MACHINE LEARNING

ASSIGNMENT 3: NEURAL NETWORKS

SARTHAK CHAKRABORTY (16CS30044)

Part 1

The objective of this assignment is to build a spam classifier. The dataset contains 5574 messages with each labelled as HAM or SPAM. A neural network was built to classify the messages.

The steps involved in build the networks included:

1. **PREPROCESS:** The given data was first extracted from the '.txt' format and two separate arrays containing the message and their class was formed. Each of the string was tokenized first, followed by the removal of stopwords and Porter stemming. Porter stemming was done using 'nltk' library of python. The data was then split into two sets - train(80%) and test(20%)

2. **DATA LOADER:** Mini batches were created for the train set and loaded in this module.

3. **TRAINING:** The specifications for the neural networks used are:
   a) Number of hidden layers = 1
   b) Number of neurons in the hidden layer = 100
   c) Activation function = **ReLU**(for hidden layer) and **Sigmoid**(for output layer)
   d) Number of neurons in input layer = Determined in program
   e) Number of neurons in the output layer = 1
   f) Threshold for classification = 0.5
   g) Loss function = Categorical Cross-Entropy Loss
   h) Number of epochs = 30
   i) Optimizer = Mini-batch Stochastic Gradient Descent (SGD)
   j) Batch size = 32
   k) Learning Rate = 0.1

4. **TESTING:** The built model was then tested using the test set and the accuracy was reported.

Results:

The model was trained for 30 epochs using mini-batch SGD. A forward pass and a backward pass is made for every batch in each epoch to train the weights and the biases. Accuracy and Loss are reported for each epoch. The result is as follows:

```
EPOCH  1        Train Error:  0.1309683177153057      Loss:  0.38943150415076333     Test Error:  0.14652014652014655
EPOCH  2        Train Error:  0.1309683177153057      Loss:  0.37523708245264187     Test Error:  0.14652014652014655
EPOCH  3        Train Error:  0.1309683177153057      Loss:  0.322942428180728       Test Error:  0.14652014652014655
EPOCH  4        Train Error:  0.0535475234270415      Loss:  0.17315714477987423     Test Error:  0.08150183150183155
EPOCH  5        Train Error:  0.02677376171352075     Loss:  0.09851558476414375     Test Error:  0.045787545787545736
EPOCH  6        Train Error:  0.018964747880410582    Loss:  0.07278328062823627     Test Error:  0.03663003663003661
EPOCH  7        Train Error:  0.015171798304328465    Loss:  0.05884780256449335     Test Error:  0.03388278388278387
EPOCH  8        Train Error:  0.013163766175814384    Loss:  0.04938709646466962     Test Error:  0.03205128205128205
EPOCH  9        Train Error:  0.011601963409192284    Loss:  0.042257051652660076    Test Error:  0.0283882783882784
EPOCH  10       Train Error:  0.009147701918786222    Loss:  0.03659496784131994     Test Error:  0.02472527472527475
EPOCH  11       Train Error:  0.00825524319500226     Loss:  0.03196572587846102     Test Error:  0.022893772893772923
EPOCH  12       Train Error:  0.006470325747434225    Loss:  0.028114732700161644    Test Error:  0.0210622710622711
EPOCH  13       Train Error:  0.0055778670236501515   Loss:  0.024879588830617787    Test Error:  0.020146520146520186
EPOCH  14       Train Error:  0.0049085229808121245   Loss:  0.02214967500604795     Test Error:  0.020146520146520186
EPOCH  15       Train Error:  0.004239178937974097    Loss:  0.01983856298745918     Test Error:  0.020146520146520186
EPOCH  16       Train Error:  0.004016064257028162    Loss:  0.0178713519232854      Test Error:  0.019230769230769273
EPOCH  17       Train Error:  0.0029004908522980433   Loss:  0.016185505383550004    Test Error:  0.019230769230769273
EPOCH  18       Train Error:  0.0029004908522980433   Loss:  0.01473114437621351     Test Error:  0.019230769230769273
EPOCH  19       Train Error:  0.0024542614904060622   Loss:  0.01346839246150839     Test Error:  0.019230769230769273
EPOCH  20       Train Error:  0.0017849174475680352   Loss:  0.012364745141859252    Test Error:  0.019230769230769273
EPOCH  21       Train Error:  0.0013386880856760541   Loss:  0.011393895262807268    Test Error:  0.019230769230769273
EPOCH  22       Train Error:  0.0013386880856760541   Loss:  0.010534462484749572    Test Error:  0.019230769230769273
EPOCH  23       Train Error:  0.001115573404730008    Loss:  0.00976966534845648     Test Error:  0.019230769230769273
EPOCH  24       Train Error:  0.001115573404730008    Loss:  0.009086145019308592    Test Error:  0.016483516483516536
EPOCH  25       Train Error:  0.001115573404730008    Loss:  0.008473293751172       Test Error:  0.016483516483516536
EPOCH  26       Train Error:  0.001115573404730008    Loss:  0.007922329882273404    Test Error:  0.015567765567765512
EPOCH  27       Train Error:  0.001115573404730008    Loss:  0.007425712984866593    Test Error:  0.015567765567765512
EPOCH  28       Train Error:  0.001115573404730008    Loss:  0.006976828594785586    Test Error:  0.015567765567765512
EPOCH  29       Train Error:  0.001115573404730008    Loss:  0.006569780816841108    Test Error:  0.015567765567765512
EPOCH  30       Train Error:  0.000892458723784073l   Loss:  0.0061993655662553335   Test Error:  0.015567765567765512

Train Set accuracy:  0.9991075412762159
Test Set Accuracy:  0.9844322344322345
```

[**Note: 'result.txt' stores the above result. The code also has a commented section where the data is being loaded from .npy files. It was done so that preprocessing can be avoided for every execution. Readme provides details of how to store the data in the .npy format.**]

Plots showing the variation of train error and test error with respect to epochs is as follows:

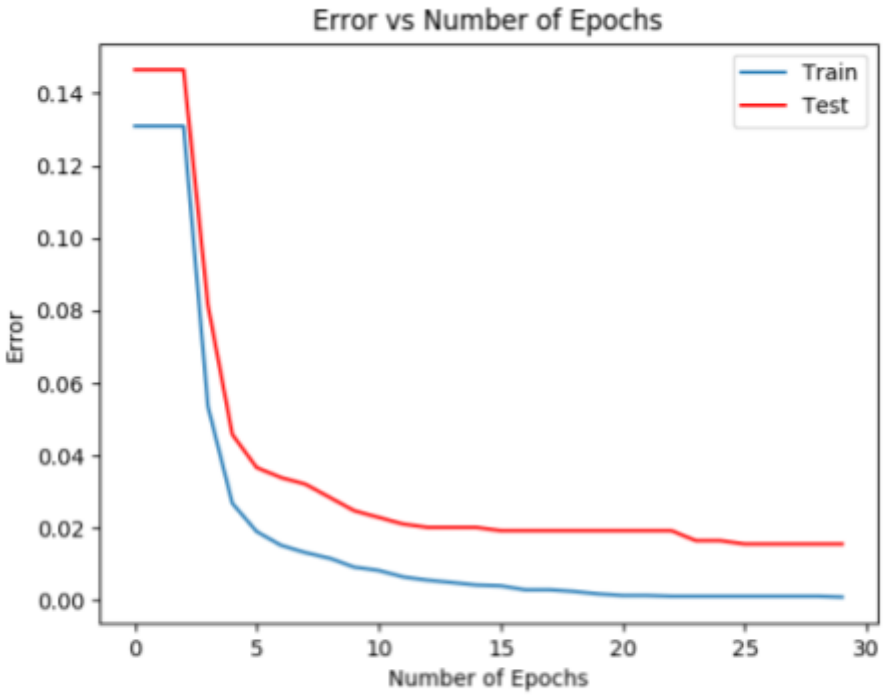

Fig. 1: Plot showing train error and test error(misclassification percentage) vs Epochs progressed.

Here is the plot showing the variation of the Cross-Entropy Loss function vs Number of epochs:

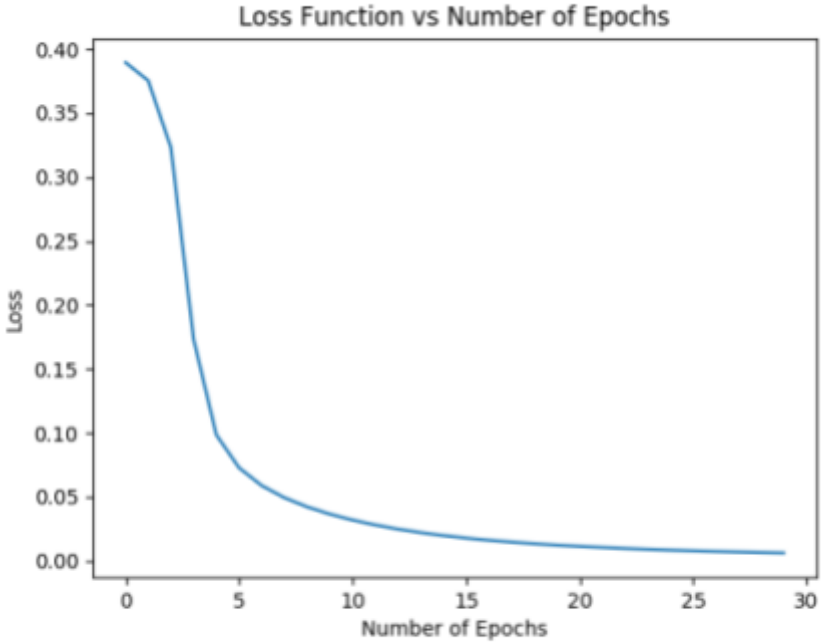

Fig. 2: Variation of Loss function vs Epochs progressed.

Final Test Accuracy: 0.9844322344322345

Part 2

The objective of this assignment is the same as the previous part, that is to build a spam classifier. The dataset contains 5574 messages with each labelled as HAM or SPAM. A neural network was built to classify the messages.

The steps involved in build the networks included:

5. **PREPROCESS:** The given data was first extracted from the '.txt' format and two separate arrays containing the message and their class was formed. Each of the string was tokenized first, followed by the removal of stopwords and Porter stemming. Porter stemming was done using 'nltk' library of python. The data was then split into two sets - train(80%) and test(20%)

6. **DATA LOADER:** Mini batches were created for the train set and loaded in this module.

7. **TRAINING:** The specifications for the neural networks used are:
   a) Number of hidden layers = 2
   b) Number of neurons in the hidden layer = <COMMAND LINE INPUT>
   c) Activation function = **Sigmoid**(for hidden layer) and **Softmax**(for output layer)
   d) Number of neurons in input layer = Determined in program
   e) Number of neurons in the output layer = 2
   f) Loss function = Categorical Cross-Entropy Loss
   g) Number of epochs = 30
   h) Optimizer = Mini-batch Stochastic Gradient Descent (SGD)
   i) Batch size = 8
   j) Learning Rate = 0.1

8. **TESTING:** The built model was then tested using the test set and the accuracy was reported.

Results:

In each epoch, a forward pass and a backward pass is made for batches successively to recompute the weights and biases as was explained in Part 1. The number of nodes in each hidden layer that has been used in the following demonstration is **[50, 10]**. Accuracy and Loss after each epoch is recorded. The result is as follows:

```
EPOCH  1        Train Error:  0.13262298798458394      Loss:  0.1331885253948541      Test Error:  0.1392949269131556
EPOCH  2        Train Error:  0.13262298798458394      Loss:  0.13434914172227416     Test Error:  0.1392949269131556
EPOCH  3        Train Error:  0.13262298798458394      Loss:  0.1355316448357774      Test Error:  0.1392949269131556
EPOCH  4        Train Error:  0.13262298798458394      Loss:  0.1367724686420357      Test Error:  0.1392949269131556
EPOCH  5        Train Error:  0.13262298798458394      Loss:  0.13806894665929575     Test Error:  0.1392949269131556
EPOCH  6        Train Error:  0.13262298798458394      Loss:  0.13933939565520376     Test Error:  0.1392949269131556
EPOCH  7        Train Error:  0.13262298798458394      Loss:  0.14037012076237124     Test Error:  0.1392949269131556
EPOCH  8        Train Error:  0.13262298798458394      Loss:  0.14076265203908653     Test Error:  0.1392949269131556
EPOCH  9        Train Error:  0.13262298798458394      Loss:  0.1398430341086073      Test Error:  0.1392949269131556
EPOCH  10       Train Error:  0.13262298798458394      Loss:  0.13638226211320284     Test Error:  0.1392949269131556
EPOCH  11       Train Error:  0.13262298798458394      Loss:  0.12847483220750439     Test Error:  0.1392949269131556
EPOCH  12       Train Error:  0.13262298798458394      Loss:  0.11975264047993506     Test Error:  0.1392949269131556
EPOCH  13       Train Error:  0.049421899795964674     Loss:  0.15607718895678868     Test Error:  0.061049011177988
EPOCH  14       Train Error:  0.02493765586034913      Loss:  0.23893717676503748     Test Error:  0.03009458297506451
EPOCH  15       Train Error:  0.017683065064611148     Loss:  0.3146526739252537      Test Error:  0.02407566638005154
EPOCH  16       Train Error:  0.011788710043074135     Loss:  0.37795597985977253     Test Error:  0.022355975924333582
EPOCH  17       Train Error:  0.009066823849467242     Loss:  0.4318593186102955      Test Error:  0.019776440240756643
EPOCH  18       Train Error:  0.005894355021537012     Loss:  0.4780485819565286      Test Error:  0.017196904557179704
EPOCH  19       Train Error:  0.004987531172069848     Loss:  0.5173776211214931      Test Error:  0.016337059329320724
EPOCH  20       Train Error:  0.0040807073226025725    Loss:  0.550770329964227       Test Error:  0.015477214101461745
EPOCH  21       Train Error:  0.002720471548401715     Loss:  0.5794333156642479      Test Error:  0.014617368873602765
EPOCH  22       Train Error:  0.0018136476989344397    Loss:  0.604537660144749       Test Error:  0.014617368873602765
EPOCH  23       Train Error:  0.001133529811834011     Loss:  0.6268499816646742      Test Error:  0.013757523645743786
EPOCH  24       Train Error:  0.001133529811834011     Loss:  0.6467747964697168      Test Error:  0.013757523645743786
EPOCH  25       Train Error:  0.0006801178871004288    Loss:  0.6646036449161283      Test Error:  0.012897678417884806
EPOCH  26       Train Error:  0.0006801178871004288    Loss:  0.6805444021513721      Test Error:  0.012897678417884806
EPOCH  27       Train Error:  0.00045341192473358216   Loss:  0.6948286292970427      Test Error:  0.012897678417884806
EPOCH  28       Train Error:  0.00045341192473358216   Loss:  0.7077635173406811      Test Error:  0.012897678417884806
EPOCH  29       Train Error:  0.00045341192473358216   Loss:  0.7196432959402179      Test Error:  0.012897678417884806
EPOCH  30       Train Error:  0.00045341192473358216   Loss:  0.7307041939278542      Test Error:  0.012897678417884806

Train Set accuracy:  0.9995465880752664
Test Set Accuracy:  0.9871023215821152
```

[**Note: 'result.txt' stores the above result. The code also has a commented section where the data is being loaded from .npy files. It was done so that preprocessing can be avoided for every execution. Readme provides details of how to store the data in the .npy format.**]

Plots showing the variation of train error and test error with respect to epochs is as follows:

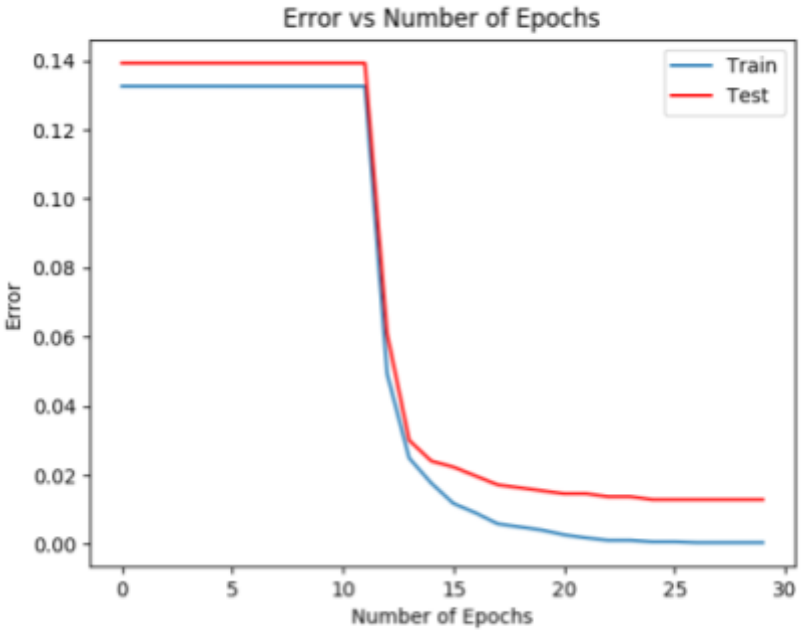

Fig. 3: Plot showing train error and test error(misclassification percentage) vs Epochs progressed.

Here is the plot showing the variation of the Cross-Entropy Loss function vs Number of epochs:

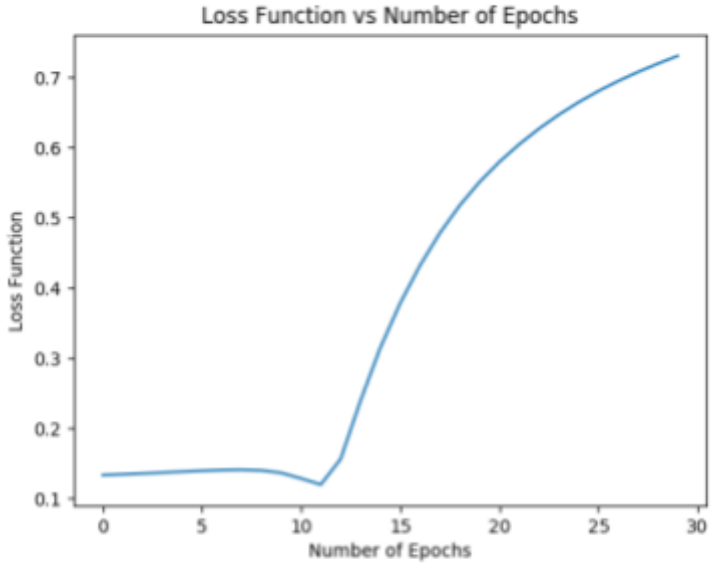

Fig. 4: Variation of Loss function vs Epochs progressed.

Final Test Accuracy: 0.9871023215821152
