Introduction to Machine Learning

Project 3 Report

UBID: sppandey@buffalo.edu

PERSON NO: 50288608

Objective:
We will be implementing four methods of machine learning to perform a task of classification on MNIST Test set and USPS Test set.

Goal :
The classification task will be that of recognizing a 28×28 grayscale handwritten digit image and identify it as a digit among 0, 1, 2, ... , 9.

Machine Learning Models used in the project:

1) Logistic Regression
2) Multilayer Perceptron Neural Network
3) Random Forest Package
4) Support Vector Machine Package

Concepts against which our models will be tested and judged:

1) **No Free Lunch Theorem :**

    a. NFL theorem focuses on proving that, if averaged over all problems, no optimization algorithm is expected to perform better than any other optimization algorithm.

    b. The number of solution candidates before reaching the optimal solution are the measure of performance.

    c. NFL theorem in simple words prove that the arithmetic mean of performance remains constant over all problems.

2) **Confusion Matrix:**

    a. When dealing with classification problems, confusion matrices are one such parameters that can be used to evaluate a model's performance.

    b. Thus, we can say that performance can be visualized using a confusion matrix.

    c. In the matrix each row of the Matrix represents the instances in a predicted class while each column represents the instances in an actual class.

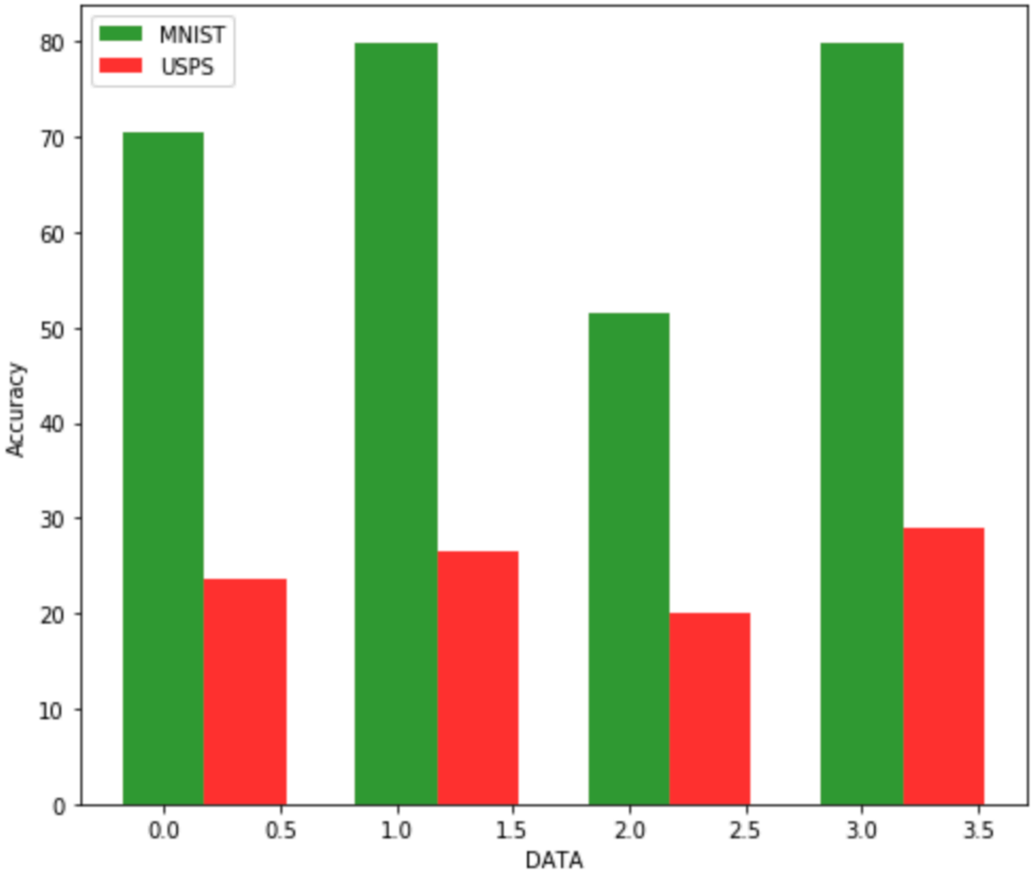
In report we will be looking at various scenarios for different models by tweaking the hyperparameters and try to get the optimal solution.

Logistic Regression:

In this special case of logistic regression where we have don't just have two outputs like [0,1] in case of sigmoid, here we use a multiclass classifier – Softmax for our regression.

Iterations	Learning Rate	Accuracy for MNIST Data	Accuracy for USPS Data
100	0.001	70.61	23.53
500	0.01	79.89	26.58
700	0.5	51.46	20.05
1000	0.1	79.92	28.96

Graph:



Observation:
One clear observation was that an increase in the number of iterations and decreasing learning rate helped in achieving better accuracy.

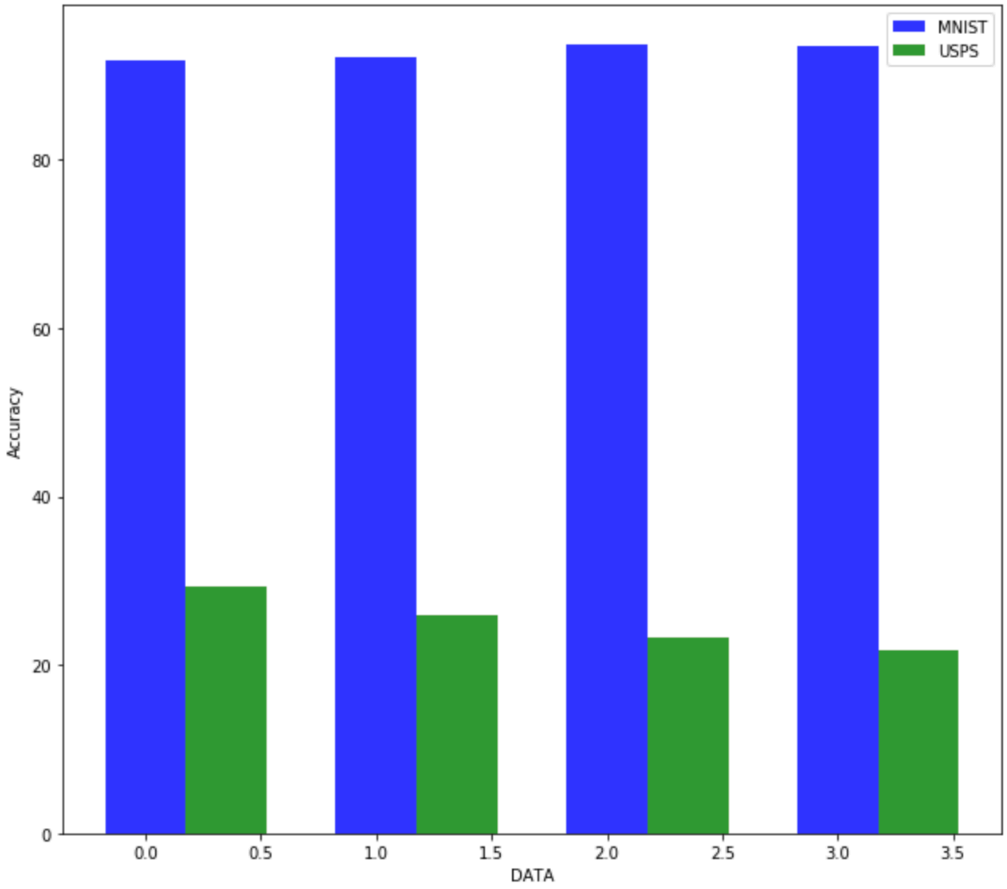
Neural Networks:

In neural networks we will be testing our model on different number of neurons and varying learning rate. For different settings we have observed the following the accuracies

For a constant learning rate of 0.01 we got the following accuracies:

Number of neurons in input layer	Number of neurons in hidden layert	Accuracy for MNIST Data	Accuracy for USPS Data
32	32	91.73	29.25
64	32	92.22	25.97
64	96	93.78	23.24
128	128	93.56	21.78

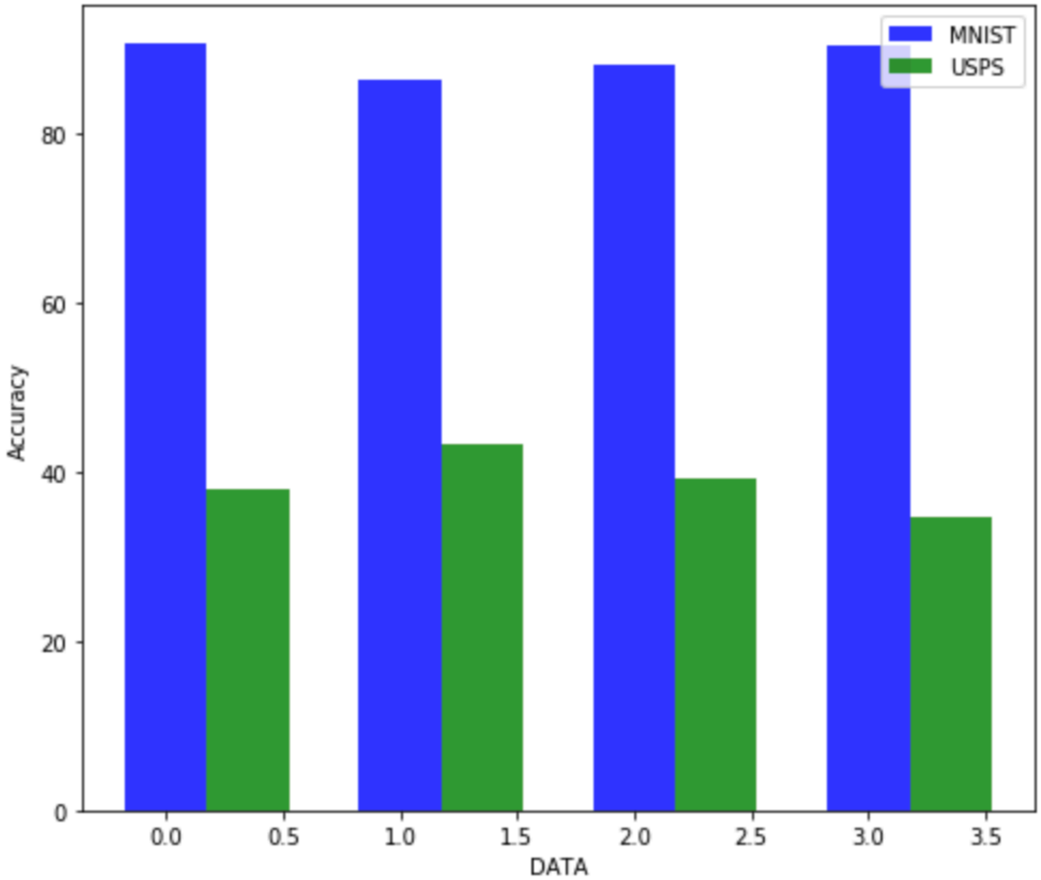
Graph:



In neural network since learning rate has far more significant impact on a constant setting of 32 neurons in input layer and 32 neurons in hidden layer we got the following accuracies.

Learning Rate	Accuracy for MNIST Data	Accuracy for USPS Data
0.01	90.78	37.89
0.1	86.33	43.33
0.05	88.19	39.21
0.02	90.46	34.79

Graph:



Observation :
For a learning rate 0.01 at constant number of neurons better results were observed. However the accuracy for USPS still didn’t give better results but were fluctuating in the observed range.

Support Vector Machine:

Support vector machines is a non-probabilistic classification machine learning model. The model is a representation of the examples as points in space such as vectors, which are mapped so that the different groups are divided by a clear gap that is as wide as possible.

For testing this model initially a huge amount of time was involved thus for the basic settings suc as kernel =’rbf, C=’2’ and Gamma = 0.05 an accuracy of 93.1 % was observed for MNIST data and 24.82 for USPS data.

Random Forest:

Random forests, also known as random decision forests, are a popular ensemble method that can be used to build predictive models for both classification and regression problems.

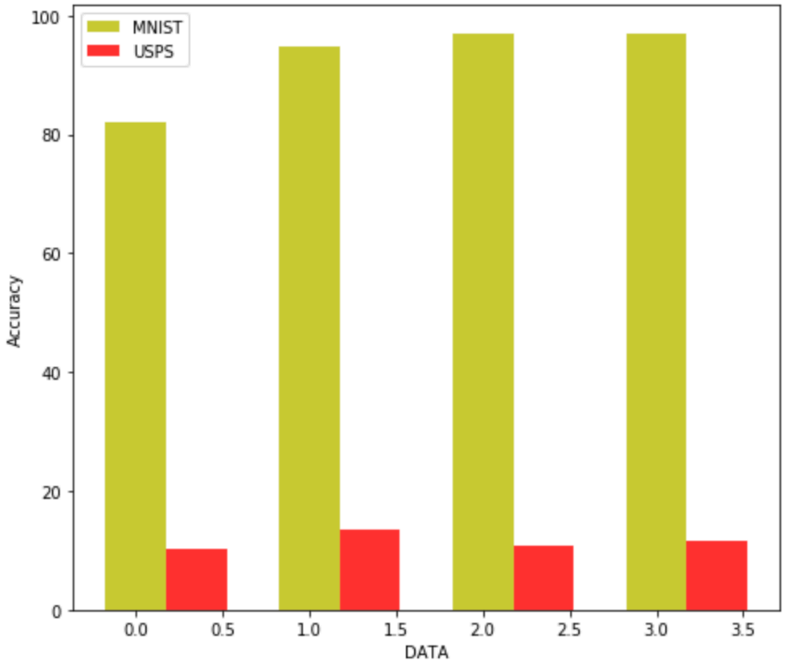
For this specific model we will be tuning the n_estimator values and max depth in some scenarios. In few of the test cases it was found that max_depth affected the accuracy in a negative fashion

Scenarios for Random Forest:

n_estimators indicate the number of trees in the forest. For testing this model we have used n_estimatiors starting from the lowest values to larger ones and the following observations were made.

n_estimators	Accuracy for MNIST Data	Accuracy for USPS Data
2	82.15	10.22
10	94.81	13.42
100	96.91	10.76
500	97.03	11.47

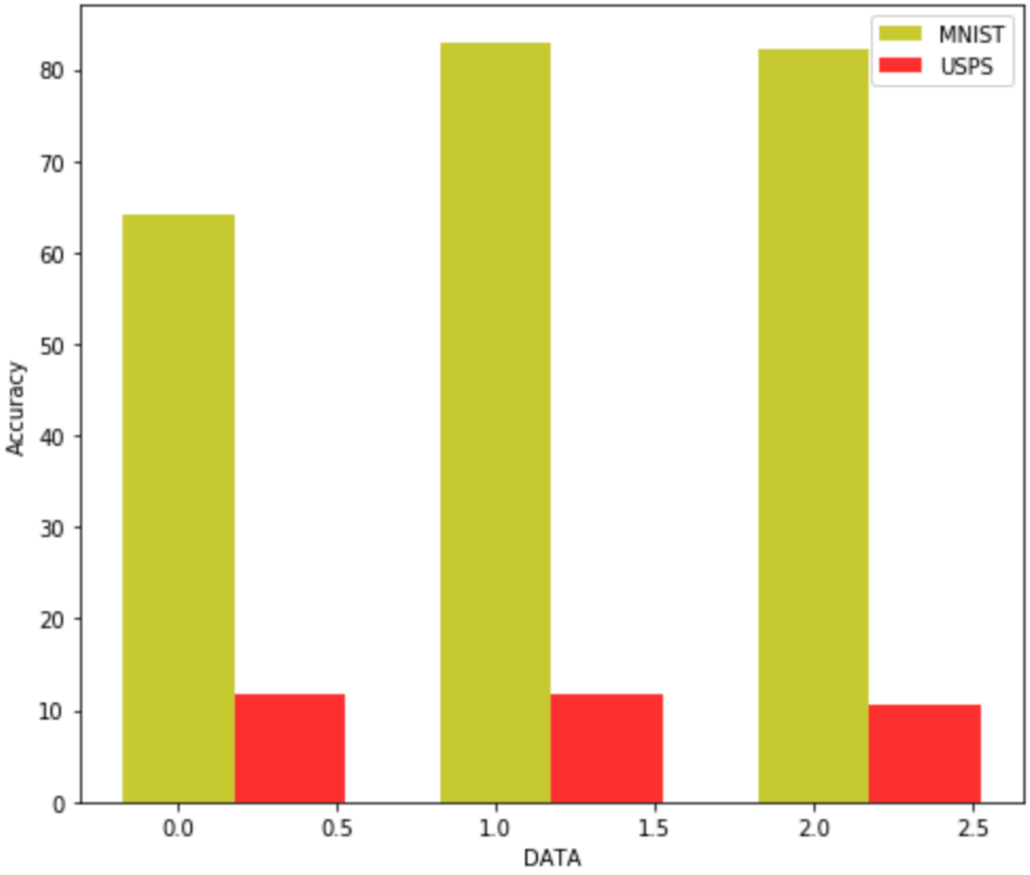
Graph:



max_depth indicates the depth of the tree. Since it was evident from previous cases that increase in number of n_estimators helps in getting a better accuracy atleast for the MNIST data thus for the cases with max depth we start with higher n_estimator values and following values were observed.

n_estimators	max_depth	Accuracy for MNIST Data	Accuracy for USPS Data
100	2	64.23	11.70
200	4	82.96	11.78
600	4	82.24	10.505

Graph :



Observation : The observed values showed that the model when tested on MNIST data gave better results when number of trees were increased. On the other hand USPS data didn’t give out promising results, testing data reflects once a maximum accuracy of approximately 28 percent was reached.

No Free Lunch Theorem Hypothesis: Since the theorem cleary states that no model is better than other but the mean of performance remains constant given for any model, here after our testing and observation we found that performance was nearly equal in all models and thus we believe that the No Free Lunch holds to be true in our project.

References:

[1] https://en.wikipedia.org/wiki/Random_forest
[2] https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/
[3] https://medium.com/@awjuliani/simple-softmax-in-python-tutorial-d6b4c4ed5c16
[4] https://www.python-course.eu/confusion_matrix.php
[5]https://scikit-learn.org/stable/auto_examples/model_selection/plot_confusion_matrix.html#sphx-glr-auto-examples-model-selection-plot-confusion-matrix-py
[6] https://www.kdnuggets.com/2016/07/softmax-regression-related-logistic-regression.html