Assignment 2 – Finding Natural Feature Sets

1. Introduction

In this paper, I will discuss the implementation of a simple neural network that is used for classification. I will be focusing on improving the performance of the classification model and to learn how the hidden nodes activate for different classes (groupings) of letters.

My hypothesis for this experiment is that different classes of letters with similar line structures (horizonal lines such as "E" and "F", diagonal lines such as "M" and "N") will activate the same hidden nodes to the same intensity.

2. Model & Experiment Design

A simple neural network (input-hidden-output) architecture utilizing the backpropagation algorithm is implemented for this experiment. The neural network contains 81 input nodes, corresponding to having a 9x9 element input grid, and a varying number of output nodes, depending on how many classes are specified. There are 26 potential classes, which corresponds to each letter in the alphabet, however we are only interested in classes that contains groups of similar line structures. These classes are determined by how similar the letters are in terms of line/angle construction (i.e. a class being composed of "M" and "N"). The input data is structured to be a 9x9 grid, consisting of 0 (white space/pixel) and 1 (the angle of the letter in the pixel). In addition, a noise parameter = 0.25 is implemented into training data to make the data appear "more realistic" as in the real world, we cannot expect to have perfect, well defined data to work with.

There will be a baseline model (81 input nodes, 7 output nodes (feature classes), 9 hidden nodes, learning rate $\eta = 0.5$, limited to a maximum of 5000 iterations). From the results of this baseline model, I will experiment with changing the number of hidden nodes and the different classes to see how the hidden nodes activate and how the training and test accuracy will improve. Note that the focus of this experiment is to not optimize the model, but to see how the different hyperparameters affect the model.

3. Model I – Baseline Model

Like mentioned earlier, the baseline model will be used as the starting point to understand the changes occurring to the model. We began with 9 hidden nodes to see whether or not we should increase/decrease the number of nodes to achieve the "sweet spot", which would show hidden node activations clearly for different feature classes. The following table below shows the 8 different classes defined in the training data:

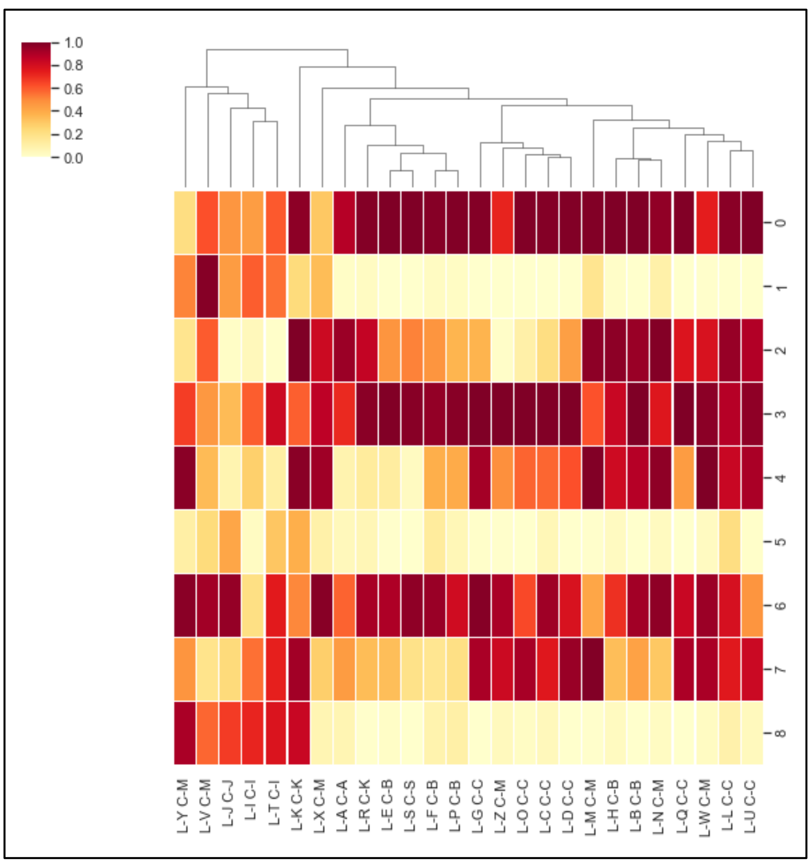
Class	Letters
0	A
1	B, E, F, H, P
2	C, D, G, L, O, Q, U
3	I, T
4	J
5	R, K
6	M, N, V, W, X, Y, Z
7	S

Here is the initial classification results table, which is performance prior to introducing the training data:

```
Desired/Selected Output Classes
=====================================================================================
=====================================================================================

         0  1  2  3  4  5  6  7  8  9  10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
Letter    A  B  C  D  E  F  G  H  I  J  K  L  M  N  O  P  Q  R  S  T  U  V  W  X  Y  Z
Desired   0  1  2  2  1  1  2  1  3  4  5  2  6  6  2  1  2  5  7  3  2  6  6  6  6  6
Selected  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2
Correct   N  N  Y  Y  N  N  Y  N  N  N  N  Y  N  N  Y  N  Y  N  N  N  Y  N  N  N  N  N
Percent Accuracy 26.92
=====================================================================================
=====================================================================================
```

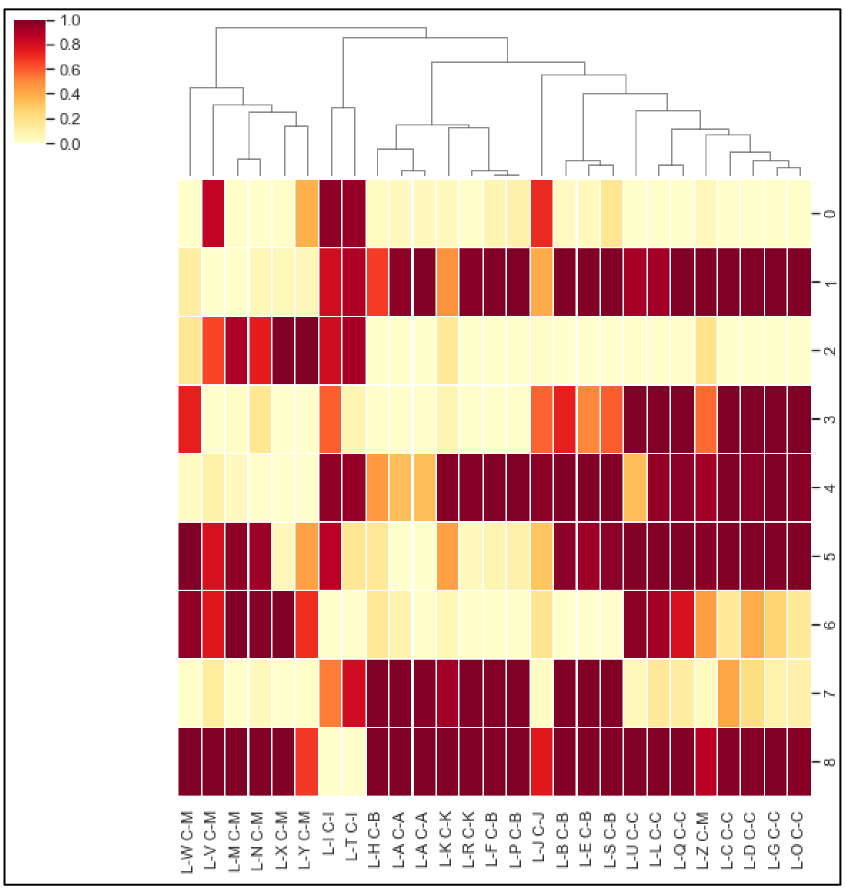
The classification accuracy is low at 34.62% and from the output table, we can also see which letters are not being classified into the desired class correctly. In fact, the model selecting feature class 2 had the highest frequency. The dendrogram below also shows performance before training – not much insight can be derived about the patterns by which the nodes are activating for the different classes as several nodes are being activated to the same extent for all class sets:



Next, the training data will be introduced with noise level of 0.25 (on a scale of 0 to 1) to mimic the noisy data that may be encountered in a real business setting. We can see that by training the data, performance has improved a lot to an accuracy of 59.26%. Based on the results table, the letters A, B, E, J, K, N, R, S, W, and Z were classified into the wrong classes. With the increased amount of accuracy, we can now derive a bit of insight as to how the user-defined classes are being picked up by the hidden node.

```
Desired/Selected Output Classes
===============================================================================
===============================================================================

          0  1  2  3  4  5  6  7  8  9  10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
Letter    A  B  C  D  E  F  G  H  I  J  K  L  M  N  O  P  Q  R  S  T  U  V  W  X  Y  Z  A
Desired   0  1  2  2  1  1  2  1  3  4  5  2  6  6  2  1  2  5  7  3  2  6  6  6  6  6  0
Selected  1  2  2  2  2  1  2  1  3  3  2  2  6  2  2  1  2  1  2  3  2  6  2  6  6  2  1
Correct   N  N  Y  Y  N  Y  Y  Y  Y  N  N  Y  Y  N  Y  Y  Y  N  N  Y  Y  Y  N  Y  Y  N  N
Percent Accuracy 59.26
===============================================================================
===============================================================================
```



Based on the dendrogram, we can identify some of the unique classes being picked up by the hidden node activations. For example, Class C is being uniquely identified as heavily activating nodes 1, 3, 4, 5, and 8, while Class M is heavily activating nodes 2, 5, 6, and 8. In addition, there are nodes that are still activating for almost all class sets, so there is a possibility of reducing the amount of hidden nodes to achieve better insight on the natural feature sets/classes.

From here, we will redefine some of the letters to match the classes better to the naturally defined classes specified by the hidden node activation patterns.

4. Model II – 7 Hidden Nodes

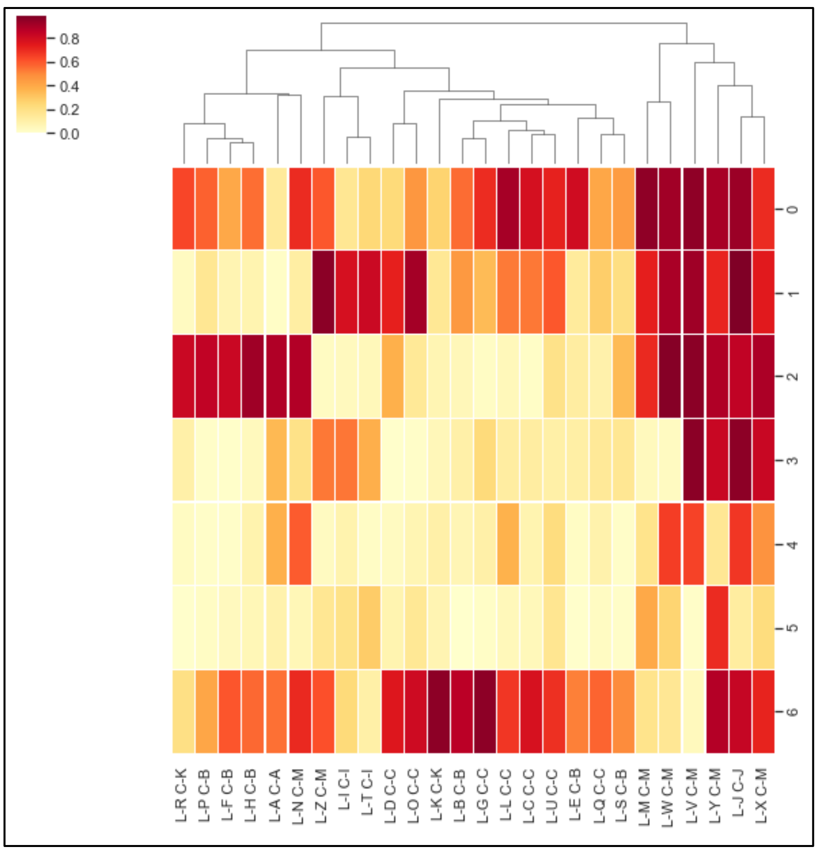
Model II is built based off of Model I and it was fine-tuned to account some changes made to the classes based of Model I's hidden node activation patterns. The changes to the classes include moving "S" to class 1, thus reducing the number of output classes. In addition, the number of hidden nodes was reduced to 7 in attempts of better finding the "sweet spot" of hidden nodes that would give the most meaningful results. The following table below shows the 7 different classes defined in the training data:

Class	Letters
0	A
1	B, E, F, H, P, S
2	C, D, G, L, O, Q, U
3	I, T
4	J
5	R, K
6	M, N, V, W, X, Y, Z

Surprisingly the performance accuracy before introducing the training data was even lower compared to Model I at 19.23%. This time, the selected classes for the model were overwhelmingly class 1 or 6, with the dendrogram not giving much insight.

```
Desired/Selected Output Classes
=====================================================================================
=====================================================================================

         0  1  2  3  4  5  6  7  8  9  10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
Letter    A  B  C  D  E  F  G  H  I  J  K  L  M  N  O  P  Q  R  S  T  U  V  W  X  Y  Z
Desired   0  1  2  2  1  1  2  1  3  4  5  2  6  6  2  1  2  5  1  3  2  6  6  6  6  6
Selected  1  1  1  1  6  1  6  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
Correct   N  Y  N  N  N  Y  N  Y  N  N  N  N  N  N  N  Y  N  N  Y  N  N  N  N  N  N  N
Percent Accuracy 19.23
=====================================================================================
=====================================================================================
```

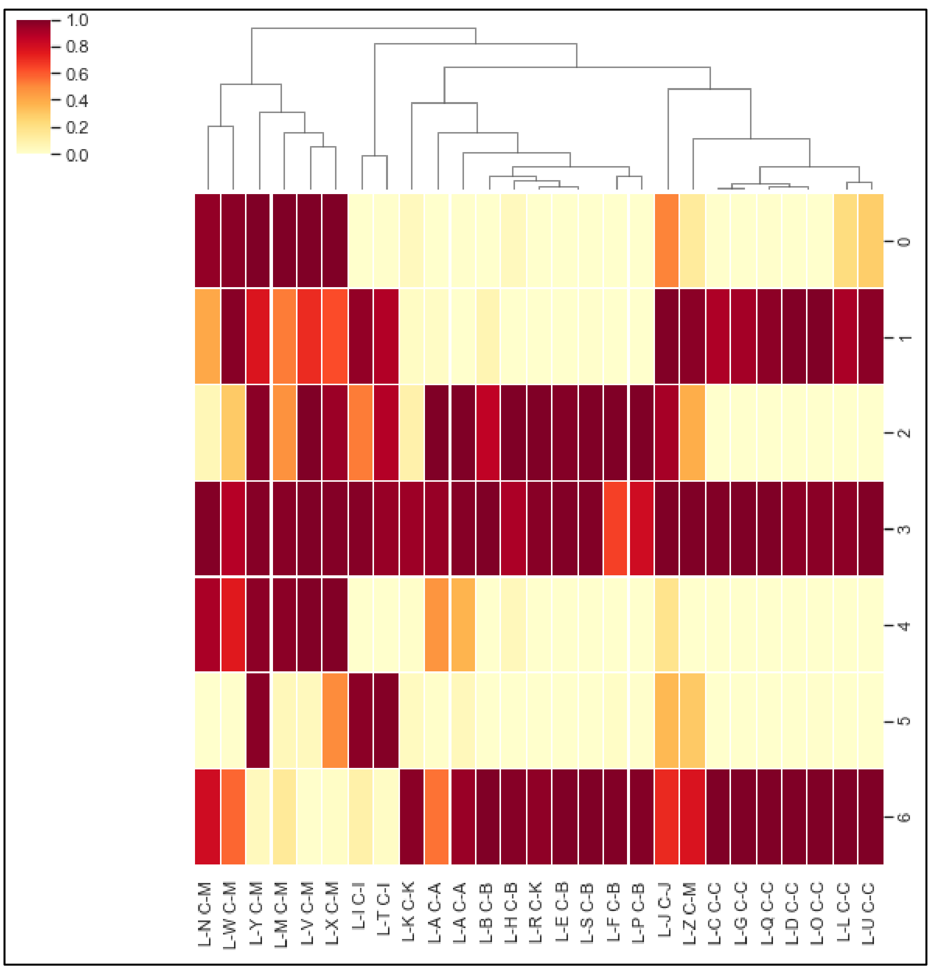


Once the training data was introduced, it looks like the changes made to the feature sets/classes yielded better performance accuracy at 77.78%. The letters A, J, K, R, and Z were classified into the wrong classes, which will be considered when looking at the next iteration of model improvements for Model III.

```
Desired/Selected Output Classes
=====================================================================================
=====================================================================================

         0  1  2  3  4  5  6  7  8  9  10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
Letter    A  B  C  D  E  F  G  H  I  J  K  L  M  N  O  P  Q  R  S  T  U  V  W  X  Y  Z  A
Desired   0  1  2  2  1  1  2  1  3  4  5  2  6  6  2  1  2  5  1  3  2  6  6  6  6  6  0
Selected  1  1  2  2  1  1  2  1  3  6  2  2  6  6  2  1  2  1  1  3  2  6  6  6  6  2  1
Correct   N  Y  Y  Y  Y  Y  Y  Y  Y  N  N  Y  Y  Y  Y  Y  Y  N  Y  Y  Y  Y  Y  Y  Y  N  N
Percent Accuracy 77.78
=====================================================================================
=====================================================================================
```

The dendrogram also shows better insight into how the patterns naturally being picked up by what hidden nodes are being activated:



For instance, Class C is clearly being defined as activating nodes 1, 3, and 6, while Class B activating nodes 2, 3, and 6, Class I activating nodes 1, 2, 3, and 5, and finally M activating 0, 1, 3, and 4. Letter Z in Class M appears to be aligning with Class C in terms of node activation, so a shift to Class C will be made. Letter R seems to be activating similarly to Class B, so another adjustment will be made there for Model III.

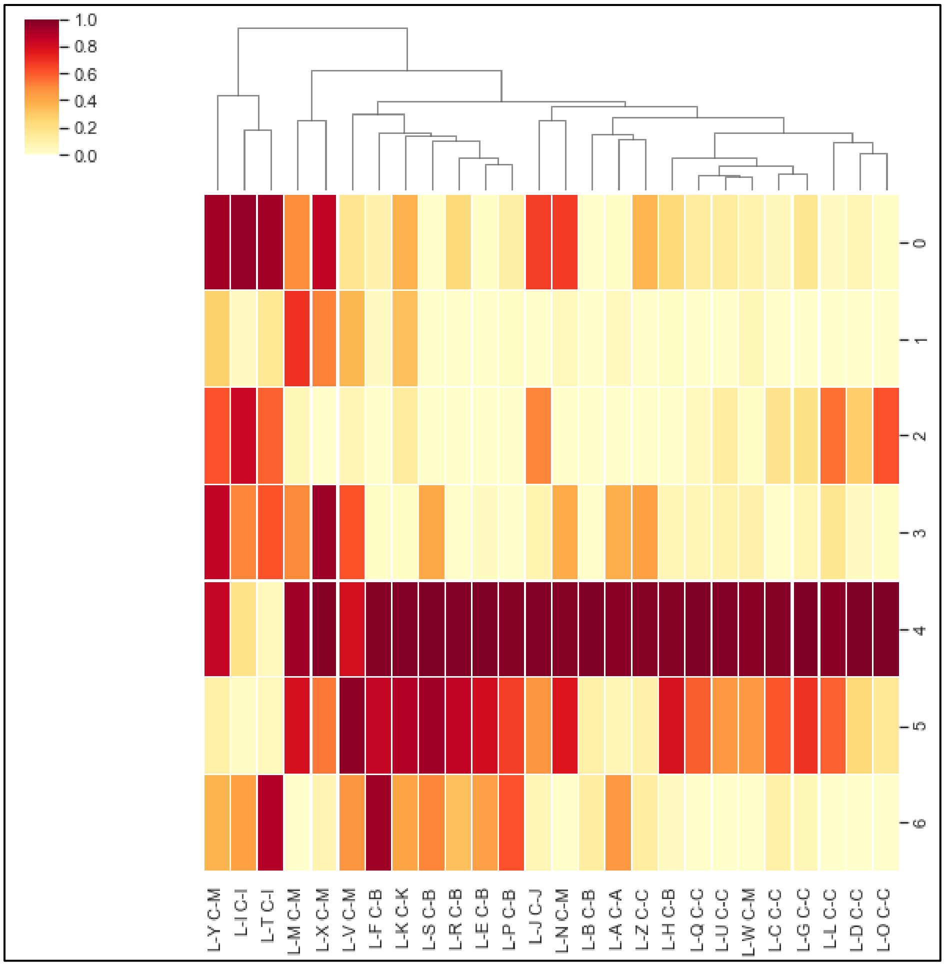
4. Model III – The Final Model

Taking in the insights gathered from Model II, a final redefinition based on the natural feature sets being picked up by hidden node activation patterns will be made to hopefully yield higher performance. This final model contains 81 input nodes, 7 hidden nodes, and 7 different output classes (feature sets). Like the other models, a noise parameter of 0.25 is introduced for the training data.

The initial results before introducing the training data shows a performance accuracy of 23.08%, which is slightly higher than Model II, all classes were predicted as Class 6, and its dendrogram is not providing much insight besides the fact that node 4 is being activated for nearly all classes, except for Class I.

```
Desired/Selected Output Classes
==========================================================================================
==========================================================================================

         0  1  2  3  4  5  6  7  8  9  10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
Letter    A  B  C  D  E  F  G  H  I  J  K  L  M  N  O  P  Q  R  S  T  U  V  W  X  Y  Z
Desired   0  1  2  2  1  1  2  1  3  4  5  2  6  6  2  1  2  1  1  3  2  6  6  6  6  2
Selected  6  6  6  6  6  6  6  6  6  6  6  6  6  6  6  6  6  6  6  6  6  6  6  6  6  6
Correct   N  N  N  N  N  N  N  N  N  N  N  N  Y  Y  N  N  N  N  N  N  N  Y  Y  Y  Y  N
Percent Accuracy 23.08
==========================================================================================
==========================================================================================
```



Introducing the training data yielded much more favorable results. The performance accuracy was 85.19%, which is the highest accuracy compared to baseline Model I and Model II. The letters A, J, and K were not classified to the right class correctly.

```
Desired/Selected Output Classes
==========================================================================================
==========================================================================================

          0  1  2  3  4  5  6  7  8  9  10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
Letter     A  B  C  D  E  F  G  H  I  J  K  L  M  N  O  P  Q  R  S  T  U  V  W  X  Y  Z  A
Desired    0  1  2  2  1  1  2  1  3  4  5  2  6  6  2  1  2  1  1  3  2  6  6  6  6  2  0
Selected   1  1  2  2  1  1  2  1  3  2  2  2  6  6  2  1  2  1  1  3  2  6  6  6  6  2  1
Correct    N  Y  Y  Y  Y  Y  Y  Y  Y  N  N  Y  Y  Y  Y  Y  Y  Y  Y  Y  Y  Y  Y  Y  Y  Y  N
Percent Accuracy 85.19
==========================================================================================
==========================================================================================
```

The final model's dendrogram shows that the user-defined classes align with the naturally defined patterns being picked up with the hidden node activations. We can see that Class M is defined by the activations of nodes 0, 1, 3, and 5. Class I is defined by the activations of nodes 0, 2, and 6. Class J is defined by nodes 0 and 2. Class C is defined by the activation of 0, 2, 4, and 5. Class K is defined by 0, 1, 2, 4, 5, and 6. Similarly to Class K, Class B is defined by nodes 0, 4, 5, and 6. Lastly, Class A is defined by nodes 0, 1, 4, 5, and 6.

5. Conclusion

The results from the set of experiments proved that my hypothesis was correct - different classes of letters with similar line structures will activate the same hidden nodes to the same intensity. With Model III, the dendrogram showed that each class has a unique set of hidden node activations. This experiment proved to be a good exercise in training data that contains some noisiness while yielding a decent performance accuracy, which businesses should learn to not be intimidated with data that has some noisiness to it - key insights about classes or clusters can still be learnt.