Machine Learning CBC - Assignment 3: Artificial Neural Networks

Chris Carter (cc1808), Mark Law (ml1909), Michael Thorpe (mt2309) and Fraser Waters (fjw08)

Network parameters

The three network parameters that we modified were learning rate, number of hidden layers and number of nodes per hidden layer. We could have modified epochs but we set that at 100 and felt that was acceptable as the tests would normally terminate before 100 epochs had passed, as well having a negligible effect on performance.

Learning rate is a measure of how much the weights and biases are changed, and is used when applying the Gradient Descent training algorithm.

Hidden layers and nodes per hidden layer are simply a count of how many layers of nodes there are between the input and output layers and the number of nodes each of those layers has.

In order to try and optimise the networks, we ran several tests to try and find the configuration which maximises the correctness of the classifier. Graphs of our findings are shown below.

<u>Single 6-output neural network</u>

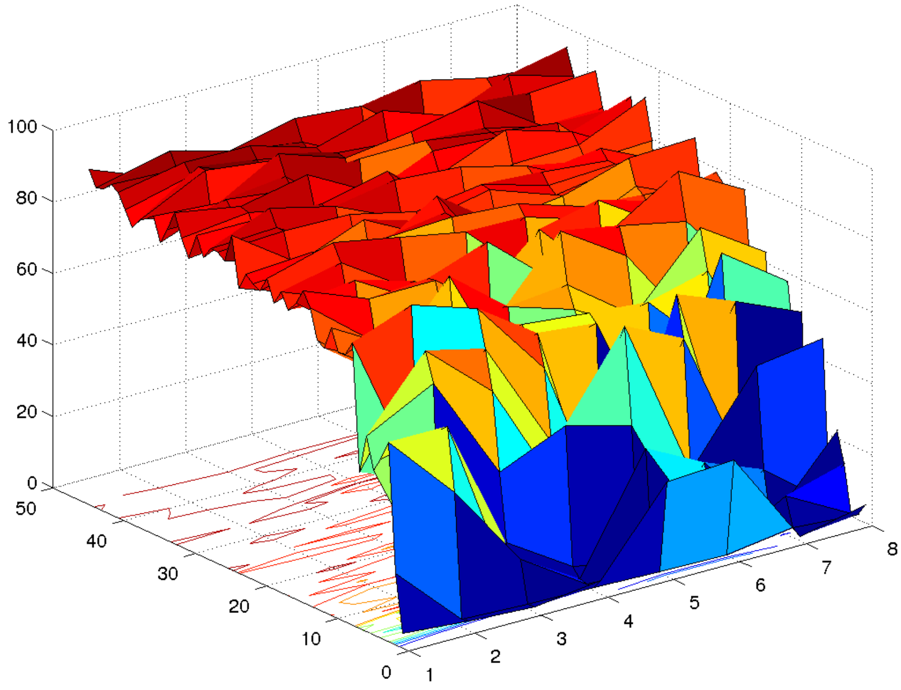


Figure 1: Number of hidden layers (x-axis) vs. number of nodes per hidden layer (y-axis) vs. percentage correct classifications (z-axis)

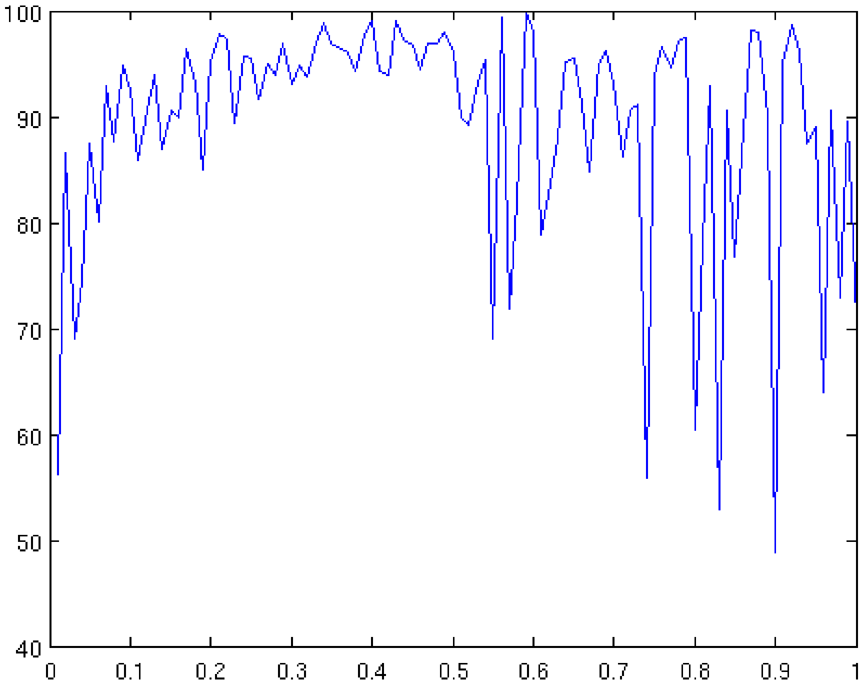


Figure 2: Learning rate parameter (x-axis) vs. percentage correct classifications (y-axis)

Six 1-output neural networks

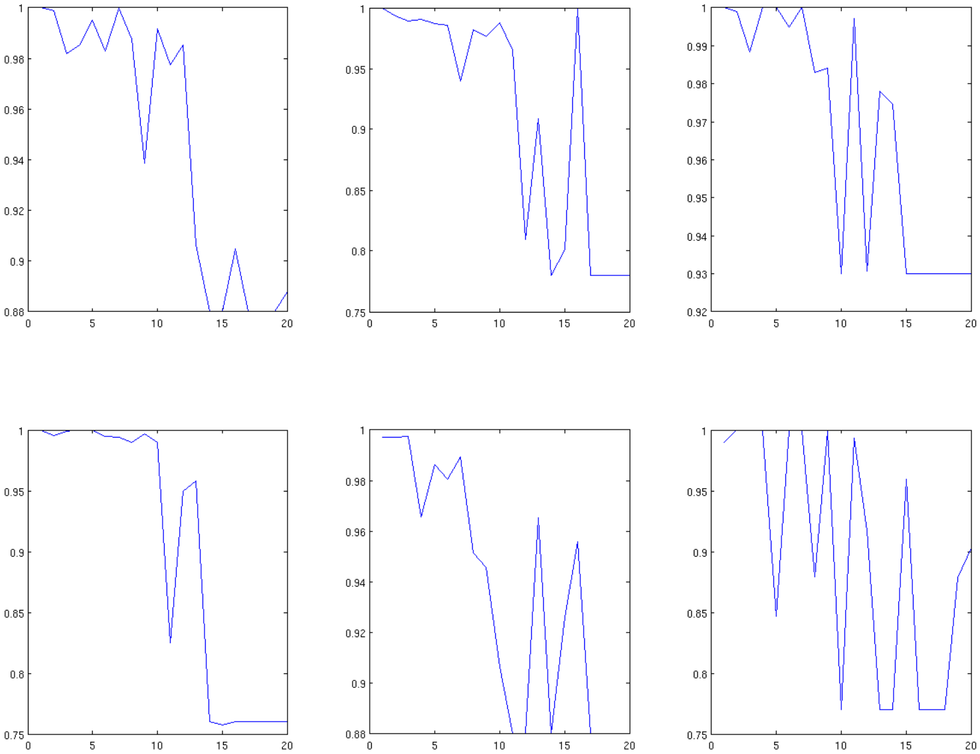


Figure 3: Number of hidden layers (x-axes) vs. fraction of correct classifications (y-axes), for each neural network

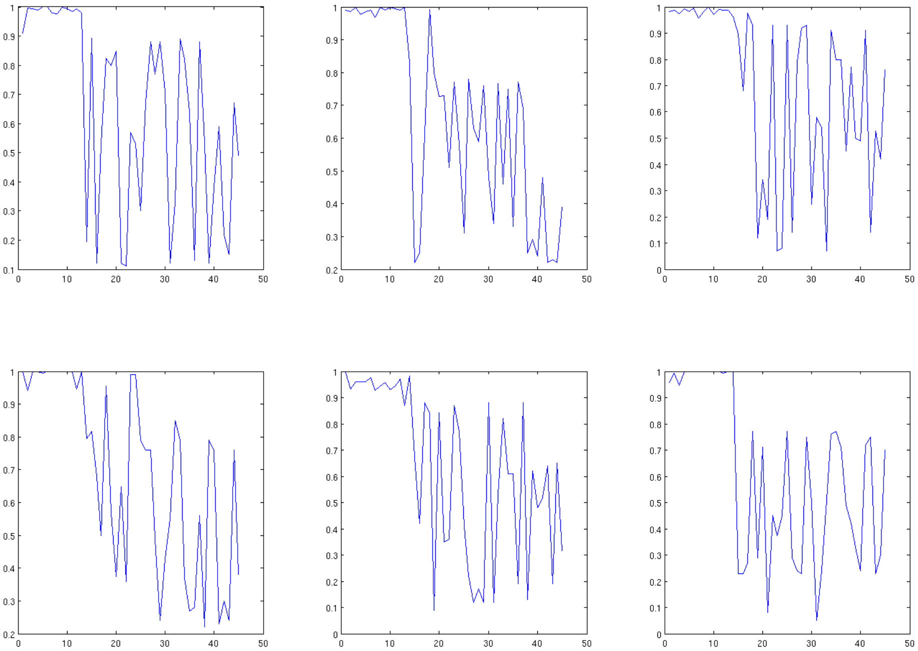


Figure 4: Number of nodes per hidden layer (x-axes) vs. fraction of correct classifications (y-axes), for each neural network

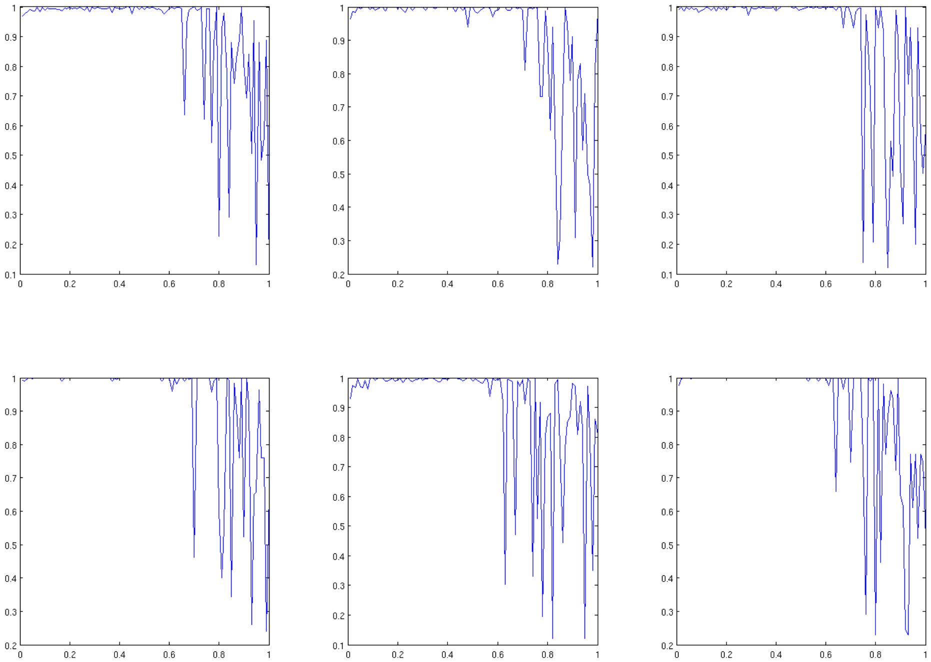
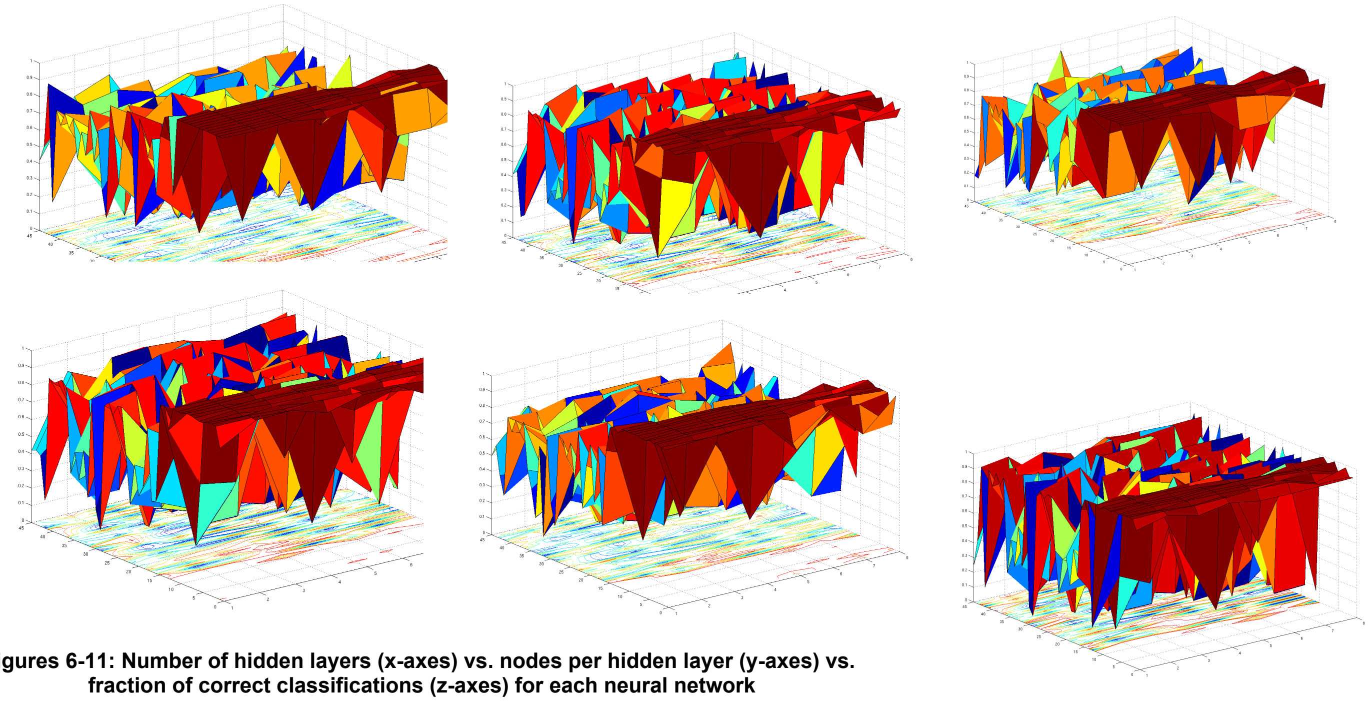


Figure 5: Learning rate parameter (x-axes) vs. fraction of correct classifications (y-axes), for each neural network



Figures 6-11: Number of hidden layers (x-axes) vs. nodes per hidden layer (y-axes) vs. fraction of correct classifications (z-axes) for each neural network

These six graphs show the variation in performance of the six single output networks against number of layers and number of nodes per layer. They represent simultaneously the data shown in figures 4 and 5. They all perform well for lower numbers of nodes per layer increasing slightly up until around ten nodes, after which it became erratic due to over fitting. For this reason we decided to use 7 nodes per layer. The number of layers did not seem to make much difference so we chose 2 layers.

In the graphs above, the percentage correctness is used to measure value is 'optimal' for the attribute we were testing. In order to calculate this, we compared the output of the neural network with the known classification result, and worked out the percentage of correct classifications. We felt that this was a suitable value to maximise to optimise the attributes, however one alternative would have been to maximise the recall and precision values, or the f1 measure.

Using the graphs, we have decided on the following optimal parameters:

	Six-output Neural Network	Six single-output Neural Networks
Number of hidden layers	2	2
Nodes per hidden layer	25	7
Learning rate	0.4	0.4

As you can see, the only difference between the two systems lies in the number of nodes in the hidden layers between the input and output layers. We feel that this may be due to the six-output system not suffering from over training until many more nodes are used. The single-output systems quickly become over trained with increasing number of nodes due to having to make a much simpler decision. This is supported by the data represented by the graphs shown above.

Using the optimal values we have found, we then proceeded to perform 10-fold cross-validation on both systems. The results of this validation is shown below:

Six-output Neural Network

	Anger (1)	Disgust (2)	Fear (3)	Happiness (4)	Sadness (5)	Surprise (6)
Anger (1)	8	0	1	0	1	0
Disgust (2)	1	21	0	0	2	0
Fear (3)	1	0	5	0	0	0
Happiness (4)	0	0	1	24	0	0
Sadness (5)	1	1	0	0	9	0
Surprise (6)	1	0	0	0	0	23

From this matrix we were then able to compute the average recall, precision values over the 10 folds, along with the F1 measure for each of the emotion values:

	Anger (1)	Disgust (2)	Fear (3)	Happiness (4)	Sadness (5)	Surprise (6)
Recall	0.4333	0.9750	0.4500	1.000	0.5500	0.9000
Precision	0.5333	0.8500	0.4500	0.9500	0.6333	0.8667
F_1 measure	0.472	0.9082	0.4500	0.9744	0.5887	0.8830

Six single-output Neural Networks

	Anger (1)	Disgust (2)	Fear (3)	Happiness (4)	Sadness (5)	Surprise (6)
Anger (1)	7	1	0	0	4	0
Disgust (2)	0	18	0	0	0	0
Fear (3)	2	0	6	0	0	0
Happiness (4)	1	1	0	24	0	0
Sadness (5)	0	1	1	0	7	0
Surprise (6)	2	1	0	0	1	23

From this matrix we were again able to compute the average recall, precision values over the 10 folds, along with the F1 measure for each of the emotion values for this system:

	Anger (1)	Disgust (2)	Fear (3)	Happiness (4)	Sadness (5)	Surprise (6)
Recall	0.3833	0.8750	0.5500	1.0000	0.4000	0.9000
Precision	0.3750	1.0000	0.5500	0.9167	0.5000	0.8167
F_1 measure	0.3791	0.9333	0.5500	0.9565	0.4444	0.8563

As you can see from the tables shown above, the single six-output neural network system seems to perform marginally better given our optimal parameters. It also appears that both systems struggle to identify anger and sadness (emotions 1 and 5 respectively), however both perform well with disgust, happiness, and surprise (emotions 2, 4, and 6 respectively).

Although a direct comparison between the f1 measures for each system for each fold would show how both systems perform with equal data sets, we feel this would be inconsequential for this data. This is because each fold only includes 10 examples, and so each fold may not contain equal numbers of each emotion. With such a tiny data set, the percentage error is vastly increased for one misclassification, and a graph representing this comparison would be erratic and resemble spaghetti. We feel this would not be a fair comparison between the systems, and would require a much larger data set in order to make a suitable analysis.

Advantages/disadvantages

The advantage of using a six output network is that exactly one emotion can be selected each time, this cannot be said for the six single output networks as multiple classifications or even no classifications are possible.

The disadvantage is that when trying to assess whether a particular classification holds given a set of attributes the six output network is less accurate than the six single output networks.

Overall which network is preferable is mainly down to the what it is being used for. If given a set of attributes the most likely emotion is required then the six output network is best. However if it is necessary to determine whether a particular emotion holds then the six single output networks are best.

Deliverables:

2. Report of approximately two pages (excluding result matrices) containing:
(a) Discussion of the network parameters (e.g. learning rate, epochs, topology, etc.). What criteria have you used to choose optimal topology/parameters of the networks? Compare the optimal parameters of both types of the networks (the single six-output network and six single-outputs networks). Explain what strategy you employed to ensure good generalisation ability of the networks and overcome the problem of overfitting, if encountered (support this by experimental results).
(b) Perform 10-fold cross-validation for both types of networks. Cross-validation should be performed in the same way as in Assignment 2 (with a script that splits the given dataset into training and test sets). 10-fold cross-validation should be performed using the optimal topology and best parameters obtained in 2(a), i.e., for six-output and single-output NNs. Note that in the case of 6 networks, each example must be classified as one of the 6 emotions. Plot the performance (only F1 measure) per fold of each network type in the same figure.
(c) Report also the average results of the 10-fold cross-validation (similarly to the trees) for both types of networks (single six-output NN and six single-outputs NNs):
o confusion matrices,
o recall and precision rates per class,
(Hint: you can derive them directly from the previously computed confusion matrix)
o the F1-measure derived from the recall and precision rates of the previous step.
(d) Is there any difference in the classification performance of the two different classification approaches. Discuss the advantages / disadvantages of using 6 single-outut NNs vs. 1 six-output NNs.