Flowers

The Iris flower data set is a "classic" data set in the context of machine learning. A subset of it is contained in iris.txt in the Software section on the course homepage, together with the backpropagation demo code that was already uploaded earlier in the Syllabus section. Each row in the data file contains the data of one flower in the following order: Sepal length, Sepal width, Petal length, Petal width, and Species. The first four items are real numbers and the fifth one, the species, is either "versicolor" or "virginica." The file contains the data of 100 flowers, 50 of each species.

Your task is to program, train, and test an artificial neural network that can successfully tell from the four measurements of a given flower to which of the two species it belongs. First of all, you need to modify the code in the mini_backprop_demo.c file so that the network has four input neurons. You could also change the number of hidden-layer units, but just keeping the two original ones is probably a good place to start.

In the next step, add a function to the code that reads the iris.txt file, which is supposed to be in the same directory as the program file, and puts the data in an appropriate data structure. Then write some code that randomly picks 80 flowers for the training set and the remaining 20 ones for the test set and trains the network until the training error reaches a given threshold or the improvement in accuracy asymptotes. You can pick a criterion of your choice. Print the average training error on the console after each epoch. After the training has terminated, let the network classify the 20 test data and record the proportion of correct classifications. Perform this whole procedure ten times with different training vs. test set selections and let the program print the average accuracy across the ten runs.

When you place the input values into the input neurons, make sure that you scale the values in a reasonable way. Input values should ideally be in the range between 0 and 1. For the output, just keep the single output neuron, and use 0.05 as the desired output value for versicolor and 0.95 the desired output value for virginica. When you test the network with the test data, interpret any output < 0.5 as versicolor and all other outputs as virginica.

Write a text file explaining the changes that you made to the code and the average accuracy value that you obtained.

For CS670 students: What happens if you train with 20 flowers and test performance on the remaining 80? Do not submit the modified code but just report the average accuracy over 10 runs that you obtain.

Question 2: The Soccer Network

Here is an idea for an ANN that would make you rich if it performed well. This ANN predicts the results of soccer matches. The network receives information about the two competing teams and the conditions of the match and is supposed to predict how many goals each team will score. With this knowledge, you could bet on the projected winner team and gain a lot of money.

Let us say that every team consists of 20 players. You are providing the following input data to the network:

- The skill level of every player on each of the two teams. Skill is rated by a group of soccer reporters on a scale from 0 ("is unable to kick the ball") to 10 ("world class player").

- The number of matches that each team has played during the last two weeks. There are never more than seven matches in that period of time.

- The statistics of former matches between the same two teams within the past 10 years (e.g., Team A won 30% of the matches, Team B 45%, and 25% of the matches were tied).

- The continent that each team comes from (North America, South America, Europe, Africa, Asia, or Australia).

- Where the match takes place (Team A's stadium, Team B's stadium, or neutral place).

- The phase of the soccer season (early season vs. late season).

You want to build and train a backpropagation network that, based on this information, is able to predict the number of goals each team will score. Describe an appropriate way of formatting the input, interpreting the output, collecting exemplars, constructing the network, training the network, and testing the network. **Give reasons** for the decisions that you make. Describe everything **in great detail** so that a computer programmer who does not know anything about ANNs would be able to successfully build this network application, predict results, and become rich. The programmer can look up the BPN equations for training and operation in a book, but needs precise explanations for everything else. Please help him/her out!

Question 3: From Gradient Descent to Backpropagation

Explain in your own words how the concepts of gradient descent and backpropagation are related to each other.

Please put your memo for Question 1 and the answers to Questions 2 and 3 in a single text file and put it in your course directory. Also add your neural network code, named iris_nn.c.