PROGRAMMING ASSIGNMENT 5
Submission Part 1: DUE: Tuesday April 14, 11:59 PM

Submission Part 2: DUE: Tuesday April 21, 11:59 PM

Submission Part 3: DUE: Tuesday April 28, 11:59 PM

Using WEKA, build a neural network. Specifically a "Multilayer Perceptron". Use the "Letter Recognition" data set in the UCI repository. This assignment is worth 300 points. The grading of each part of this assignment will be focused on whether the information requested was provided fully and clearly, and whether the discussion paragraph in each section shows depth of understanding of the process being followed to build the neural network and of the results obtained.

NOTE: The 2 groups that find the neural network with the best accuracy on the test set (see PART 3 below) will receive 15 % bonus (45 points) on the entire assignment.

PART 1: Worth 30% of total assignment points

Use WEKA to divide the 20,000 records into two groups. One group, "DATA SET 1", should include **a random selection** of 85% of the records and it is to be used to train the network. The second group, "DATA SET 2", should include the other 15% of the records,

Your submission for Part 1 should be a pdf file with 3 sections.

Section 1 should include your name and your partner's name at the top. It should then include data in four columns, each column properly labeled: the first column should be letters A through Z. The second column should show the percentage of the 20,000 records that includes data for recognition of that letter. The third column should show the percentage of data set 1 records that includes data for recognition of that letter. The fourth column should show the percentage of data set 2 records that includes data for recognition of that letter.

Section 2 should include the results that WEKA provides when you use data set 1 to train the network selecting **"use training set" option in the "test options" window**. The following items are to be included and clearly described in this section 2:

a. The overall accuracy achieved by the trained network.

b. The accuracy by which each letter was classified. Use a format similar to what you used in section 1.
c. The number of layers in the network and the number of neurons in each layer
d. Save the network in case this is the best network you will find.
e. A paragraph discussing the results obtained including -- best accuracy and worst accuracy on a letter, and any other points that you want to include to show that you understand the process of building the network.

Section 3 should include the results that WEKA provides when data set 2 records are tested on the network described on section 2. The following items are to be included and clearly described in this section 3:

f. The overall accuracy achieved by the network on this test file
g. The accuracy by which each letter was classified. Use a format similar to that used in sections 1 and 2.
h. A paragraph discussing the results obtained including – do the best and worst accuracies correspond to the same letters from section 2 and any other points that you want to include to show that you understand the process of building the network.

PART 2: Worth 30% of total assignment points

Your submission for Part 2 should be a pdf file with 2 sections.

Section 1 should include your name and your partner's name at the top. Then it should include the results that WEKA provides when you use data set 1 to train the network selecting **"percentage split" option in the "test options" window** and setting **the percentage split to 80**. The following items are to be included and clearly described in this section 1:

a. The overall accuracy achieved by the trained network.
b. The accuracy by which each letter was classified. Use a format similar to what you used in Part 1.
c. The number of layers in the network and the number of neurons in each layer
d. Save the network in case this is the best network you will find.
e. A paragraph discussing the results obtained – compare results obtained with this method vs those obtained in PART 1 – for example "is the architecture similar?", "accuracy" "best and worse accuracies on letters" etc.

Section 2 should include the results that WEKA provides when data set 2 records are tested on the network described on section 1 and saved. The following results are to be included and clearly described in this section:

f. The overall accuracy achieved by the network on this test file

g. The accuracy by which each letter was classified. Use a format similar to what you used in sections 1.
h. A paragraph discussing the results obtained -- compare results obtained with this method vs those obtained in PART 1.

PART 3: Worth 40% of total assignment points

Your submission for Part 3 should be a zip file that includes;

1. a pdf with 2 sections,
2. the file for the model of the best network found and
3. the .arff file that includes data set 2

Section 1 should include your name and your partner's name at the top. Then it should include the results that WEKA provides when, using the **"percentage split" option in the "test options" window** and setting **the percentage split to 80**, you experiment trying to find the "best" network by varying the number of neurons in the hidden layers, varying the number of epochs, and increasing the number of hidden layers. You should at least try up to 3 hidden layers. For each of 3 different networks that you build searching for the "best", include the following items.

a. The number of layers in the network and the number of neurons in each layer
b. The overall accuracy achieved by the network on the training set (data set 1) and on the test set (data set 2).
c. The accuracy by which each letter was classified. Use a format similar to what you used in sections 1 and 2.
d. At the end of the section, include a paragraph discussing the results obtained. The discussion should show depth of understanding of the process of building neural networks.

Section 2 should include your selection of the best network that you found. The following should be included and clearly described in this section:

i. A description of the network, including number of layers and neurons in each layer, and the overall accuracy of the network on the testing set.
j. The accuracy by which each letter was classified by this best network. One column for the accuracy on the training set (data set 1) and one column for the accuracy on the test set (data set 2). Use a format similar to what you used in Part 1.
k. A paragraph with a clear justification for why you selected the network as being "best".
l. A paragraph discussing the strengths and weaknesses of the entire method followed, as well as strengths and weaknesses of WEKA as a tool.