HANDWRITTEN DIGIT CLASSIFICATION SYSTEM

INTRODUCTION

- The objective of this project is to solve the classification problem where the data is in the form of “handwritten digit images”
- Various ML techniques learnt are each applied on the data sets provided which are the MNIST dataset and the USPS dataset
- The Methods used are:
  Neural networks (in the form of a perceptron network)
  Random Forest Classifier
  Support Vector Machine classifier
- The “No Free Lunch” theorem is also examined by testing the trained networks obtained from these techniques and testing them on a different Data set (USPS)

BRIEF OVERVIEW OF THE DATASETS USED

- The MNIST dataset:
- Popular dataset which is publicly available in the form of handwritten digit images
- From this dataset, we take 60,000 samples for training and 10,000 samples for testing the performance of network

- The USPS dataset:
- In-House dataset developed at UB which also contains 2000 samples for each digit from 0 to 9 (20,000 in all)
- This dataset is used for testing purpose

BRIEF OVERVIEW OF THE METHODS USED

- Logistic Regression:
- It is a form of predictive analysis which uses binary dependent variables
- It is based on a sigmoid function, also known as a logistic function.
- It is used to determine the probability of an data element being in a class.

- A threshold can be set to determine whether a data point falls into a class or not.
- Higher the probability of a data for a class, more the confidence that it belongs to it.

- Neural Networks:
- A Multilayer Perceptron Network which has ten layers and takes an input in the form of 28X28 image for training purpose
- Number of classes is 10
- It is a feed-forward network that consists of input, hidden and output layers.
- They are good, even for data that is not linearly separable.
- An appropriate activation function can be decided according to data set

- Support Vector Machine:
- Support vector machine is a form of supervised learning in the form of a linear classifier
- Given a set of samples, a linear classifier draws a decision boundary in the sample space dividing the entire sample set into either one of the samples.
- SVM can be either hard margin or soft margin classifiers.
- A good SVM chooses a separation boundary which maintains a wide margin from both the set of points.
- This is done with the help of the support vectors which are the data points which are closest to boundary.

- Random Forest:
- It is another supervised learning technique which utilizes decision trees to form a "forest" of trees (models for prediction).
- It is nothing but a structure which utilizes top down structure for classification of data by splitting into many edges (taking a binary decision) until the condition is met.
- The data passes through the tree in a top down approach and proximities are computed for each pair.
- When there are no more splits, the node is known as a "leaf"

EXPERIMENT

- The four techniques have been applied on the data sets. For MNIST, 60,000 samples are for training and 10,000 samples for testing.

- The main training set is MNIST. Testing Dataset is USPS

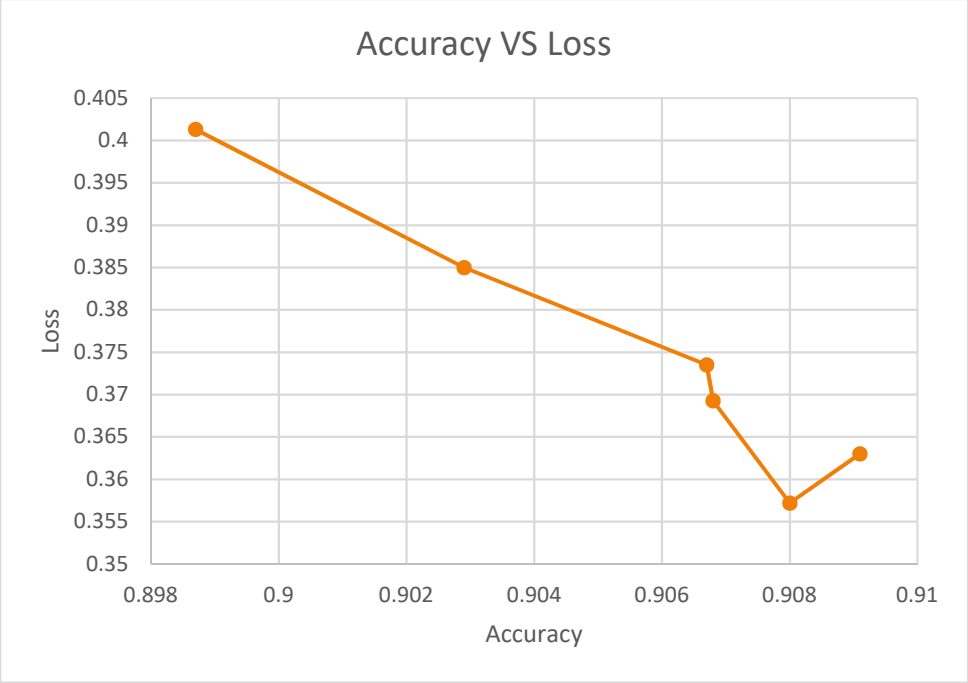
NEURAL NETWORKS

- Neural network produces results with high performance.
- The network is run for multiple iterations and the accuracy and loss values are recorded

Accuracy and loss values obtained for Neural network on MNIST dataset:

```
loss value 0.35727326385974884
Accuracay: 0.908

loss value 0.3630147574901581
Accuracay: 0.9091

loss value 0.4013926750421524
Accuracay: 0.8987

loss value 0.3693095678329468
Accuracay: 0.9068

loss value 0.3850791241168976
Accuracay: 0.9029

loss value 0.3735261542320251
Accuracay: 0.9067
```

Performance graph of Neural Net



CONFUSION MATRIX

- Confusion matrix is a visualization metric which can be used to display the overall output of a machine learning system
- An unequal number of classifications in a particular class can lead to some discrepancy
- A confusion matrix is a better way to judge the performance of a system.
- For each class, a confusion matrix provides the predicted results in comparison to the actual results.
- The number of incorrect to correct predictions over the total values is summarized.

SUPPORT VECTOR MACHINE

- The various important parameters for non-linear SVM are:
- C: For soft margin SVM, C is defined as the cost function. It is essentially a trade-off between the correct classifications and marginalization.
- A smaller value of C will encourage a larger margin
- Gamma: This parameter affects the influence of a particular training sample

SVM Performance analysis:

Generated Confusion matrix for SVM on MNIST dataset

```
[[ 958    0    3    3    1    5    4    0    3    2]
 [   0 1119    3    2    0    0    5    0    5    1]
 [  15    3  971   15    2    1    7    7   10    1]
 [   6    1   30  919    0   23    0    9   16    6]
 [   9    2   17    0  909    3   14    2    6   20]
 [  17    5    9   52   12  773    7    2    8    7]
 [  18    4   11    3   13   18  882    0    7    2]
 [   5    7   32    6    2    2    0  953    2   19]
 [  12    6   26   24   20   18    6    2  848   12]
 [   9    6    7   10   34    9    2   14   10  908]]
```

Generated Confusion matrix for SVM on USPS dataset

```
[[   0 2000    0    0    0    0    0    0    0    0]
 [   0 2000    0    0    0    0    0    0    0    0]
 [   0 1999    0    0    0    0    0    0    0    0]
 [   0 2000    0    0    0    0    0    0    0    0]
 [   0 2000    0    0    0    0    0    0    0    0]
 [   0 2000    0    0    0    0    0    0    0    0]
 [   0 2000    0    0    0    0    0    0    0    0]
 [   0 2000    0    0    0    0    0    0    0    0]
 [   0 2000    0    0    0    0    0    0    0    0]
 [   0 2000    0    0    0    0    0    0    0    0]]
```

RANDOM FORESEST

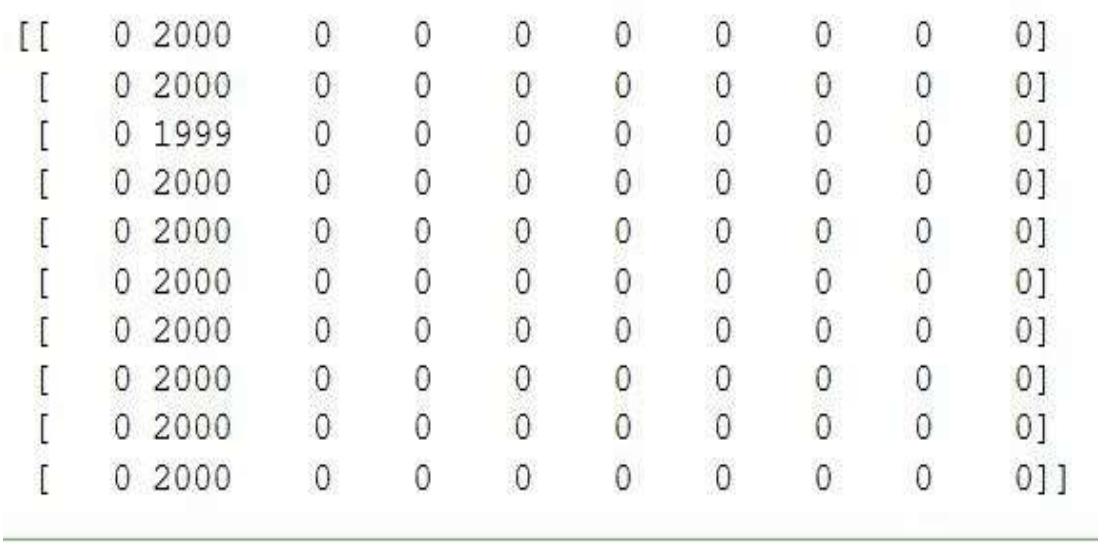
- N_estimators is an important parameter which

Random Forest Performance Analysis:

Generated Confusion matrix for Random Forest on MNIST dataset

```
[[ 967    1    1    0    1    3    3    2    1    0]
 [   0 1121    0    7    1    1    2    1    2    0]
 [  14    3  974    8    4    1    6   12   10    0]
 [   3    2   13  956    2   11    1    7   11    4]
 [   2    3    5    2  932    2    5    1    3   27]
 [   6    2    1   28    4  829    7    2    8    5]
 [  12    2    4    0    8   10  919    1    2    0]
 [   2    7   23    7    6    2    0  969    2   10]
 [   3    1   12   14    9   17    8    4  889   17]
 [   5    7    2   13   27    5    1    9   11  929]]
```

Generated Confusion matrix for Random Forest on USPS dataset



```
[[   0 2000    0    0    0    0    0    0    0    0]
 [   0 2000    0    0    0    0    0    0    0    0]
 [   0 1999    0    0    0    0    0    0    0    0]
 [   0 2000    0    0    0    0    0    0    0    0]
 [   0 2000    0    0    0    0    0    0    0    0]
 [   0 2000    0    0    0    0    0    0    0    0]
 [   0 2000    0    0    0    0    0    0    0    0]
 [   0 2000    0    0    0    0    0    0    0    0]
 [   0 2000    0    0    0    0    0    0    0    0]
 [   0 2000    0    0    0    0    0    0    0    0]]
```

QUESTIONS TO BE ANSWERED AND CONCLUSION

- According to “No Free lunch theorem”, just because a network is trained on a dataset for a particular problem and performs well, it will perform well for other datasets/problems in the same way.
- This means that when the performance of one method is averaged out over different Datasets/problems, there is no single superior method.
- Upon running the various models on the two datasets, we can see that the results clearly support the “No free Lunch Theorem”.
- For example, we can observe that there is significant performance difference between MNIST and USPS for SVM and Random Forest
- From confusion matrices, we can see that for MNIST (for which the system has trained), there is a high accuracy (refer diagonal elements). But for USPS, the performance is poor.
- We can observe that the performance of Neural Nets is generally the best. For MNIST dataset, a non-linear “Random Forest” is slightly better than “Support Vector Machine”.
- As the number of “n_estimators” increases, the accuracy of Random Forest increases.
- For SVM, as the value of “C” is minimized, it favors a wider margin.
- Overall, the methods combined produce a better result through majority voting rather than any single classifier.

REFERENCES

- https://scikit-learn.org/stable/index.html
- www.wikipedia.org
- https://ml-cheatsheet.readthedocs.io/en/latest/logistic_regression.html