CS7015: DEEP LEARNING
Assignment 1 Report

Team : 7

Authors:
Rahul Chakwate:AE16B005
Soham Dixit:CS16B006
Pranav Gadikar:CS16B115

March 10, 2019

CONTENTS

1 Function Approximation Task

1.1 Introduction

Neural Networks, though designed mainly for the classification tasks are also useful in regression tasks like the function approximation task. If a single neuron is present in the output layer, it can be used to approximate a functional output of the input features.

1.2 Dataset

The given dataset for group 7 is Concrete Compressive Strength Data Set.

Data Set Characteristics: Multivariate
Number of Instances: 1030
Number of Attributes/ input features: 8
Number of output variables: 1
Train:Test split = 70:30

1.3 Baseline Configuration

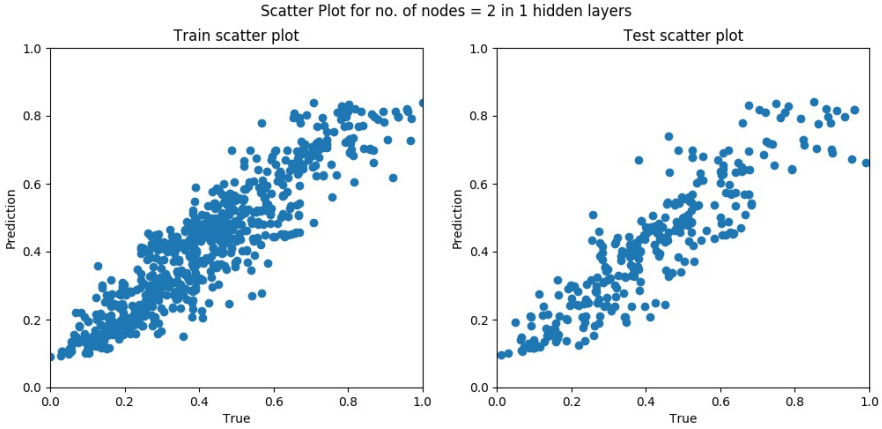
A baseline model is designed to decide on the number of hidden layers, number of nodes in those layers and the slope of the logistic function.

This baseline model was found out using Pattern mode, normalized features, sum of squared error, logistic function and delta rule as the primary configuration. Number of nodes and number of hidden layers were varied one at a time to find out the best architecture. Then the slope of the activation function was varied to choose the best slope for least sum of squared error.

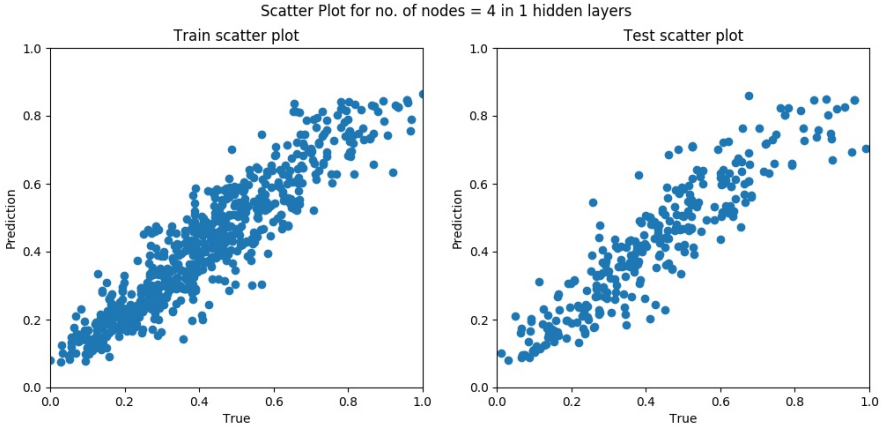
1.3.1 Number of nodes in 1 hidden layer

The first experiment is carried out by changing the number of nodes in 1 hidden layer. 2,4,8,12 and 20 nodes in the hidden layer are experimented.

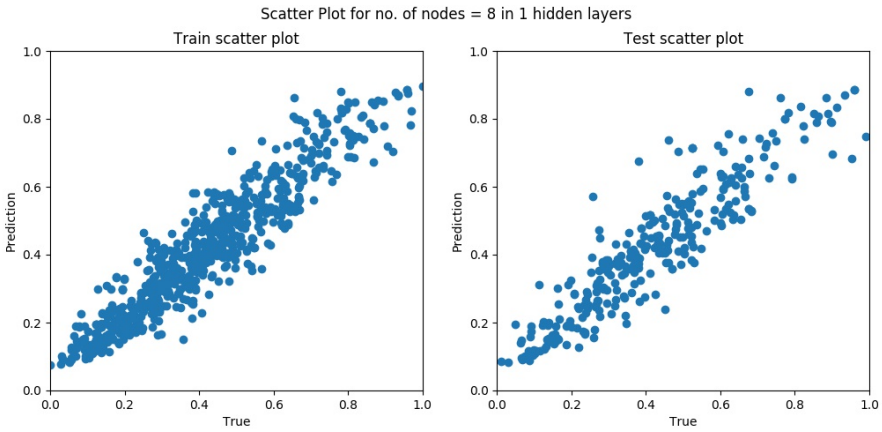
The following are the experiments.


Scatter Plot for no. of nodes = 2 in 1 hidden layers
Train scatter plot
Test scatter plot
Prediction
True
0.0
0.2
0.4
0.6
0.8
1.0

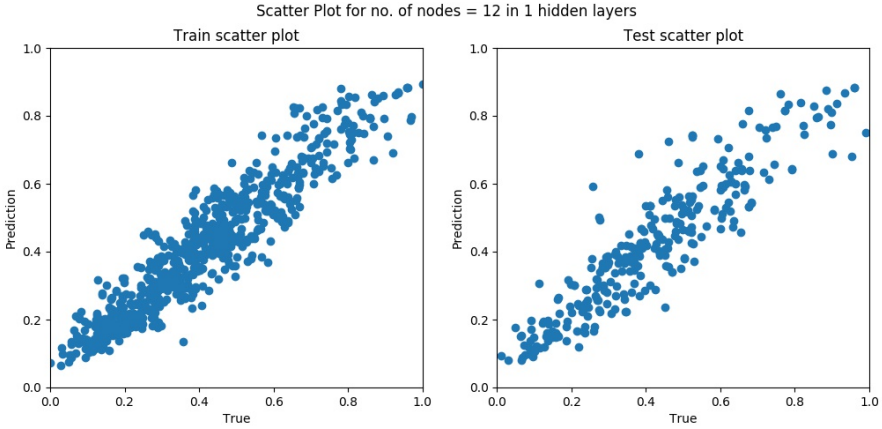
(a)


Scatter Plot for no. of nodes = 4 in 1 hidden layers
Train scatter plot
Test scatter plot
Prediction
True
0.0
0.2
0.4
0.6
0.8
1.0

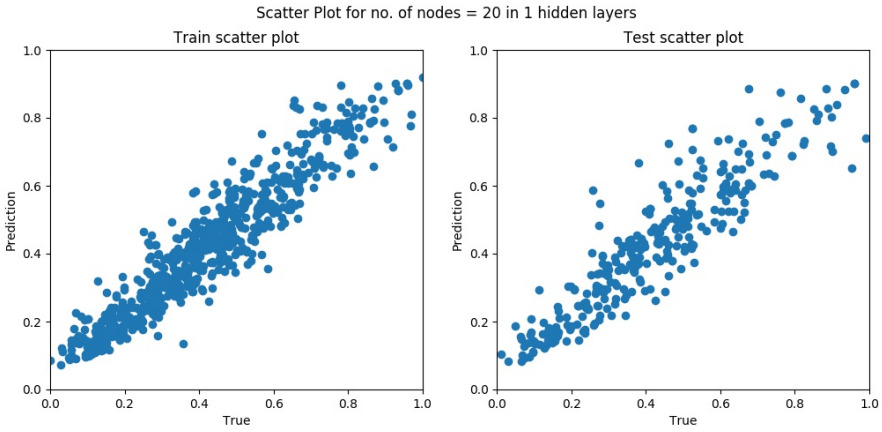
(b)


Scatter Plot for no. of nodes = 8 in 1 hidden layers
Train scatter plot
Test scatter plot
Prediction
True
0.0
0.2
0.4
0.6
0.8
1.0

(c)



(d)



(e)

Figure 1: (a)-(e) Scatter Plots for different no. of nodes in 1 Hidden layer

Table 1: Convergence Epoch and MSE for different no. of nodes in 1 hidden layer.

No of hidden nodes	Convergence epoch	Train MSE	Test MSE
2	640	0.00706	0.00803
4	711	0.00593	0.00721
8	1153	0.00504	0.00687
12	1061	0.00505	0.00713
20	1347	0.0046	0.00699

The variation of train MSE and test MSE versus epoch with different no. of hidden nodes is shown in the graph below.

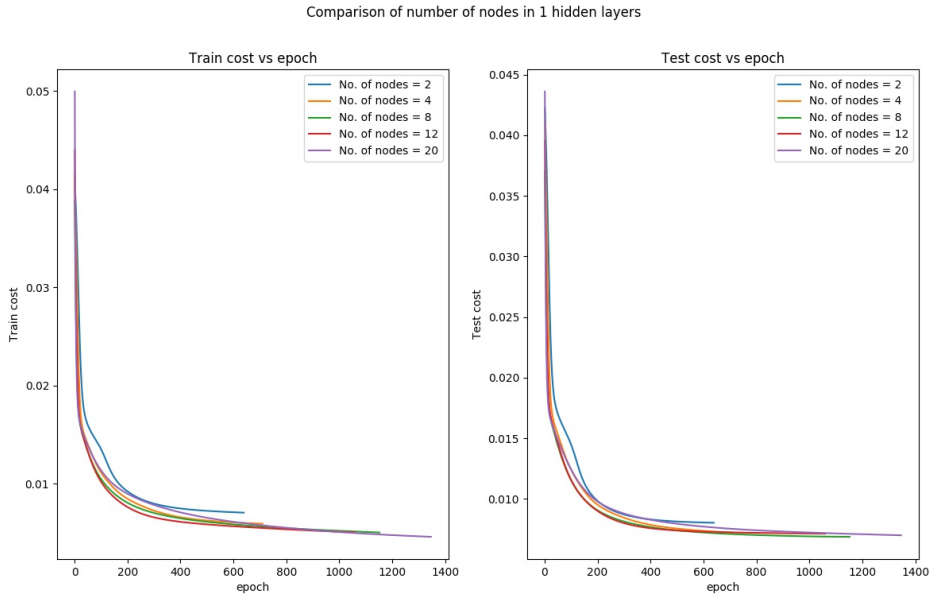
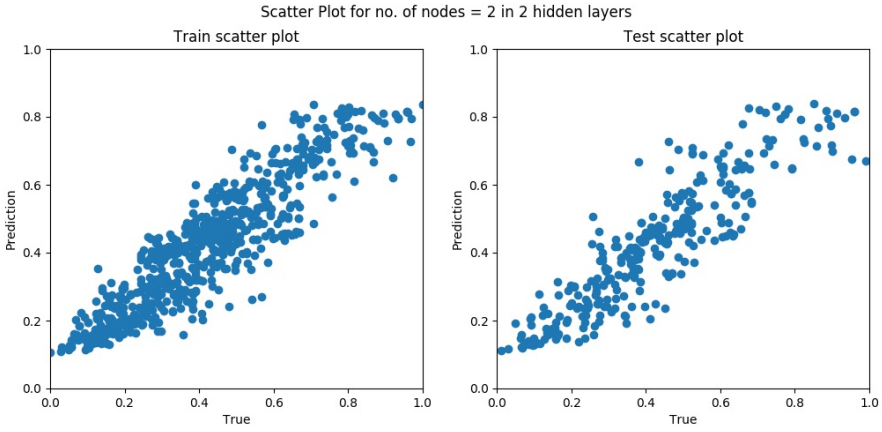


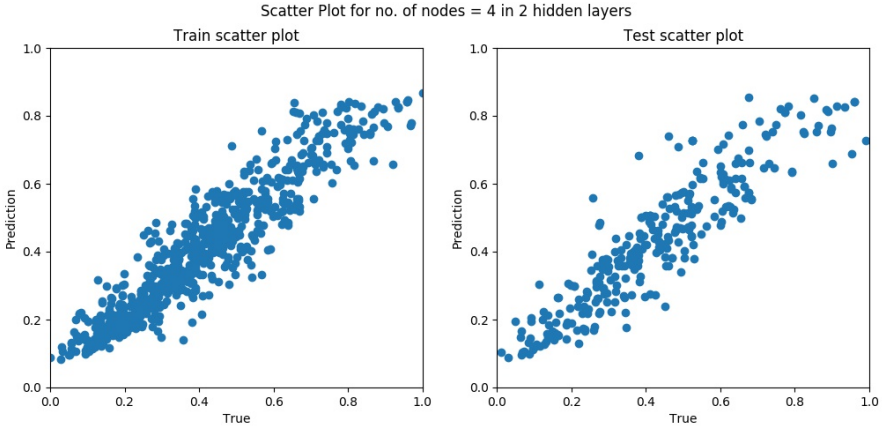
Figure 2: Effect of different no. of nodes in 1 hidden layer on train and test MSE

1.3.2 Number of nodes in 2 hidden layers

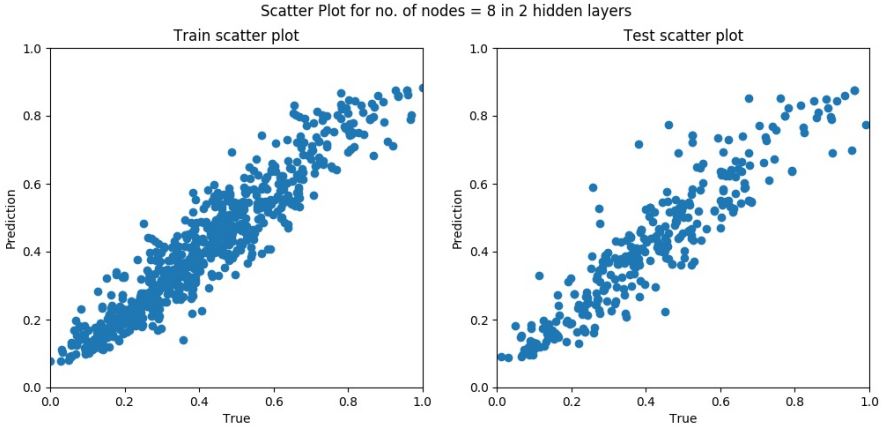
Number of hidden layers is increased to two and similar experiments are performed.



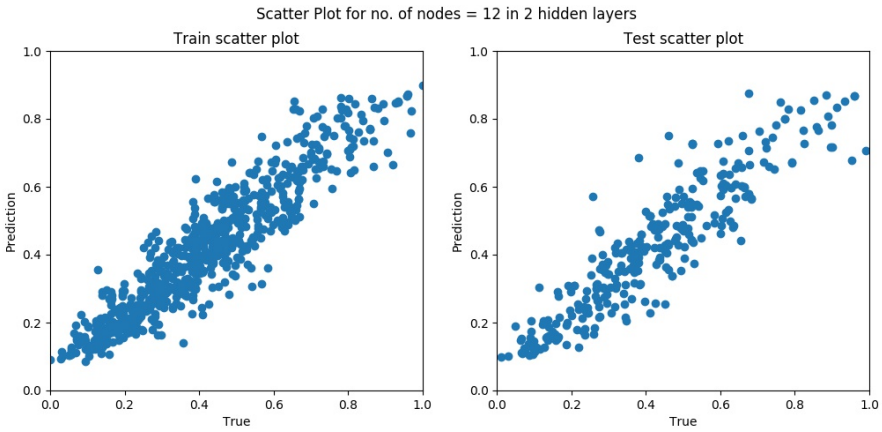
(a)


Scatter Plot for no. of nodes = 4 in 2 hidden layers
Train scatter plot
Test scatter plot
Prediction
True
0.0
0.2
0.4
0.6
0.8
1.0

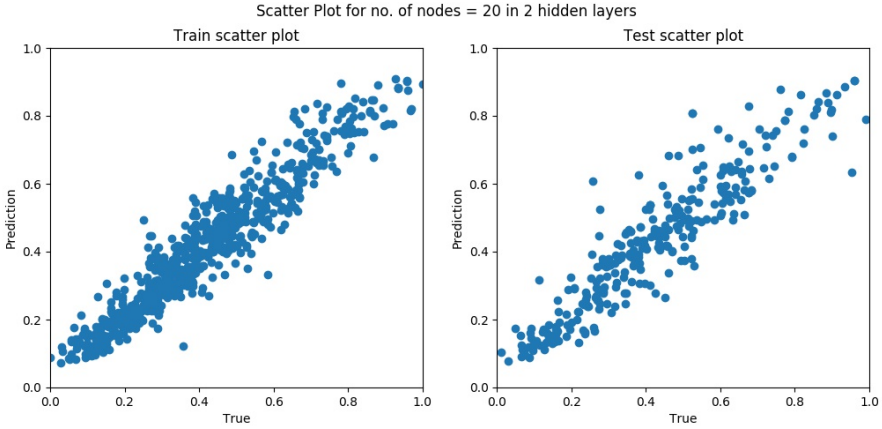
(b)


Scatter Plot for no. of nodes = 8 in 2 hidden layers
Train scatter plot
Test scatter plot
Prediction
True
0.0
0.2
0.4
0.6
0.8
1.0

(c)


Scatter Plot for no. of nodes = 12 in 2 hidden layers
Train scatter plot
Test scatter plot
Prediction
True
0.0
0.2
0.4
0.6
0.8
1.0

(d)



(e)

Figure 3: (a)-(e) Scatter Plot for different no. of nodes in 2 Hidden Layers

Table 2: Convergence Epoch and MSE for different no. of nodes in 2 hidden layers.

No of hidden nodes	Convergence epoch	Train MSE	Test MSE
2	909	0.00702	0.00797
4	873	0.00556	0.0072
8	1329	0.00465	0.00706
12	898	0.00572	0.0073
20	1700	0.00387	0.00651

The variation of train MSE and test MSE versus epoch with different no. of nodes in 2 hidden layers is shown in the graph below.

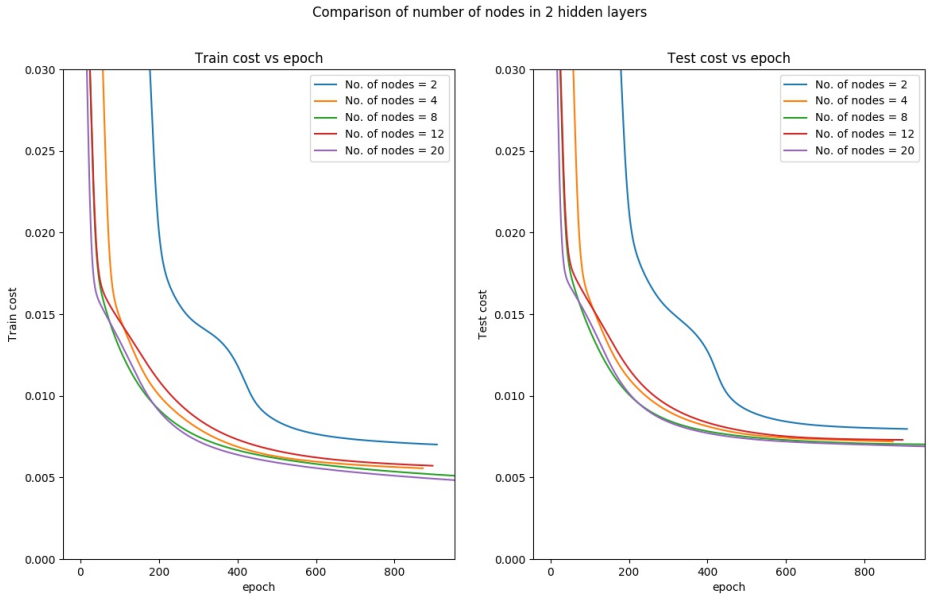
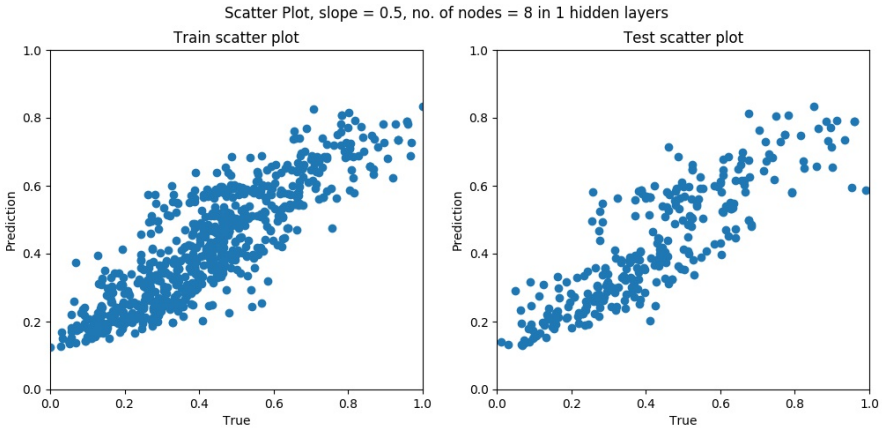


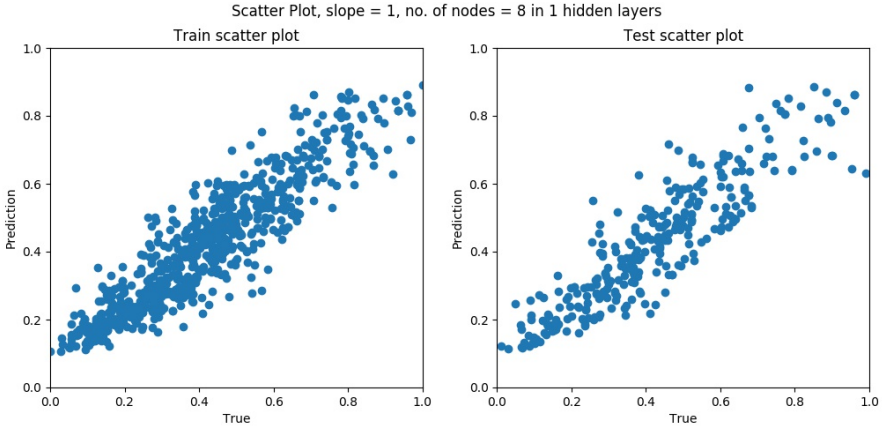
Figure 4: Effect of different no. of nodes in 2 Hidden Layers on train and test MSE

1.3.3 SLOPE OF LOGISTIC FUNCTION

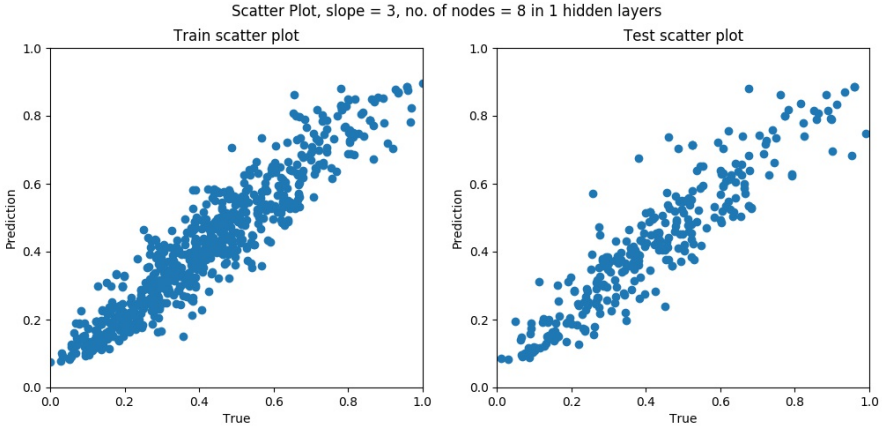
Keeping the activation function the same as logistic function slope of the logistic function is varied.



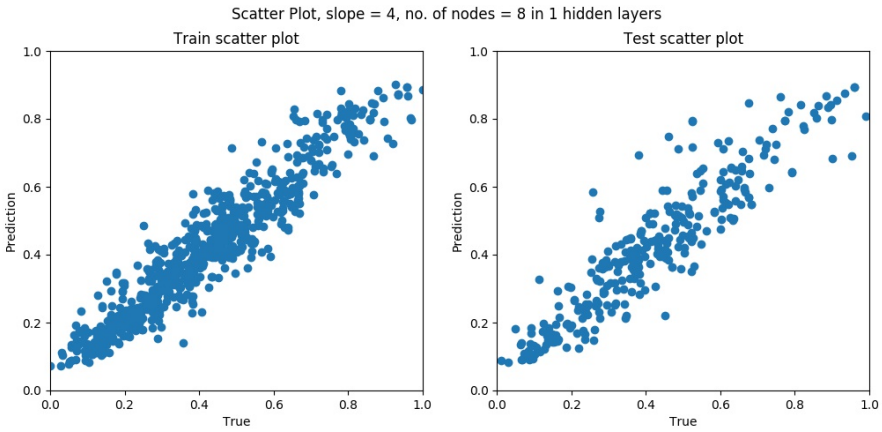
(a)


Scatter Plot, slope = 1, no. of nodes = 8 in 1 hidden layers
Train scatter plot
Test scatter plot
Prediction
True
0.0
0.2
0.4
0.6
0.8
1.0

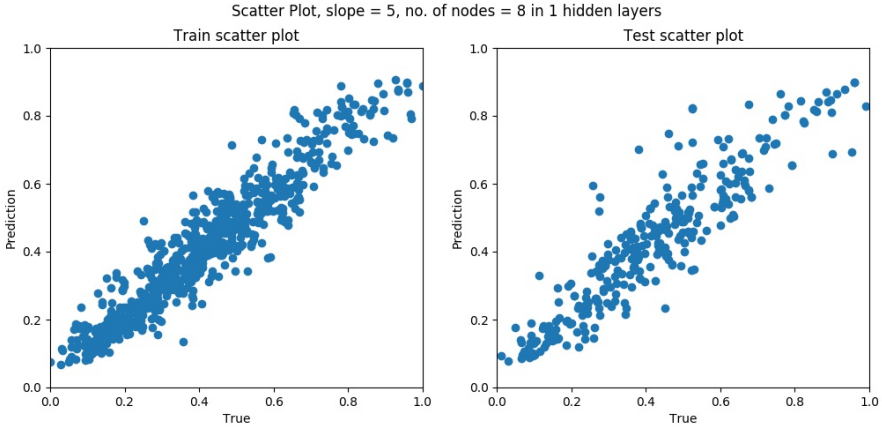
(b)


Scatter Plot, slope = 3, no. of nodes = 8 in 1 hidden layers
Train scatter plot
Test scatter plot
Prediction
True
0.0
0.2
0.4
0.6
0.8
1.0

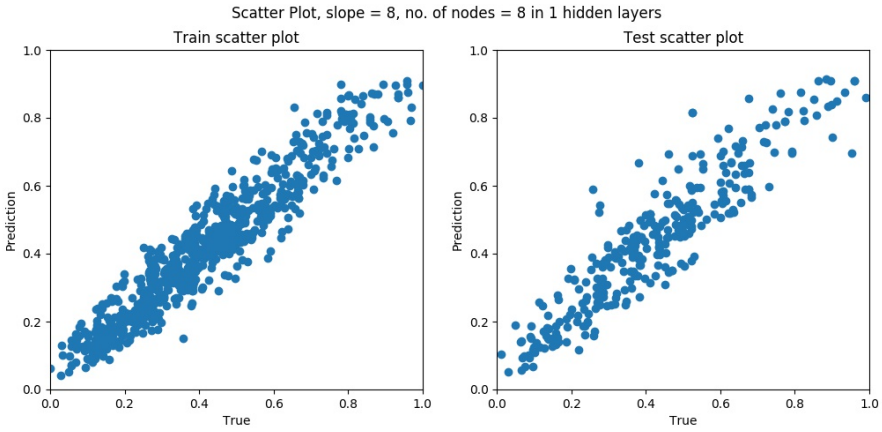
(c)


Scatter Plot, slope = 4, no. of nodes = 8 in 1 hidden layers
Train scatter plot
Test scatter plot
Prediction
True
0.0
0.2
0.4
0.6
0.8
1.0

(d)



(e)



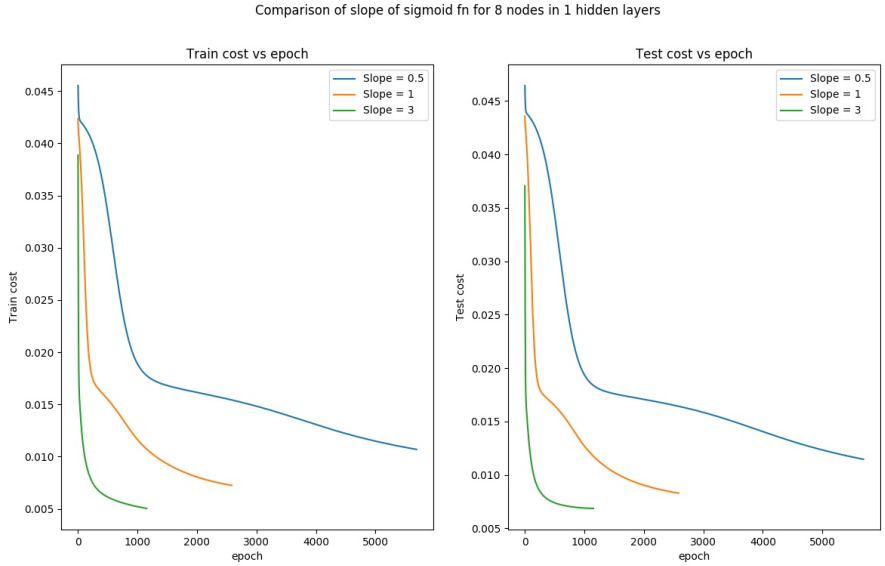
(f)

Figure 5: (a)-(f) Scatter Plot for different slopes of Logistic function

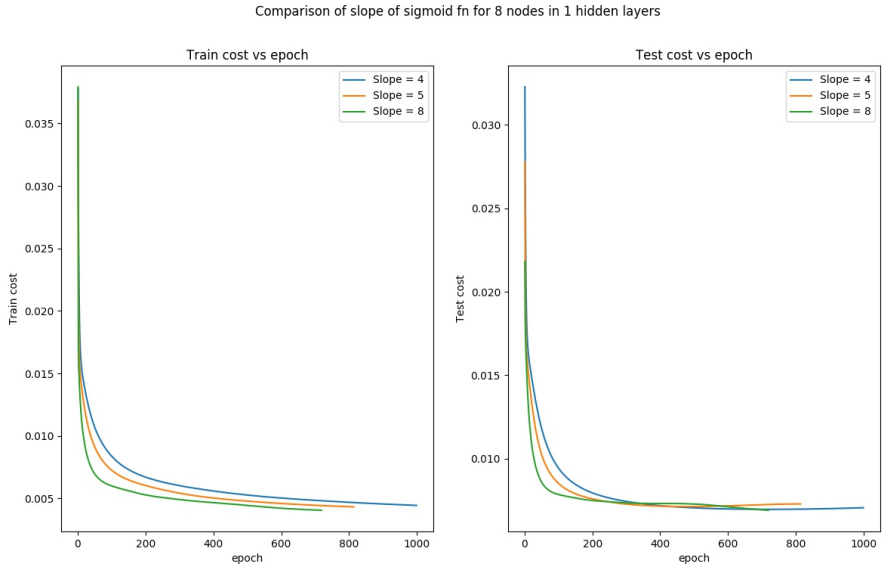
Table 3: Convergence Epoch and MSE for different slope of logistic function applied to all layers.

slope	Convergence Epoch	Train MSE	Test MSE
0.5	5691	0.01069	0.01147
1.0	2585	0.00724	0.0083
3.0	1153	0.00504	0.00687
4.0	1000	0.00445	0.00705
5.0	815	0.00433	0.00728
8.0	720	0.00406	0.0069

Comparison of behavior of train cost and test cost versus epochs is made for different slopes of the sigmoid function.



(a)



(b)

Figure 6: (a)-(b) Effect of different slopes on train and test MSE

1.3.4 Conclusion

From the above experiments we can conclude the following:

- As number of nodes increases, convergence epoch also increases, train MSE decreases monotonically but test MSE decreases to a certain point. (Refer: Figure 1., Figure2. and Table 1.)
- As number of hidden layers is increased to 2, test MSE is greater than that corresponding to 1 hidden layer for each number of nodes. Convergence epoch also increases as more parameters have to converge to a certain value. (Refer: Figure 3., Figure 4. and Table 2.)
- Hence, we choose model with 8 nodes in 1 hidden layer as it gives least test MSE with less over-fitting.
- As the slope of logistic function is increased, convergence epoch decreases, train MSE decreases, test MSE also decreases upto some point but over-fitting increases. (Refer: Figure 5., Figure 6. and Table 3.)
- Hence optimal slope for logistic function is 3 which has least test MSE and decent over-fitting and convergence epoch.

1.3.5 Final Baseline Configuration

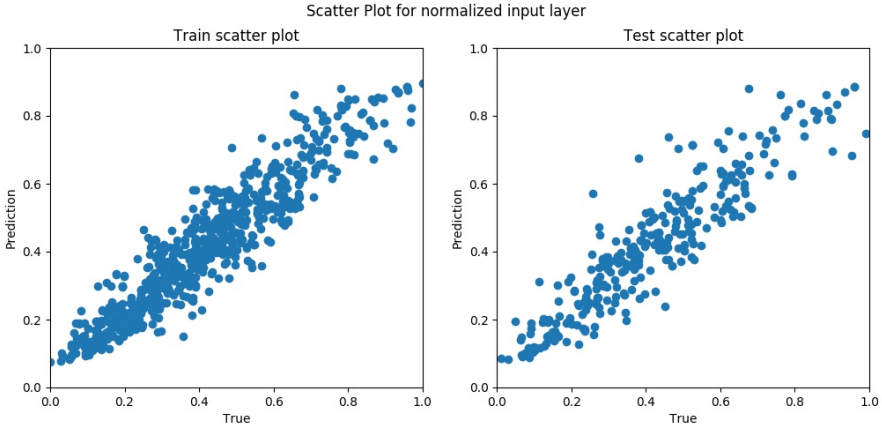
From the above, we can conclude that the model containing 8 nodes in 1 hidden layer with logistic slope of 3 performs better than the other models in terms of convergence epoch, test cost and not over-fitting. Hence this configuration is chosen as the baseline configuration.

1.4 Experiments

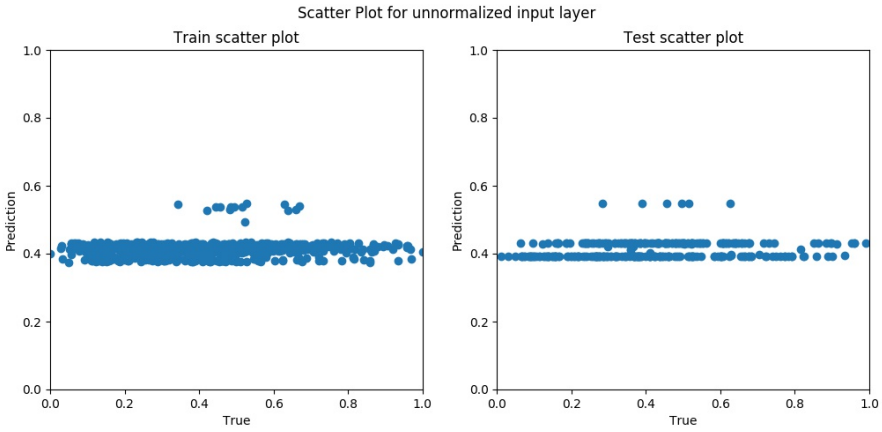
For conducting the experiments, the above final baseline configuration is used. Loss function used for function approximation task is the sum of squared error loss.

1.4.1 Normalized versus Unnormalized input features

In this section, comparison is made between feeding the inputs directly to the network versus normalizing the inputs before feeding the network.



(a)



(b)

Figure 7: (a)-(b) Scatter Plot for normalized v/s unnormalized input features

Table 4: Convergence Epoch and MSE for normalized v/s unnormalized input features.

Input Type	Convergence Epoch	Train MSE	Test MSE
unnormalized	3	0.04183	0.04455
normalized	1153	0.00504	0.00687

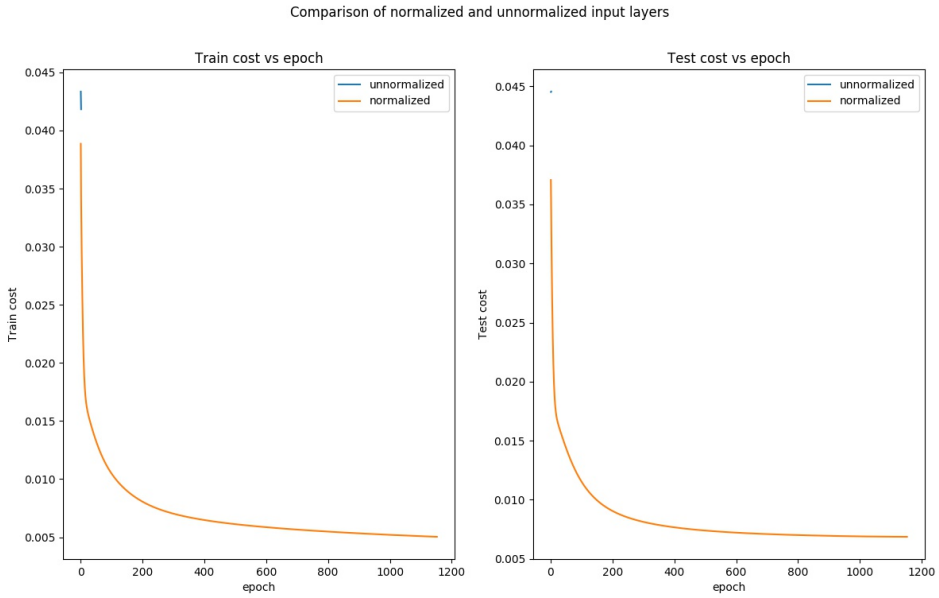


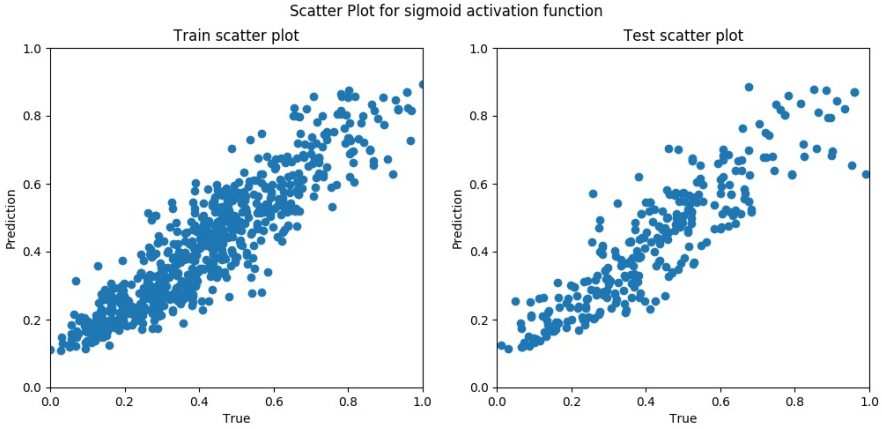
Figure 8: Effect of normalization on train and test MSE

Conclusion:

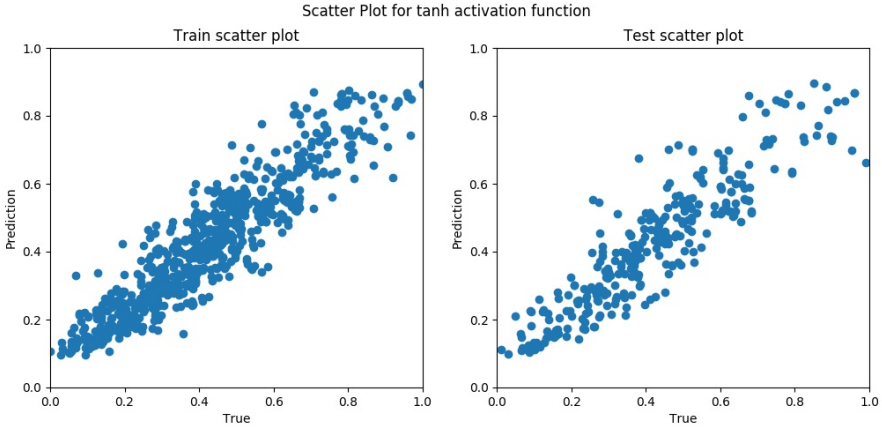
- From Figure 7. and Table 4. clearly normalizing input features outperforms unnormalized features. This is because the logistic function performs well with normalized inputs.
- If inputs are not normalized, output of the node will mostly lie in the saturation region.
- If inputs are normalized, logistic function will operate in the active region and hence give better output.

1.4.2 Activation Functions

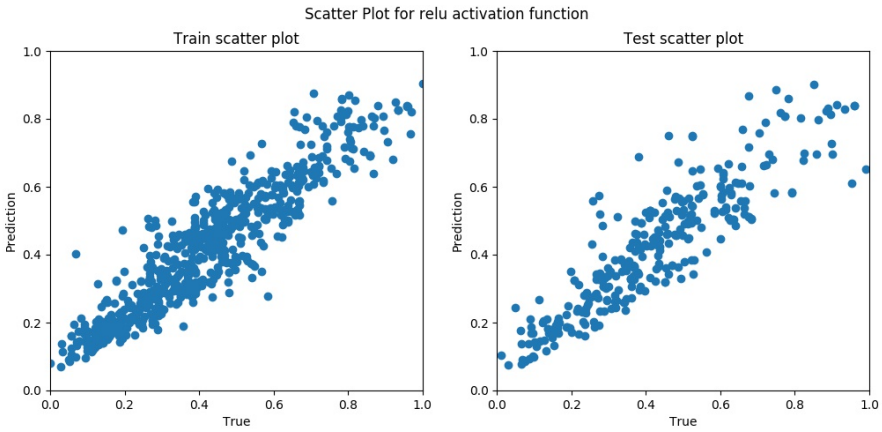
Various activations functions including Logistic, Tanh, ReLU, Softplus and ELU are compared.


Scatter Plot for sigmoid activation function
Train scatter plot
Test scatter plot
Prediction
True
0.0
0.2
0.4
0.6
0.8
1.0

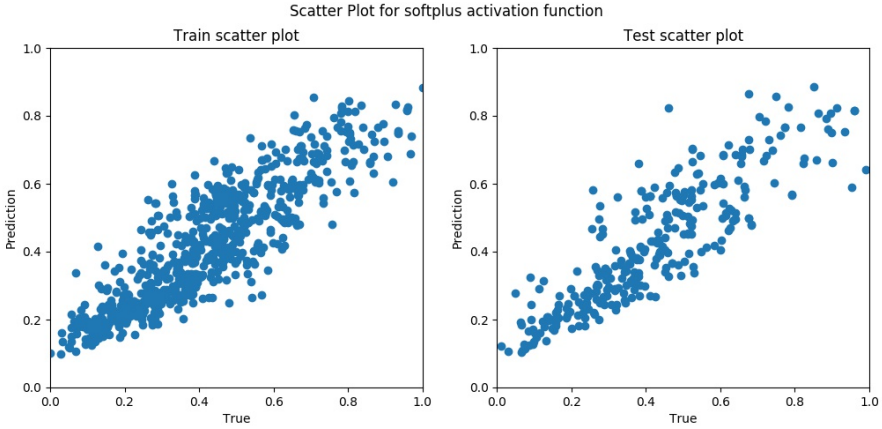
(a)


Scatter Plot for tanh activation function
Train scatter plot
Test scatter plot
Prediction
True
0.0
0.2
0.4
0.6
0.8
1.0

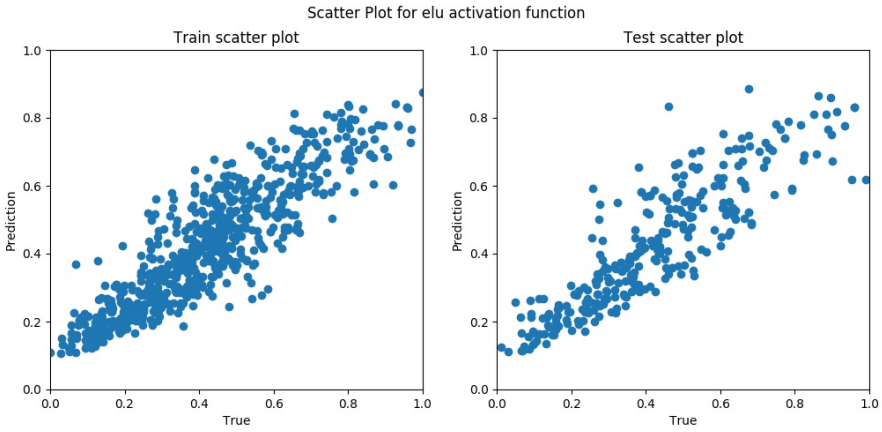
(b)


Scatter Plot for relu activation function
Train scatter plot
Test scatter plot
Prediction
True
0.0
0.2
0.4
0.6
0.8
1.0

(c)



(d)



(e)

Figure 9: (a)-(e) Scatter Plot for different Activation Functions

Table 5: Convergence Epoch and MSE for different Activation Functions.

Activation Function	Convergence Epoch	Train MSE	Test MSE
sigmoid	251	0.00745	0.0085
tanh	141	0.00622	0.00765
relu	247	0.00621	0.00858
softplus	386	0.00942	0.01071
elu	387	0.00878	0.01006

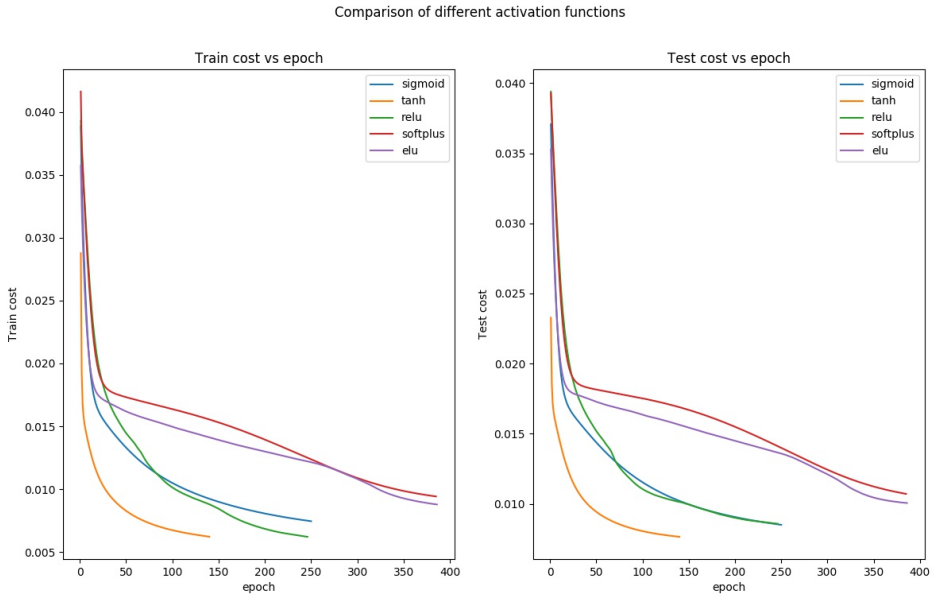


Figure 10: Effect of Activation function on train and test MSE

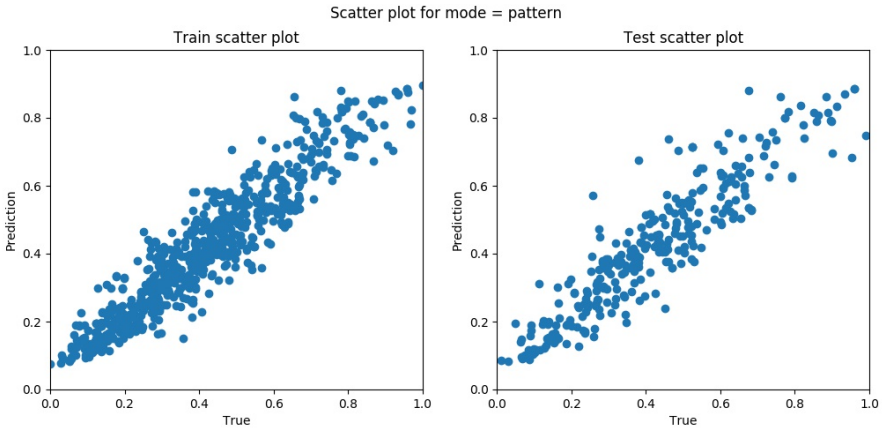
Conclusion:

- Each activation functions perform differently on different datasets. In many image classification models, ReLU has shown to perform better than other models.
- However, for the given dataset for function approximation task, from Figure 10. and Table 5. it is clear that "tanh" activation function outperforms every other activation function giving least test MSE in least no. of epochs for convergence.

1.4.3 Learning Mode

Pattern Mode and Batch Mode of learning are studied in this experiment.

(a)



(b)

Figure 11: (a)-(b) Scatter Plot for different modes of learning

Table 6: Convergence Epoch and MSE for different modes of learning.

Learning Mode	Convergence Epoch	Train MSE	Test MSE
pattern	1153	0.00504	0.00687
batch	7775	0.01932	0.01999

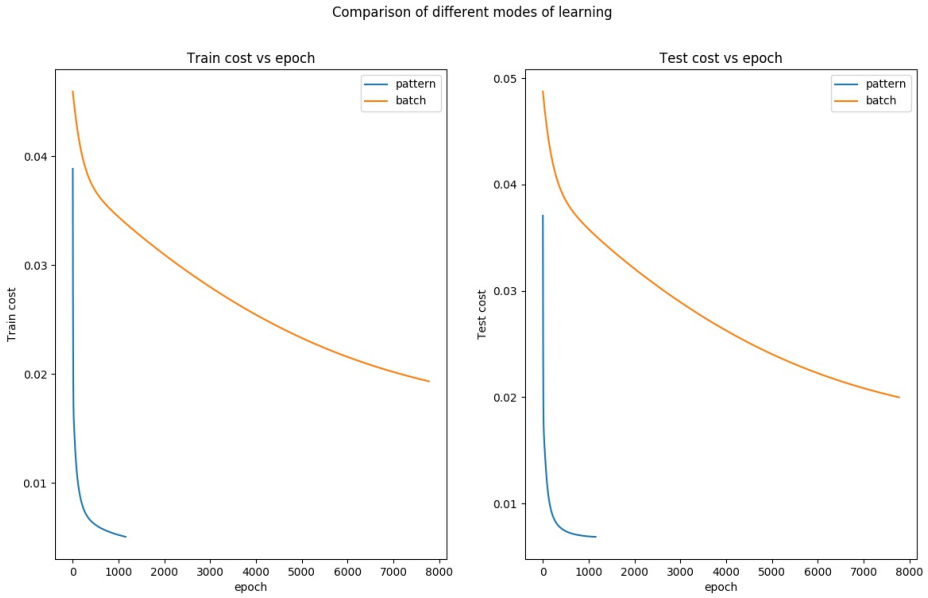


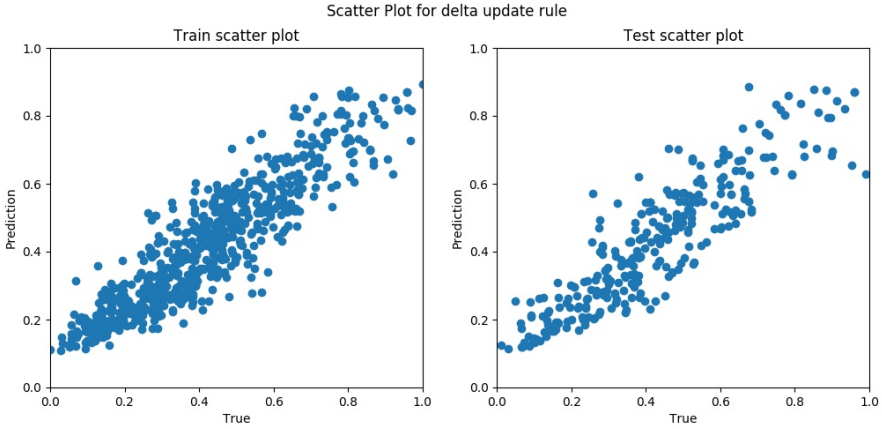
Figure 12: Effect of different modes of learning on train and test MSE

Conclusion:

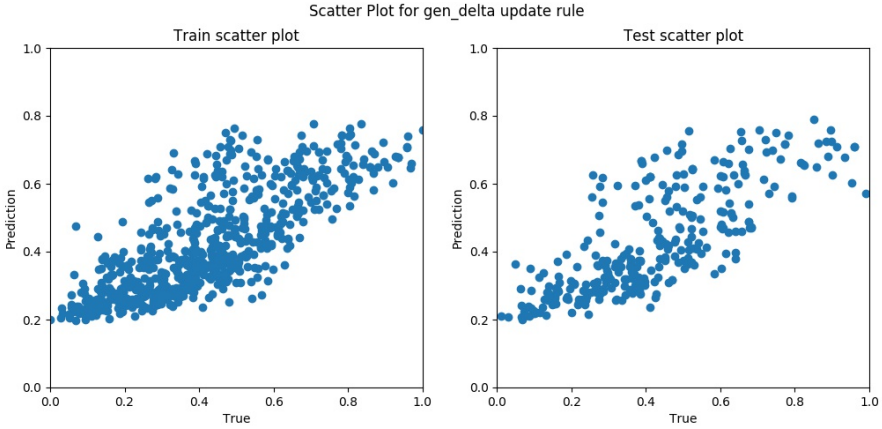
- Referring to Figure 11., Table 6. and Figure 12. one can make out that pattern mode significantly works better and faster than batch mode.
- Pattern mode converges 7 times faster in terms of number of epochs for convergence.
- It converges to test MSE which is about a third of test MSE of batch mode.
- However, one epoch of pattern mode takes lot more time than one epoch of batch mode which updates all the weights simultaneously.

1.4.4 Weight Update Rules

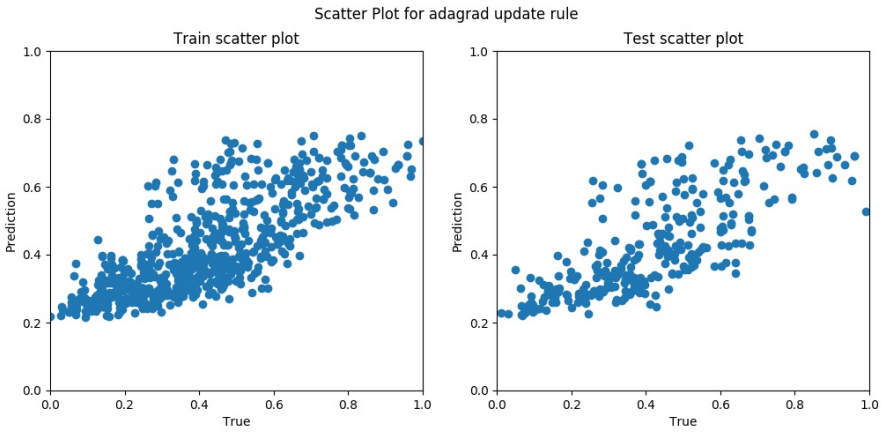
Comparison between the performance of different optimizers is made in this experiment.


Scatter Plot for delta update rule
Train scatter plot
Test scatter plot
Prediction
True
0.0
0.2
0.4
0.6
0.8
1.0

(a)


Scatter Plot for gen_delta update rule
Train scatter plot
Test scatter plot
Prediction
True
0.0
0.2
0.4
0.6
0.8
1.0

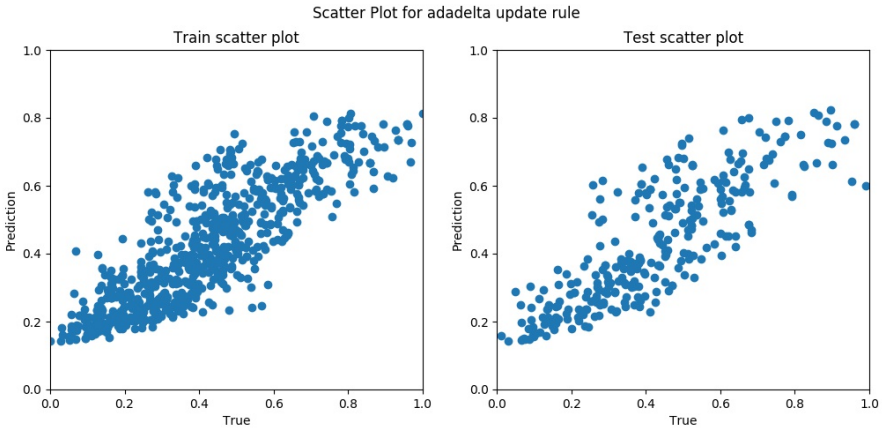
(b)


Scatter Plot for adagrad update rule
Train scatter plot
Test scatter plot
Prediction
True
0.0
0.2
0.4
0.6
0.8
1.0

(c)


Scatter Plot for rmsprop update rule
Train scatter plot
Test scatter plot
Prediction
True
0.0
0.2
0.4
0.6
0.8
1.0

(d)


Scatter Plot for adadelta update rule
Train scatter plot
Test scatter plot
Prediction
True
0.0
0.2
0.4
0.6
0.8
1.0

(e)

(f)

Figure 13: (a)-(f) Scatter Plot for different update rules.

Table 7: Convergence Epoch and MSE for different update rules.

Optimizer	Convergence Epoch	Train MSE	Test MSE
delta	251	0.00745	0.0085
gen_delta	364	0.01682	0.01792
adagrad	305	0.0178	0.01833
rmsprop	83	0.00491	0.00691
adadelta	324	0.01205	0.01323
adam	74	0.00473	0.00725

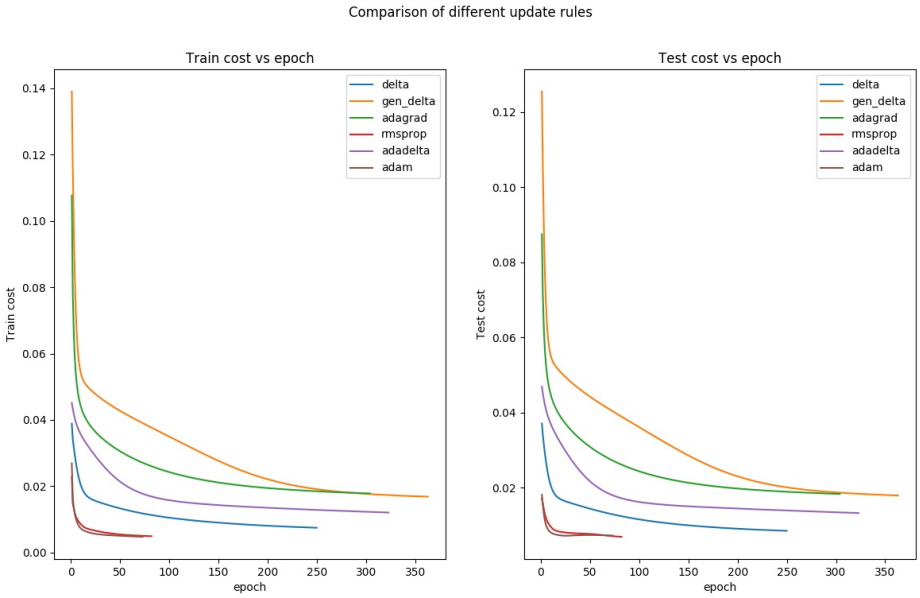


Figure 14: Effect of different optimizers on train and test MSE

Conclusion:

- From Figure 14. and Table 7. one can infer that both Adam and RMSProp optimizers outperform the other weight update rules.
- RMSProp takes into account the running average of of the previous L gradients which improves the weight updates.
- Adam, due to its effective usage of first order and second order moments performs better weight updates than other optimizers.
- When few more experiments were performed to break the tie, Adam turned out to be the best optimizer for this problem.

1.5 Best Configuration and Inferences

The best configuration for the given Function Approximation task are as follows:

- Number of Hidden Layers: 1
- Number of nodes in a hidden layer: 8
- Slope of logistic function: 3
- Type of input features: Normalized
- Activation Function: Tanh

- Learning mode: Pattern
- Weight Update Rule: Adam

The loss function used in the experiments is Sum of Squared Error since the given task is a regression task. Below are the scatter plots of the best configuration for train and test data, obtained from experiments conducted above.

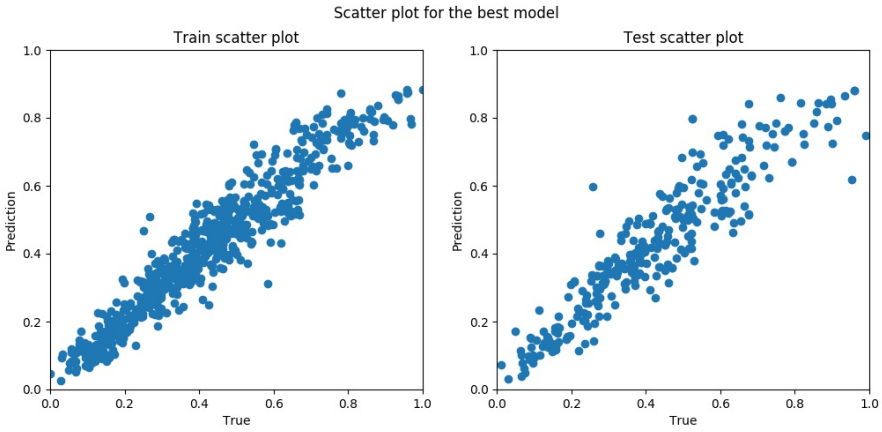


Figure 15: Scatter Plots for the best configuration

The train MSE and test MSE versus epochs for the best configuration is as follows.

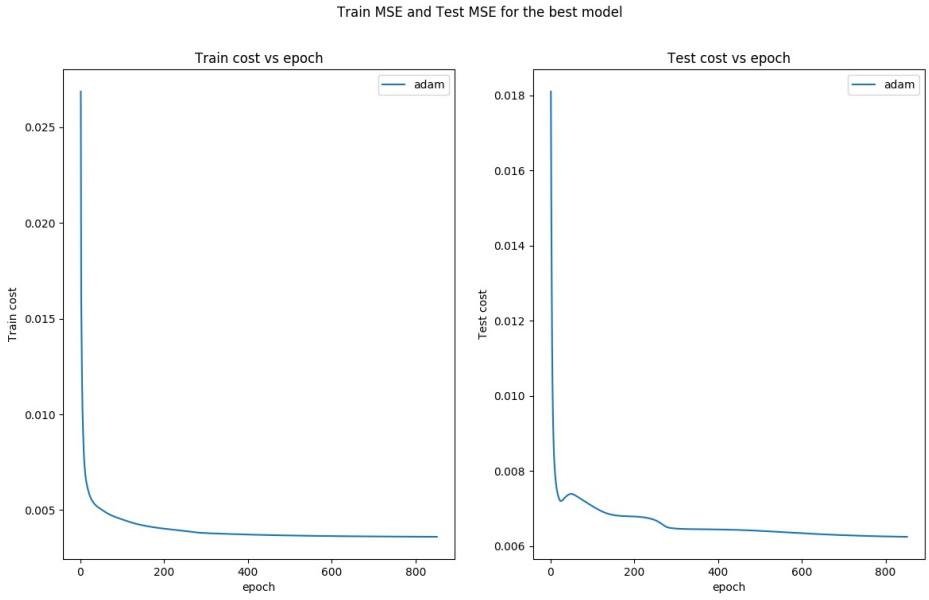


Figure 16: Train and Test MSE v/s epochs for the best configuration

The exact values for the above plot are given in the table below.

Table 8: Convergence criteria, Convergence Epoch and MSE for the Best Configuration.

Convergence threshold	Convergence_epoch	MSE Train	MSE Test
1e-7	852	0.00360	0.00624

2 SINGLE-LABEL MULTI-CLASS CLASSIFICATION

2.1 INTRODUCTION:

Baseline model was found out using Unnormalized features, Pattern mode, Cross entropy error,logistic function and delta rule as the primary configuration. Number of nodes and number of hidden layers were varied to find out the best architecture. Variation of various components on this architecture is studied by varying them independently one at a time.

2.2 BASELINE CONFIGURATION:

Features are unnormalized,Pattern mode, Cross entropy error, Logistic Function and Delta rule,Learning rate for Pattern mode is 0.001. Threshold for convergence is 0.000005

Table 9: Hidden Layers=1,Number of nodes=4,Number of Epochs=189

Training Confusion					Testing Confusion				
Coast	Forest	Highway	Street	Building	Coast	Forest	Highway	Street	Building
60%	9%	11%	3.6%	14%	43.8%	13.3%	15.2%	11.4%	16.1%
0.8%	91%	0%	4.8%	2.8%	1%	82%	0%	8.2%	8.2%
28.4%	3.1%	59%	6.7%	2.5%	31.1%	0%	52%	6.4%	10%
1%	8.4%	3.1%	82.1%	4.9%	0%	13.9%	1.2%	75%	8.8%
12.8%	17.2%	1.6%	13.6%	55%	10%	24.1%	5.8%	16.7%	43.4%

Table 10: Hidden Layers=1,Number of nodes=6,Number of Epochs=655

Training Confusion					Testing Confusion				
Coast	Forest	Highway	Street	Building	Coast	Forest	Highway	Street	Building
74%	6.8%	12.6%	2.1%	4.3%	57.2%	6.2%	16.7%	8.3%	11.4%
2%	89.4%	0%	6.9%	2.4%	4%	78%	0%	14.4%	3%
9%	0%	76.5%	7.9%	5.8%	16.8%	2.2%	64%	8.9%	7.8%
2.2%	1.3%	3.9%	88%	4.4%	5.4%	4.1%	1.3%	86.3%	2.7%
10.7%	10.3%	1.5%	14.7%	62.5%	15.4%	14.6%	8.9%	16.2%	44.7%

Table 11: Hidden Layers=1,Number of nodes=5,Number of Epochs=481

Training Confusion					Testing Confusion				
Coast	Forest	Highway	Street	Building	Coast	Forest	Highway	Street	Building
74.2%	2.9%	12.8%	1.1%	8%	56%	12%	18%	3%	11%
5%	85%	0%	3.2%	5.3%	7%	77%	0%	10%	6%
15.5%	1%	74.8%	2.6%	5.8%	25%	1.1%	57%	3.4%	13.6%
1.8%	1.8%	5.4%	79%	11.8	3.6%	8.5%	3.6%	75%	8.5%
15.7%	8.2%	4.8%	8.9%	62.2%	12%	12%	7.4%	12.9%	55.5%

Table 12: Hidden Layers=2,Number of nodes=5:5,Number of Epochs=158

Training Confusion					Testing Confusion				
Coast	Forest	Highway	Street	Building	Coast	Forest	Highway	Street	Building
82%	2.1%	0%	1.4%	14.3%	77.3%	5.1%	0%	3%	14.4%
1.2%	54.8%	0%	20.9%	23%	2.8%	51.8%	0%	24.5%	20.7%
93%	0%	0%	1.5%	4.7%	84%	1.1%	0%	1.1%	12.9%
6.7%	57.8%	0%	8%	27.3%	9%	67.5%	0%	7.7%	15.5%
36.5%	13.3%	0%	6.8%	43.5%	35.4%	18.5%	0%	8.8%	37.1%

Table 13: Hidden Layers=2,Number of nodes=5:8,Number of Epochs=171

Training Confusion					Testing Confusion				
Coast	Forest	Highway	Street	Building	Coast	Forest	Highway	Street	Building
78.7%	4.3%	0%	3.9%	12.9%	69.7%	7.2%	0%	1%	21.8%
0%	87%	0%	8.2%	4.1%	4.9%	79.2%	0%	12.8%	2.9%
87.8%	2.1%	0%	1%	8.9%	85.5%	5.5%	0%	1.1%	7.7%
3.9%	70%	0%	8.4%	38%	4%	71.6%	0%	8%	16%
36%	17%	0%	6.9%	38.3%	34.1%	22.2%	0%	10%	33.3%

Table 14: Hidden Layers=2,Number of nodes=4:11,Number of Epochs=145

Training Confusion					Testing Confusion				
Coast	Forest	Highway	Street	Building	Coast	Forest	Highway	Street	Building
91%	1.8%	0%	0%	6.8%	83.7%	3.6%	0%	0%	13.2%
2.9%	69%	0%	7.9%	20.4%	5.7%	69%	0%	4.7%	20.9%
96%	0%	0%	1.5%	1%	92%	1.1%	0%	0%	5.9%
17.3%	52.3%	0%	5%	25	11.1%	62.9%	0%	7.4%	18.5%
50.9%	9.8%	0%	1.8%	37.4%	46%	14.5%	0%	2.7%	37%

Table 15: Hidden Layers=2,Number of nodes=6:6,Number of Epochs=229

Training Confusion					Testing Confusion				
Coast	Forest	Highway	Street	Building	Coast	Forest	Highway	Street	Building
41%	6.4%	32%	4.2%	16%	39%	13.4%	21%	5.1%	21.6%
0%	94%	0%	4%	2.45%	2.1%	92%	0%	2.1%	4.2%
21.7%	0%	68.7%	6.1%	2.7%	15.6%	3.1%	70.8%	4.1%	6.2%
5.7%	20%	2.1%	58.7%	13.1	4%	13.5%	4%	77%	1.3%
17.8%	32%	3.1%	8.2%	39.4%	22.2%	29%	11.1%	5.9%	31.6%

Table 16: Hidden Layers=2,Number of nodes=7:6,Number of Epochs=198

Training Confusion					Testing Confusion				
Coast	Forest	Highway	Street	Building	Coast	Forest	Highway	Street	Building
91.2%	3.2%	1.8%	1%	2%	82.6%	9.1%	2%	3%	3%
7.2%	77%	0%	6%	9.3%	8.6%	75.2%	0%	7.5%	8.6%
77.8%	0%	16.4%	2.8%	2.8%	72.4%	3%	20%	3%	1%
4.3%	6.9%	0%	85.3%	3%	9.7%	4.1%	2.8%	79%	4.1%
35.7%	25.3%	1.57%	11.9%	25.4%	36%	22%	0%	28.2%	12%

Baseline is 5 Hidden layer nodes as it has overall better performance on test data and takes lesser number of epochs to converge than 6 hidden nodes,whereas 4 nodes is underfit model. Number of hidden layer is 1.

2.3 Experiments:

2.3.1 Varying Slope of Sigmoid:

Hidden layers=1, Number of nodes=5, Mode of learning= Pattern, Update Rule=Delta, Convergence Criteria=0.0005

Table 17: Hidden Layers=1,Number of nodes=5,Number of Epochs=314,Slope=0.2

Training Confusion					Testing Confusion				
Coast	Forest	Highway	Street	Building	Coast	Forest	Highway	Street	Building
68.3%	5.1%	11.7%	1.8%	12.8%	59%	11%	12%	7%	11%
2.8%	83.1%	0%	5.3%	8.6%	0%	84%	0%	5%	11%
24.5%	0.5%	64.7%	3.2%	6.9%	29.5%	1.1%	54.5%	2.2%	12.5%
0.9%	5.9%	3.1%	77.7%	12.2%	3.6%	3.6%	1.2%	86.5%	4.8%
13.8%	8.9%	4.1%	7.1%	65.9%	11.1%	19.4%	5.5%	8.3%	55.5%

Table 18: Hidden Layers=1,Number of nodes=5,Number of Epochs=150,Slope=1

Training Confusion					Testing Confusion				
Coast	Forest	Highway	Street	Building	Coast	Forest	Highway	Street	Building
59.9%	6.2%	11%	1.4%	22%	52%	14%	11%	4%	19%
0.8%	87.6%	0%	2.4%	9%	0%	80%	0%	5%	15%
21%	1%	65.7%	3.7%	8%	17%	1.1%	59%	2.2%	20%
0.9%	5%	5%	77.7%	11.3%	3.6%	9.7%	1.2%	78%	7.3%
7.4%	8.2%	4.1%	7.1%	73%	11.1%	16.7%	4.6%	11.1%	56.5%

Table 19: Hidden Layers=1,Number of nodes=5,Number of Epochs=8,Slope=10

Training Confusion					Testing Confusion				
Coast	Forest	Highway	Street	Building	Coast	Forest	Highway	Street	Building
24.6%	0%	0%	0%	75.3%	23%	0%	0%	0%	77%
0%	0%	0%	0%	100%	1%	0%	0%	0%	99%
32%	0%	0%	0%	67.9%	21.5%	0%	0%	0%	78.4%
7.7%	0%	0%	0%	92.3	6%	0%	0%	0%	93.9%
7.1%	0%	0%	0%	92.8%	5.5%	0%	0%	0%	94.4%

OBSERVATION AND INFERENCES:

- Slope 10 saturates very fast leading to fast convergence and poor accuracy.
- Slope 0.2 and 1 have almost same accuracies but 1 takes lesser number of epochs for convergence,hence 1 is best slope.

2.3.2 VARYING MODE OF LEARNING:

Hidden layers=1, Number of nodes=5, Activation=Logistic, Features=Unnormalized, Update Rule= Delta, Slope=1,Convergence Criteria=0.000005, Learning Rate=0.004

Table 20: Pattern Mode,Number of Epochs=45

Training Confusion					Testing Confusion				
Coast	Forest	Highway	Street	Building	Coast	Forest	Highway	Street	Building
56.2%	18.3%	8.4%	0.3%	16.5%	47%	17%	15%	4%	17%
0.4%	95%	0%	1.2%	3.2%	0%	92%	0%	4%	4%
28.3%	4.8%	53.5%	5.3%	8%	28.4%	9%	47.7%	3.4%	11.36%
1.3%	14%	2.2%	75%	7.2%	2.4%	15.8%	2.4%	74.4%	4.8%
7.5%	26.2%	4.5%	7.5%	54.3%	12.9%	37%	1.8%	10%	37.9%

Table 21: Batch Mode,Number of Epochs=5001

Training Confusion					Testing Confusion				
Coast	Forest	Highway	Street	Building	Coast	Forest	Highway	Street	Building
80.9%	9.1%	0%	0%	9.9%	77%	9%	0%	0%	14%
4.5%	84.7%	0%	1.6%	9%	1%	86%	0%	1%	12%
86%	5.8%	0%	0.5%	7.4%	84%	4.5%	0%	1.1%	10.2%
13.6%	27.27%	0%	36.8%	22.27%	15.8%	21.9%	0%	41.4%	20.7%
26.6%	25.5%	0%	3.3%	44.5%	34.3%	34.3%	0%	0%	31.5%

OBSERVATION AND INFERENCES:

- Pattern mode turns out to be better than Batch mode both in terms of Number of Epochs as well as accuracy after convergence.
- Learning rate for this experiment alone is higher than others because bath mode was taking very large number of epochs for convergence and if the convergence criteria was relaxed then it converged very fast with poor accuracy due to very small changes in the loss.

2.3.3 VARYING UPDATE RULE:

Hidden layers=1, Number of nodes=5, Activation=Logistic, Features=Unnormalized, Slope=1, Mode of learning= Pattern,Learning Rate=0.001, Threshold=0.000005

Table 22: Delta Rule,Number of Epochs=481

Training Confusion					Testing Confusion				
Coast	Forest	Highway	Street	Building	Coast	Forest	Highway	Street	Building
74.2%	2.9%	12.8%	1.1%	8%	56%	12%	18%	3%	11%
5%	85%	0%	3.2%	5.3%	7%	77%	0%	10%	6%
15.5%	1%	74.8%	2.6%	5.8%	25%	1.1%	57%	3.4%	13.6%
1.8%	1.8%	5.4%	79%	11.8	3.6%	8.5%	3.6%	75%	8.5%
15.7%	8.2%	4.8%	8.9%	62.2%	12%	12%	7.4%	12.9%	55.5%

Table 23: Generalized Delta Rule,Number of Epochs=209,Alpha=0.2

Training Confusion					Testing Confusion				
Coast	Forest	Highway	Street	Building	Coast	Forest	Highway	Street	Building
63.9%	3.7%	6.7%	2.6%	22.9%	53.4%	8.9%	6.9%	1%	29.7%
1.6%	85.5%	0%	0%	12.8%	0%	73.9%	0%	1.04%	25%
38.9%	4.3%	45.4%	3.2%	8.1%	43.8%	1.1%	44.9%	2.24%	7.8%
1.4%	12.7%	2.3%	71%	12.3%	1.1%	12%	4.4%	67%	15.4%
12.8%	12.5%	0%	2.6%	71.9%	12.8%	17.8%	2.9%	8.9%	57.4%

Table 24: Adagrad Rule,Number of Epochs=2000

Training Confusion					Testing Confusion				
Coast	Forest	Highway	Street	Building	Coast	Forest	Highway	Street	Building
73.6%	3.3%	11.2%	2.2%	9.3%	64.3%	12.8%	8.9%	1.9%	11.8%
2.4%	86.3%	0%	4.5%	6.6%	5.2%	76%	0%	5.2%	613.5
28.6%	1%	57.8%	5.9%	6.4%	33%	3.3%	53.9%	1.1%	7.8%
0.9%	8.5%	3.3%	79.6%	7.5%	2.1%	6.5%	4.3%	76.9%	9.8%
15.1%	12.1%	3.7%	3.4%	65.5%	12.8%	16.8%	8.9%	11.9%	49.5%

Table 25: RMS Propagation Rule,Number of Epochs=24,Epsilon=0.00001

Training Confusion					Testing Confusion				
Coast	Forest	Highway	Street	Building	Coast	Forest	Highway	Street	Building
61.7%	7.7%	18.7%	2.2%	9.5%	53%	9%	25%	5%	8%
3.7%	81.9%	0.8%	5.7%	7.8%	1%	77%	0%	8%	14%
25.1%	3.7%	62.6%	2.1%	6.4%	22.7%	2.2%	63.6%	4.5%	6.8%
2.2%	10.4%	4.5%	74.5%	8.1%	1.2%	12.19%	4.8%	70.7%	10%
14.2%	8.9%	4.8%	10.1%	61.8%	11.1%	19.4%	12%	11.1%	46.3%

Table 26: Adadelta Rule ,Number of Epochs=40,ro=0.9,epsilon=0.0001

Training Confusion					Testing Confusion				
Coast	Forest	Highway	Street	Building	Coast	Forest	Highway	Street	Building
66.1%	6.9%	13.2%	1.4%	12.1%	54%	12%	14%	7%	13%
3.7%	83.5%	0%	2.8%	9.8%	4%	75%	0%	8%	13%
26.2%	1.6%	62%	3.2%	6.9%	28.4%	3.4%	53.4%	5.6%	9%
2.7%	10%	3.6%	70.9%	12.7%	1.2%	7.3%	4.8%	74.3%	12.1%
16.1%	4.8%	1.1%	7.4%	70.4%	17.6%	19.4%	4.6%	9.2%	49%

Table 27: Adam Rule,Number of Epochs=284,ro1=0.92,ro2=0.99999

Training Confusion					Testing Confusion				
Coast	Forest	Highway	Street	Building	Coast	Forest	Highway	Street	Building
68.9%	5%	17.9%	1.1%	7%	49.6%	6.8%	32.4%	3.4%	7.6%
4.4%	84%	0%	3.2%	8.4%	7.4%	69.1%	3.1%	8.5%	11.7%
3.5%	0%	87.4%	1%	8%	20%	1.4%	60%	8.5%	10%
0.9%	1.9%	5.2%	83.6%	8.1%	4%	3%	5%	77.7%	10%
8.3%	5.6%	4.1%	3.4%	78.4%	12.2%	12.2%	18.4%	4%	53%

OBSERVATION AND INFERENCES:

- Adam turns out to be the best update rule due to use of first and second order moments.
- Generalized Delta is better than Delta due to momentum factor which prevents large number of oscillations.
- Adagrad gives better accuracy at the cost of very large number of epochs as the gradients saturate after some point.
- RMS Propagation gives fast results but the accuracy is comparatively low.
- Adadelta gives better accuracy than RMS Prop with approximately same number of epochs.

2.3.4 Varying Activation Function:

Hidden layers=1, Number of nodes=5, Features=Unnormalized, Slope=1, Mode of learning= Pattern, Update Rule=Delta Rule

Table 28: Logistic,Number of Epochs=481

Training Confusion					Testing Confusion				
Coast	Forest	Highway	Street	Building	Coast	Forest	Highway	Street	Building
74.2%	2.9%	12.8%	1.1%	8%	56%	12%	18%	3%	11%
5%	85%	0%	3.2%	5.3%	7%	77%	0%	10%	6%
15.5%	1%	74.8%	2.6%	5.8%	25%	1.1%	57%	3.4%	13.6%
1.8%	1.8%	5.4%	79%	11.8	3.6%	8.5%	3.6%	75%	8.5%
15.7%	8.2%	4.8%	8.9%	62.2%	12%	12%	7.4%	12.9%	55.5%

Table 29: Tanh,Number of Epochs=351,Slope=0.5

Training Confusion					Testing Confusion				
Coast	Forest	Highway	Street	Building	Coast	Forest	Highway	Street	Building
70.9%	3.3%	15.8%	0.7%	9.1%	63%	12%	12%	5%	8%
4.1%	88.5%	0%	2.4%	4.9%	5%	81%	0%	2%	12%
21.4%	0%	72.1%	2.1%	4.2%	25%	2.2%	62.5%	2.2%	7.9%
1.8%	4%	6.8%	72.3%	15%	2.4%	7.3%	2.4%	70.7%	17%
12.7%	7.4%	10.1%	3.3%	66.3%	12%	13.9%	17.6%	8.3%	48.1%

Table 30: Relu,Number of Epochs=119

Training Confusion					Testing Confusion				
Coast	Forest	Highway	Street	Building	Coast	Forest	Highway	Street	Building
52.2%	8%	18.4%	1.8%	19.5%	42%	9%	22%	8%	19%
1.2%	87.6%	0%	5.3%	5.7%	2%	87%	0%	6%	5%
10.7%	0%	69%	8.5%	11.7%	12.5%	1.1%	65.9%	6.8%	13.6%
0.4%	4%	4%	84.5%	6.8%	1.2%	3.6%	3.6%	86.5%	4.8%
14.2%	24.3%	8.9%	22.4%	29.9%	6.4%	29.6%	13.9%	16.7%	33.3%

Table 31: Elu,Number of Epochs=148,Delta=0.7

Training Confusion					Testing Confusion				
Coast	Forest	Highway	Street	Building	Coast	Forest	Highway	Street	Building
53.6%	6.6%	21.7%	1.8%	16.1%	46%	12%	25%	6%	11%
2%	81.8%	0%	7.8%	8.2%	1%	78%	0%	7%	14%
15.5%	0%	73.8%	5.8%	4.8%	6.8%	4.5%	70.5%	5.6%	12.5%
0.4%	4%	4%	81.8%	9.5%	2.4%	4.8%	6%	79.2%	7.3%

Coast	Forest	Highway	Street	Building	Coast	Forest	Highway	Street	Building
9.7%	7.1%	7.4%	10.1%	65.5%	3.7%	12%	12%	12.9%	59.2%

Table 32: Softplus,Number of Epochs=588

Training Confusion					Testing Confusion				
Coast	Forest	Highway	Street	Building	Coast	Forest	Highway	Street	Building
68.4%	4%	12.5%	0.7%	14.3%	46%	8%	23%	5%	18%
4.9%	81.9%	0%	2.8%	10.2%	2%	81%	0%	4%	13%
16.6%	0%	68.4%	5.3%	9.6%	26.1%	3.4%	55.7%	7.9%	6.8%
0.4%	3.6%	2.2%	76.8%	16.8%	2.4%	6%	2.4%	75.6%	13.4%
11.2%	4.1%	8.2%	5.6%	70.8%	6.5%	14.8%	10.2%	7.4%	61.1%

OBSERVATION AND INFERENCES:

- Tanh gives better accuracy than logistic function with lesser number of epochs for convergence.
- Relu gives almost similar accuracy with much lesser number of epochs.
- ELU takes lesser epochs than sigmoid with better test accuracy.
- Softplus takes a little more number of epochs but gives better accuracy than others.

2.3.5 VARYING FEATURE NORMALIZATION:

Hidden layers=1, Number of nodes=5, Slope=1, Mode of learning= Pattern, Update Rule=Delta, Activation = Logistic, Learning Rate=0.004,Threshold=0.000005

Table 33: Number of Epochs=45,Unnormalized Features

Training Confusion					Testing Confusion				
Coast	Forest	Highway	Street	Building	Coast	Forest	Highway	Street	Building
56.2%	18.4%	8.4%	0.4%	16.5%	47%	17%	15%	4%	17%
0.4%	95%	0%	1.2%	3.2%	0%	92%	0%	4%	4%
28.3%	4.8%	53.4%	5.3%	8%	28.4%	9%	47.7%	3.4%	11%
1.3%	14.1%	2.2%	75%	7.2%	2.4%	15.8%	2.4%	74.4%	4.8%
7.5%	26.2%	4.4%	7.5%	54.3%	12.9%	37%	1.8%	10.1%	37.9%

Table 34: Number of Epochs=1000,Normalized Features

Training Confusion					Testing Confusion				
Coast	Forest	Highway	Street	Building	Coast	Forest	Highway	Street	Building
73.5%	6.6%	12.5%	1.1%	6.2%	66%	10%	15%	3%	6%

Coast	Forest	Highway	Street	Building	Coast	Forest	Highway	Street	Building
5.7%	87.2%	0%	2.4%	4.5%	3%	92%	0%	0%	5%
25.1%	0%	64.1%	5.8%	4.8%	26.1%	0%	56.8%	3.4%	13.6%
1.8%	0.4%	2.7%	85%	10%	0%	2.4%	2.4%	87.8%	7.3%
11.6%	6.7%	5.2%	10.4%	65.9%	12%	9.2%	8.3%	11.1%	59.2%

OBSERVATION AND INFERENCES:

- Normalized gives better accuracy than unnormalized features at the cost of more epochs for convergence.
- This is due to saturation of gradients in sigmoid function.

3 SINGLE-LABEL MULTI-CLASS CLASSIFICATION WITH SUM OF SQUARED LOSS

3.1 INTRODUCTION:

Baseline model was found out using Unnormalized features, Pattern mode, logistic function and delta rule as the primary configuration. Number of nodes and number of hidden layers were varied to find out the best architecture. Variation of various components on this architecture is studied by varying them independently one at a time.

3.2 BASELINE CONFIGURATION:

Learning rate for Pattern mode is 0.001. Threshold for convergence is 0.000005

Table 35: Hidden Layers=1,Number of nodes=8,Number of Epochs=277

Training Confusion					Testing Confusion				
Coast	Forest	Highway	Street	Building	Coast	Forest	Highway	Street	Building
82.2%	8.3%	3.6%	1.4%	4.3%	77.3%	3.5%	5.9%	4.7%	8.3%
3.7%	89.8%	0.4%	1.8%	4.1%	5.3%	91%	0%	0.9%	2.6%
46.2%	2.2%	44%	3.4%	4%	45.8%	2.3%	36.5%	3.5%	11%
2.8%	12.6%	2.3%	75.11%	7%	11.4%	12.6%	2.5%	62%	11.4%
21.8%	21.8%	3.7%	7.5%	44.9%	22%	27.9%	5%	1.6%	43.2%

Table 36: Hidden Layers=1,Number of nodes=6,Number of Epochs=209

Training Confusion					Testing Confusion				
Coast	Forest	Highway	Street	Building	Coast	Forest	Highway	Street	Building
55.9%	18.6%	21.4%	3.9%	0%	54.6%	16.6%	25.9%	2.7%	0%
3.3%	86%	0%	10%	0%	4.3%	88%	0%	7%	0%
21.5%	3.1%	65.7%	9.5%	0%	24.2%	7.1%	60%	8.5%	0%
2.9%	8.8%	4.9%	83.3%	0%	4.5%	17%	0%	78.4%	0%
25.8%	51.2%	7.6%	14.8%	0.4%	24.1%	48.3%	10.3%	16.6%	0%

Table 37: Hidden Layers=1,Number of nodes=7,Number of Epochs=945

Training Confusion					Testing Confusion				
Coast	Forest	Highway	Street	Building	Coast	Forest	Highway	Street	Building
73.2%	5%	8.5%	3.1%	10%	61.7%	3.9%	16.7%	0.9%	16.7%
1.8%	86.9%	0%	2.3%	8.8%	2.6%	84.2%	0%	5.2%	7.8%
17.3%	2.3%	68.8%	2.8%	8.6%	31%	1.1%	47.1%	9.1%	11.5%
1.9%	5.2%	1.9%	82.3%	8.6%	4.8%	12%	4.8%	62.6%	15.6%
6.8%	8.7%	3.4%	2.6%	78.4%	17.4%	9.7%	7.6%	7.6%	57.6%

Table 38: Hidden Layers=2,Number of nodes=7:6,Number of Epochs=43

Training Confusion					Testing Confusion				
Coast	Forest	Highway	Street	Building	Coast	Forest	Highway	Street	Building
100%	0%	0%	0%	0%	100%	0%	0%	0%	0%
100%	0%	0%	0%	0%	100%	0%	0%	0%	0%
100%	0%	0%	0%	0%	100%	0%	0%	0%	0%
100%	0%	0%	0%	0%	100%	0%	0%	0%	0%
100%	0%	0%	0%	0%	100%	0%	0%	0%	0%

Table 39: Hidden Layers=2,Number of nodes=7:7,Number of Epochs=772

Training Confusion					Testing Confusion				
Coast	Forest	Highway	Street	Building	Coast	Forest	Highway	Street	Building
79.5%	5.7%	6.4%	1.4%	6.8%	75.6%	3.6%	4.8%	7.3%	8.5%
2.8%	84.7%	0%	4.7%	7.6%	9.2%	79.8%	0%	5.8%	5%
37.2%	0.5%	52.3%	6.4%	3.4%	50%	3.4%	31.8%	6.8%	7.9%
1.8%	8.9%	1.8%	81.2%	6.1%	5%	15.2%	5%	69.6%	5%
18.7%	29.7%	2.4%	6.5%	42.7%	17.2%	30.9%	2.7%	9%	40%

Table 40: Hidden Layers=2,Number of nodes=7:8,Number of Epochs=410

Training Confusion					Testing Confusion				
Coast	Forest	Highway	Street	Building	Coast	Forest	Highway	Street	Building
96.7%	0%	0%	3.2%	0%	96.5%	0%	0%	3.5%	0%
89.7%	0%	0%	10.3%	0%	90.6%	0%	0%	9.4%	0%
18.3%	0%	0%	81.7%	0%	21.5%	0%	0%	78.4%	0%
92.1%	0%	0%	7.8%	0%	91.2%	0%	0%	8.7%	0%

Table 41: Hidden Layers=2,Number of nodes=6:6,Number of Epochs=906

Training Confusion					Testing Confusion				
Coast	Forest	Highway	Street	Building	Coast	Forest	Highway	Street	Building
72.6%	4.7%	9.1%	2.7%	10.7%	62.9%	6.5%	12.9%	4.6%	12.9%
0.8%	86.6%	0.4%	2.6%	9.3%	3.8%	79.8%	0.9%	5.7%	9.6%
32%	0%	62.5%	3.2%	2.1%	48.7%	3.9%	35.5%	9.2%	2.6%
2.9%	9.7%	3.9%	79%	4.3%	4.5%	14.9%	5.7%	70.1%	4.6%
24.9%	18.9%	5.1%	3.1%	47.8%	28.1%	26.2%	3.8%	9.7%	32%

Table 42: Hidden Layers=2,Number of nodes=6:8,Number of Epochs=435

Training Confusion					Testing Confusion				
Coast	Forest	Highway	Street	Building	Coast	Forest	Highway	Street	Building
60.8%	0%	0%	5%	34.1%	62.7%	0%	0%	5.8%	31.3%
40.1%	0%	0%	21%	38.8%	36.8%	0%	0%	18.4%	44.7%
64.1%	0%	0%	16.1%	19.6%	65.5%	0%	0%	11.5%	22.9%
8.1%	0%	0%	85.1%	6.6%	9.6%	0%	0%	77.1%	13.2%
41.6%	0%	0%	10.9%	47.3%	41.3%	0%	0%	17.4%	41.3%

Table 43: Hidden Layers=2,Number of nodes=5:6,Number of Epochs=314

Training Confusion					Testing Confusion				
Coast	Forest	Highway	Street	Building	Coast	Forest	Highway	Street	Building
95.2%	0%	0%	4.8%	0%	98.1%	0%	0%	1.8%	0%
90.2%	0%	0%	9.7%	0%	93.5%	0%	0%	6.5%	0%
89%	0%	0%	10.9%	0%	89.7%	0%	0%	10%	0%
20.7%	0%	0%	79.2%	0%	17.6%	0%	0%	82.3%	0%
89.7%	0%	0%	10.3%	0%	84.5%	0%	0%	15.4%	0%

Table 44: Hidden Layers=2,Number of nodes=5:8,Number of Epochs=1219

Training Confusion					Testing Confusion				
Coast	Forest	Highway	Street	Building	Coast	Forest	Highway	Street	Building
49%	6.7%	12.2%	3.5%	28.4%	43.9%	8.4%	16.8%	2.8%	28%
0.8%	90.2%	0%	2.5%	6.3%	0%	85.8%	1%	4.3%	8.6%
21.3%	1%	68.2%	5.7%	3.6%	22.4%	2.9%	50%	13.2%	11.7%
3.4%	8.4%	6.4%	75.6%	5.9%	2.1%	18.7%	3.2%	69.2%	6.5%
14.8%	23.3%	3.3%	2.9%	55.5%	21.6%	21.6%	7.5%	5.8%	43.3%

Baseline is 1 hidden layer with 7 nodes as it gives better test accuracy as well as train accuracy than other architectures. Some architectures having 2 hidden layers give good accuracy at the cost of higher number of parameters and more overfitting.

4 MULTI-LABEL CLASSIFICATION

4.1 INTRODUCTION:

The baseline model was found using Unnormalized features, Pattern mode of learning, Sum of Squares Error and Adam rule as the primary configuration. Number of nodes and number of hidden layers were varied to find out the best architecture. Variations of the various components are studied by varying them independently.

4.2 BASELINE CONFIGURATION:

Features are unnormalized, Pattern mode is used, Sum of Squares Error, Logistic Function,Beta=1 and Adam rule are used.Learning rate is 0.0001 and threshold for convergence is 0.00003.

Table 45: Hidden Layers=1, Beta=1

		Train			Test		
Epochs	Number of Nodes	Precision	Recall	F-measure	Precision	Recall	F-measure
74	10	0.7263	0.5706	0.6391	0.7203	0.5850	0.6456
54	25	0.7329	0.5826	0.6492	0.7380	0.5856	0.6530
34	50	0.7279	0.5687	0.6385	0.7261	0.5624	0.6339

Table 46: Hidden Layers=2, Beta=1

		Train			Test		
Epochs	Number of Nodes	Precision	Recall	F-measure	Precision	Recall	F-measure
67	10:4	0.7177	0.5707	0.6358	0.7108	0.5811	0.6395
72	10:8	0.7162	0.5661	0.6323	0.7154	0.5720	0.6357
68	10:10	0.7157	0.5712	0.6353	0.7210	0.5806	0.6332
75	25:4	0.7163	0.5810	0.6416	0.7291	0.5820	0.6473
62	25:8	0.7232	0.5748	0.6405	0.7303	0.5774	0.6449
50	25:10	0.7169	0.5832	0.6432	0.7193	0.5775	0.6406
50	50:4	0.7123	0.5635	0.6292	0.7297	0.5689	0.6394
47	50:8	0.7142	0.5881	0.6451	0.7059	0.5830	0.6386
50	50:10	0.7161	0.5860	0.6446	0.7209	0.5829	0.6446

The baseline chosen is 1 hidden layer with 25 nodes in the layer as it performs better on the test data.

4.3 EXPERIMENTS:

4.3.1 VARYING SLOPE OF SIGMOID:

Hidden layers=1, Number of Nodes=25, Update rule is Adam, Pattern mode of learning

Table 47:

		Train			Test		
Epochs	Beta	Precision	Recall	F-measure	Precision	Recall	F-measure
197	0.2	0.7263	0.5519	0.6272	0.7397	0.5581	0.6362
54	1	0.7329	0.5826	0.6492	0.7380	0.5856	0.6530
19	10	0.7318	0.5964	0.6572	0.7345	0.5951	0.6575

OBSERVATIONS AND INFERENCE:

Higher the value of beta, fewer is the number of epochs for convergence.

4.4 VARYING MODE OF LEARNING:

Hidden layers=1, Number of nodes=25, Activation=Logistic, Features=Unnormalised, Update rule=Adam, Beta=1, Learning rate=0.0001, Convergence Criteria=0.00003

Table 48:

		Train			Test		
Learning Mode	Epochs	Precision	Recall	F-measure	Precision	Recall	F-measure
Pattern	54	0.7329	0.5826	0.6492	0.7380	0.5856	0.6530
Batch	464	0.7470	0.6354	0.6867	0.7307	0.6111	0.6655

OBSERVATIONS AND INFERENCE:

The batch mode is much slower than pattern mode because of using the average error in the update step.

4.5 VARYING UPDATE RULE:

Hidden Layers=1, Number of Nodes=25, Acivation=Logistic, Features are unnormalised, Beta=1, Pattern mode of learning,Learning rate=0.0001, Threshold=0.00003

Table 49:

		Train			Test		
Rule	Epochs	Precision	Recall	F-measure	Precision	Recall	F-measure
Delta	31	0.5879	0.5885	0.5882	0.5780	0.5815	0.5798
Generalised Delta	30	0.5873	0.5900	0.5886	0.5777	0.5839	0.5808
AdaGrad	329	0.6199	0.5316	0.5724	0.6086	0.5254	0.5640
RmsProp	18	0.6257	0.5493	0.5850	0.6159	0.5421	0.5766
AdaDelta	63	0.6665	0.4975	0.5697	0.6566	0.4895	0.5609
Adam	54	0.7329	0.5826	0.6492	0.7380	0.5856	0.6530

OBSERVATIONS AND INFERENCE:

The precision,recall increase as we use adagrad, rmsprop, adadelta and adam. However, no such pattern is observed for number of epochs for convergence.

4.6 VARYING ACTIVATION FUNCTION:

Hidden layers=1, Number of Nodes=25, Features=Unnormalised, Beta=1, Pattern mode of learning, Adam update rule

Table 50:

		Train			Test		
Function	Epochs	Precision	Recall	F-measure	Precision	Recall	F-measure
Logistic	54	0.7329	0.5826	0.6492	0.7380	0.5856	0.6530
Tanh	37	0.7206	0.5885	0.6479	0.7286	0.5883	0.6510
Relu	102	0.7342	0.6200	0.6723	0.7335	0.6216	0.6729
Softplus	42	0.7240	0.5783	0.6430	0.7302	0.5833	0.6486
Elu	111	0.7354	0.6193	0.6724	0.7353	0.6197	0.6726

OBSERVATIONS AND INFERENCE:

The precision and recall increase as we move from logistic to Elu function in the table.

4.7 VARYING FEATURE NORMALISATION:

Hidden layers=1, Number of Nodes=25, Beta=1, Pattern mode, Adam update rule, Activation=Logistic,Learning rate=0.0001

Table 51:

		Train			Test		
Features	Epochs	Precision	Recall	F-measure	Precision	Recall	F-measure
Unnormalised	54	0.7329	0.5826	0.6492	0.7380	0.5856	0.6530
Normalised	43	0.7323	0.5683	0.6400	0.7345	0.5712	0.6426

OBSERVATIONS AND INFERENCE:

Normalised data converges in fewer epochs than the unnormalised case.However, there isn't much change in the precision and recall of the model.