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 = 512
   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:  0.593089259311   Test Error:  0.14652014652014655
EPOCH  2        Train Error:  0.1309683177153057        Loss:  0.529108679844   Test Error:  0.14652014652014655
EPOCH  3        Train Error:  0.1309683177153057        Loss:  0.487310345314   Test Error:  0.14652014652014655
EPOCH  4        Train Error:  0.1309683177153057        Loss:  0.45930755833    Test Error:  0.14652014652014655
EPOCH  5        Train Error:  0.1309683177153057        Loss:  0.44007935305    Test Error:  0.14652014652014655
EPOCH  6        Train Error:  0.1309683177153057        Loss:  0.426575948285   Test Error:  0.14652014652014655
EPOCH  7        Train Error:  0.1309683177153057        Loss:  0.416900519394   Test Error:  0.14652014652014655
EPOCH  8        Train Error:  0.1309683177153057        Loss:  0.409842608009   Test Error:  0.14652014652014655
EPOCH  9        Train Error:  0.1309683177153057        Loss:  0.404610076305   Test Error:  0.14652014652014655
EPOCH  10       Train Error:  0.1309683177153057        Loss:  0.400671592859   Test Error:  0.14652014652014655
EPOCH  11       Train Error:  0.1309683177153057        Loss:  0.397663639971   Test Error:  0.14652014652014655
EPOCH  12       Train Error:  0.1309683177153057        Loss:  0.395332221777   Test Error:  0.14652014652014655
EPOCH  13       Train Error:  0.1309683177153057        Loss:  0.393496094882   Test Error:  0.14652014652014655
EPOCH  14       Train Error:  0.1309683177153057        Loss:  0.392024612506   Test Error:  0.14652014652014655
EPOCH  15       Train Error:  0.1309683177153057        Loss:  0.39082168793    Test Error:  0.14652014652014655
EPOCH  16       Train Error:  0.1309683177153057        Loss:  0.389815762018   Test Error:  0.14652014652014655
EPOCH  17       Train Error:  0.1309683177153057        Loss:  0.388953057251   Test Error:  0.14652014652014655
EPOCH  18       Train Error:  0.1309683177153057        Loss:  0.388192373523   Test Error:  0.14652014652014655
EPOCH  19       Train Error:  0.1309683177153057        Loss:  0.387500679777   Test Error:  0.14652014652014655
EPOCH  20       Train Error:  0.1309683177153057        Loss:  0.386852458833   Test Error:  0.14652014652014655
EPOCH  21       Train Error:  0.1309683177153057        Loss:  0.38622644868    Test Error:  0.14652014652014655
EPOCH  22       Train Error:  0.1309683177153057        Loss:  0.385604723123   Test Error:  0.14652014652014655
EPOCH  23       Train Error:  0.1309683177153057        Loss:  0.38497220248    Test Error:  0.14652014652014655
EPOCH  24       Train Error:  0.1309683177153057        Loss:  0.384315774454   Test Error:  0.14652014652014655
EPOCH  25       Train Error:  0.1309683177153057        Loss:  0.383623766086   Test Error:  0.14652014652014655
EPOCH  26       Train Error:  0.1309683177153057        Loss:  0.382885161656   Test Error:  0.14652014652014655
EPOCH  27       Train Error:  0.1309683177153057        Loss:  0.382089110805   Test Error:  0.14652014652014655
EPOCH  28       Train Error:  0.1309683177153057        Loss:  0.38122478688    Test Error:  0.14652014652014655
EPOCH  29       Train Error:  0.1309683177153057        Loss:  0.380282297348   Test Error:  0.14652014652014655
EPOCH  30       Train Error:  0.1309683177153057        Loss:  0.379250420097   Test Error:  0.14652014652014655

Train Set accuracy:  0.8690316822846943
Test Set Accuracy:  0.8534798534798534
```

[**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:

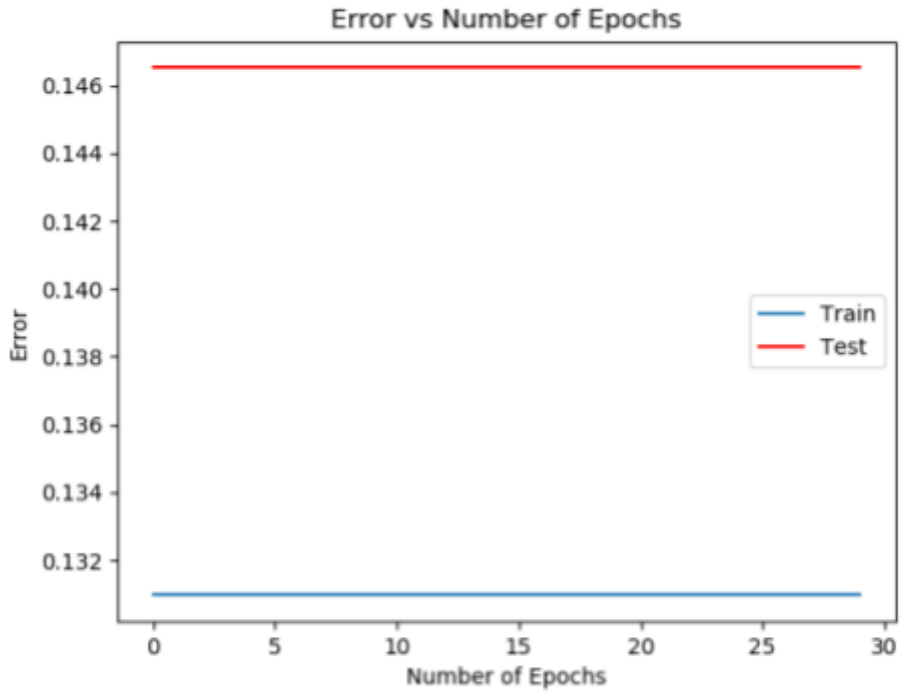

Fig. 1: Plot showing train error and test error(misclassification percentage) vs Epochs progressed.

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

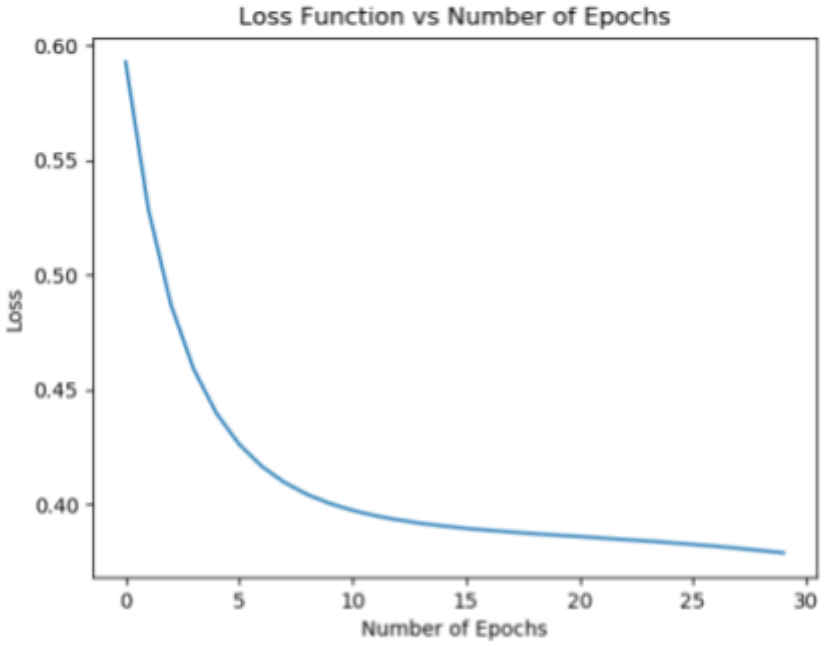

Fig. 2: Variation of Loss function vs Epochs progressed.

Final Test Accuracy: 0.8705035971223022

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 = 512
   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:  0.1284258210645527        Loss:  0.243986448191   Test Error:  0.1553062985332183
EPOCH  2        Train Error:  0.1284258210645527        Loss:  0.174198875115   Test Error:  0.1553062985332183
EPOCH  3        Train Error:  0.1284258210645527        Loss:  0.153253535615   Test Error:  0.1553062985332183
EPOCH  4        Train Error:  0.1284258210645527        Loss:  0.145489405919   Test Error:  0.1553062985332183
EPOCH  5        Train Error:  0.1284258210645527        Loss:  0.142374338484   Test Error:  0.1553062985332183
EPOCH  6        Train Error:  0.1284258210645527        Loss:  0.141083895213   Test Error:  0.1553062985332183
EPOCH  7        Train Error:  0.1284258210645527        Loss:  0.140542046835   Test Error:  0.1553062985332183
EPOCH  8        Train Error:  0.1284258210645527        Loss:  0.140313114903   Test Error:  0.1553062985332183
EPOCH  9        Train Error:  0.1284258210645527        Loss:  0.140216022136   Test Error:  0.1553062985332183
EPOCH  10       Train Error:  0.1284258210645527        Loss:  0.1401746642     Test Error:  0.1553062985332183
EPOCH  11       Train Error:  0.1284258210645527        Loss:  0.140156902918   Test Error:  0.1553062985332183
EPOCH  12       Train Error:  0.1284258210645527        Loss:  0.140149138986   Test Error:  0.1553062985332183
EPOCH  13       Train Error:  0.1284258210645527        Loss:  0.140145612788   Test Error:  0.1553062985332183
EPOCH  14       Train Error:  0.1284258210645527        Loss:  0.14014388405    Test Error:  0.1553062985332183
EPOCH  15       Train Error:  0.1284258210645527        Loss:  0.140142918671   Test Error:  0.1553062985332183
EPOCH  16       Train Error:  0.1284258210645527        Loss:  0.140142278392   Test Error:  0.1553062985332183
EPOCH  17       Train Error:  0.1284258210645527        Loss:  0.140141777461   Test Error:  0.1553062985332183
EPOCH  18       Train Error:  0.1284258210645527        Loss:  0.140141337145   Test Error:  0.1553062985332183
EPOCH  19       Train Error:  0.1284258210645527        Loss:  0.140140924061   Test Error:  0.1553062985332183
EPOCH  20       Train Error:  0.1284258210645527        Loss:  0.140140524056   Test Error:  0.1553062985332183
EPOCH  21       Train Error:  0.1284258210645527        Loss:  0.14014013112    Test Error:  0.1553062985332183
EPOCH  22       Train Error:  0.1284258210645527        Loss:  0.140139742698   Test Error:  0.1553062985332183
EPOCH  23       Train Error:  0.1284258210645527        Loss:  0.140139357701   Test Error:  0.1553062985332183
EPOCH  24       Train Error:  0.1284258210645527        Loss:  0.140138975657   Test Error:  0.1553062985332183
EPOCH  25       Train Error:  0.1284258210645527        Loss:  0.14013859636    Test Error:  0.1553062985332183
EPOCH  26       Train Error:  0.1284258210645527        Loss:  0.140138219713   Test Error:  0.1553062985332183
EPOCH  27       Train Error:  0.1284258210645527        Loss:  0.140137845669   Test Error:  0.1553062985332183
EPOCH  28       Train Error:  0.1284258210645527        Loss:  0.140137474199   Test Error:  0.1553062985332183
EPOCH  29       Train Error:  0.1284258210645527        Loss:  0.140137105283   Test Error:  0.1553062985332183
EPOCH  30       Train Error:  0.1284258210645527        Loss:  0.140136738905   Test Error:  0.1553062985332183

Train Set accuracy:  0.8715741789354473
Test Set Accuracy:  0.8446937014667817
```

[**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:

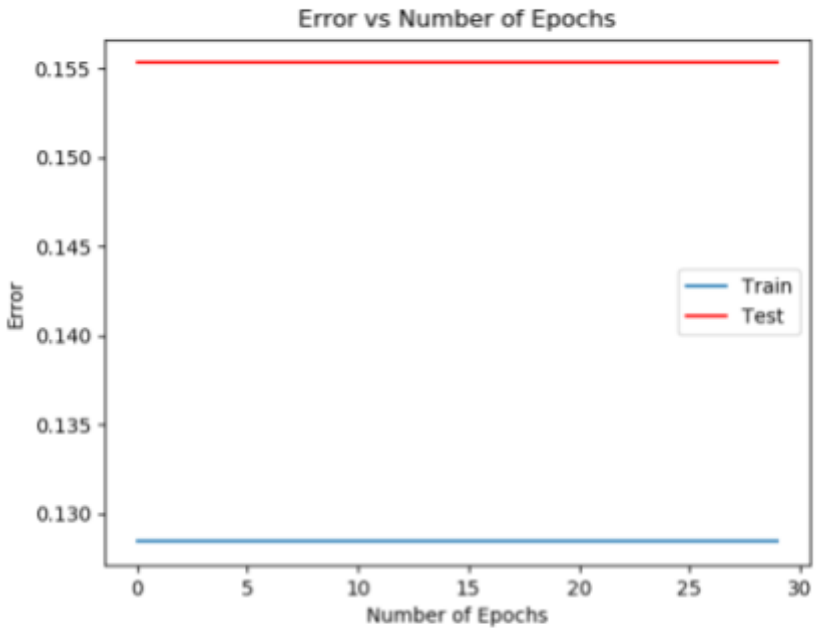

Fig. 3: Plot showing train error and test error(misclassification percentage) vs Epochs progressed.

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

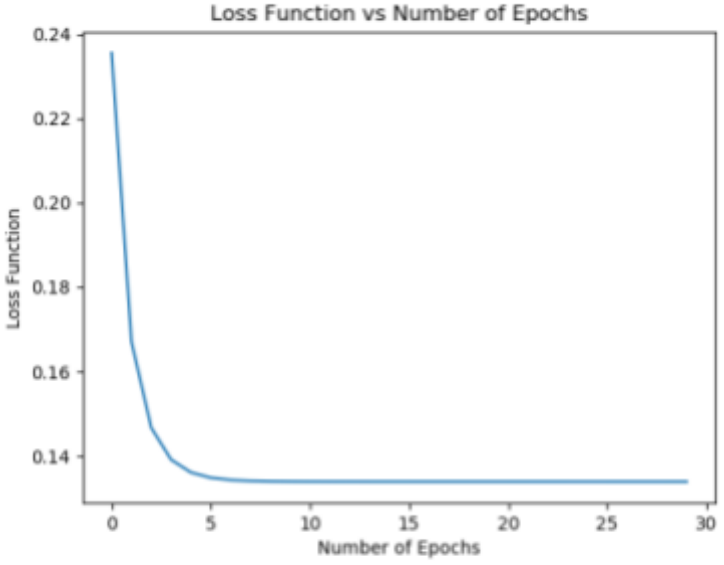

Fig. 4: Variation of Loss function vs Epochs progressed.

Final Test Accuracy: 0.8446937014667817
