ASSIGNMENT-3

UDAYIN BISWAS, 2014CS50296

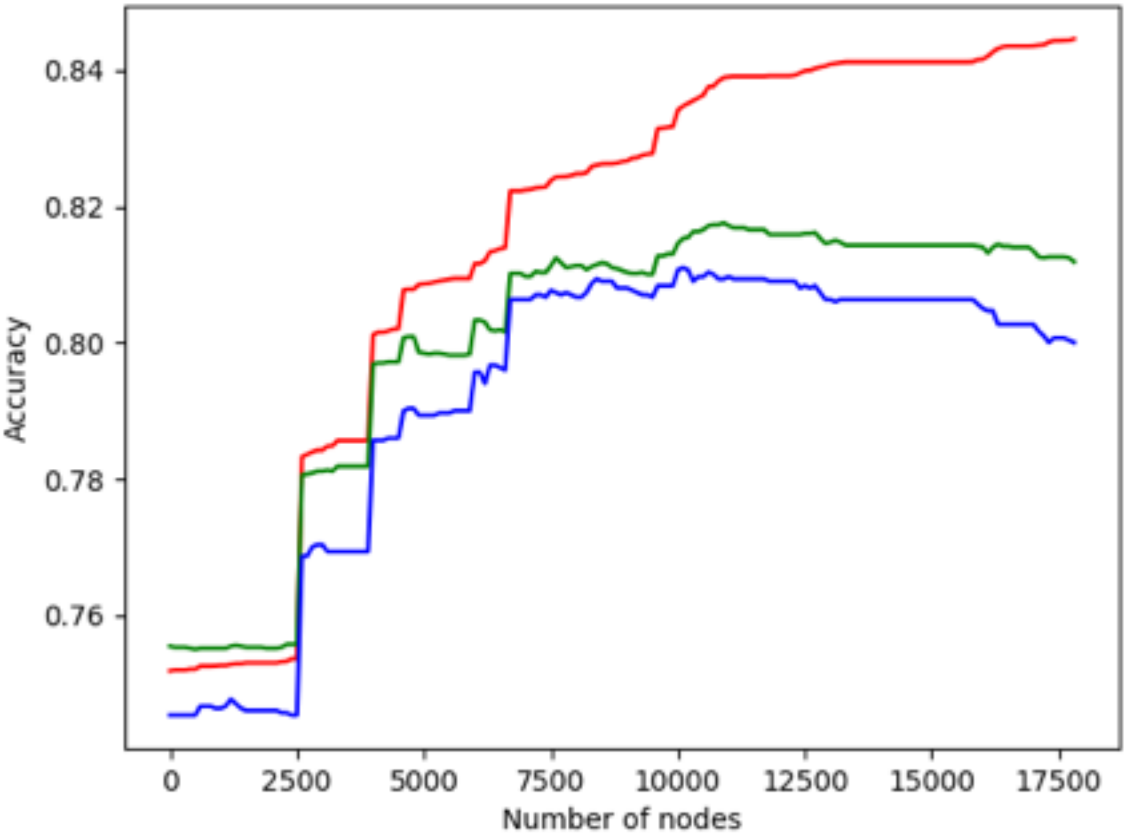
QUESTION 1:

(a) Building the decision tree and checking accuracies:

Decision tree built by choosing the attribute with the maximum information gain for splitting. Multiway split for discrete attributes and median conversion of numerical attributes as per statement.

Formula for splitting: Choose attribute which maximises:

I(X,Y)= H(Y)- [For all splits on X,(P(Xj)*H(Y|Xj))]



Training accuracy : between 77%-85%

Test accuracy: between 77%-81%

Validation Set accuracy: between 73%-80%

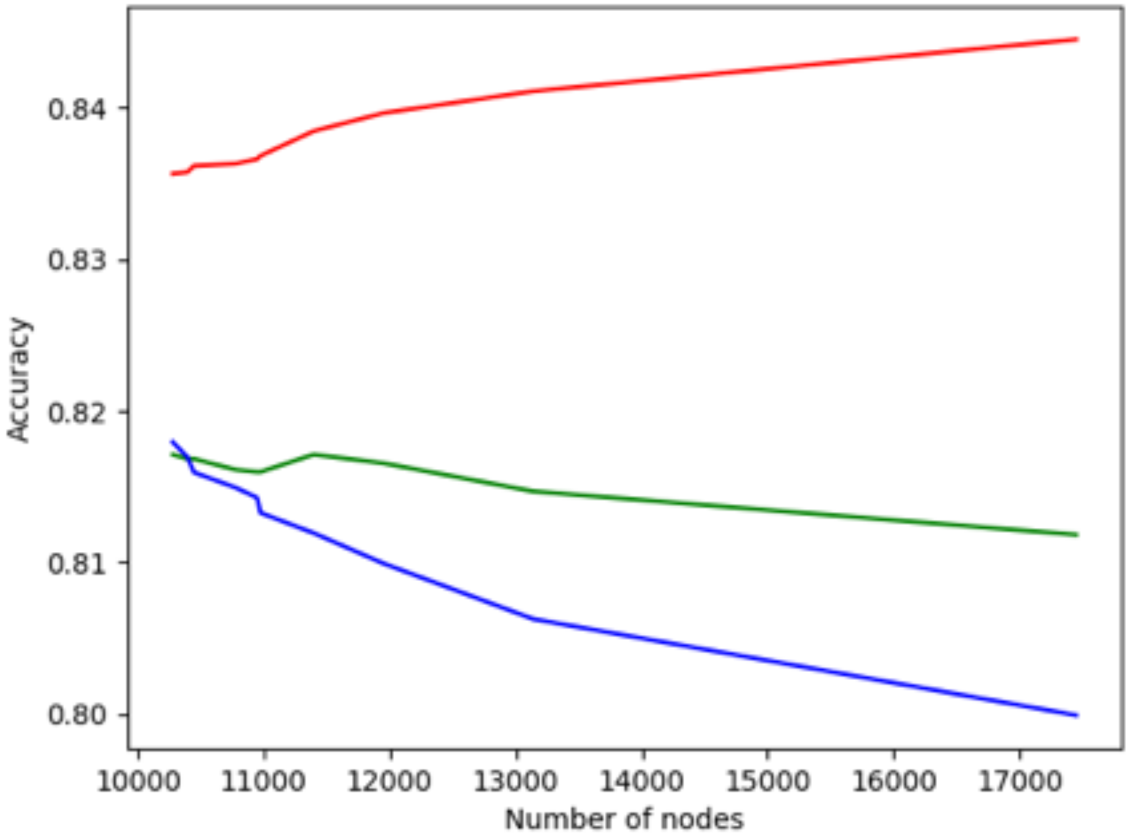
Observations:

(1) Training accuracy, test accuracy and validation set accuracy tend to gradually increase but at some points with a given number of nodes, they all peak together indicating that adding those extra nodes helped to improve accuracy.

(2) Training accuracy is the most as expected, test accuracy intermediate and validation set accuracy the least.

(3) After adding many nodes, the change in accuracy makes no effect and test accuracy and validation set accuracy decrease a bit.

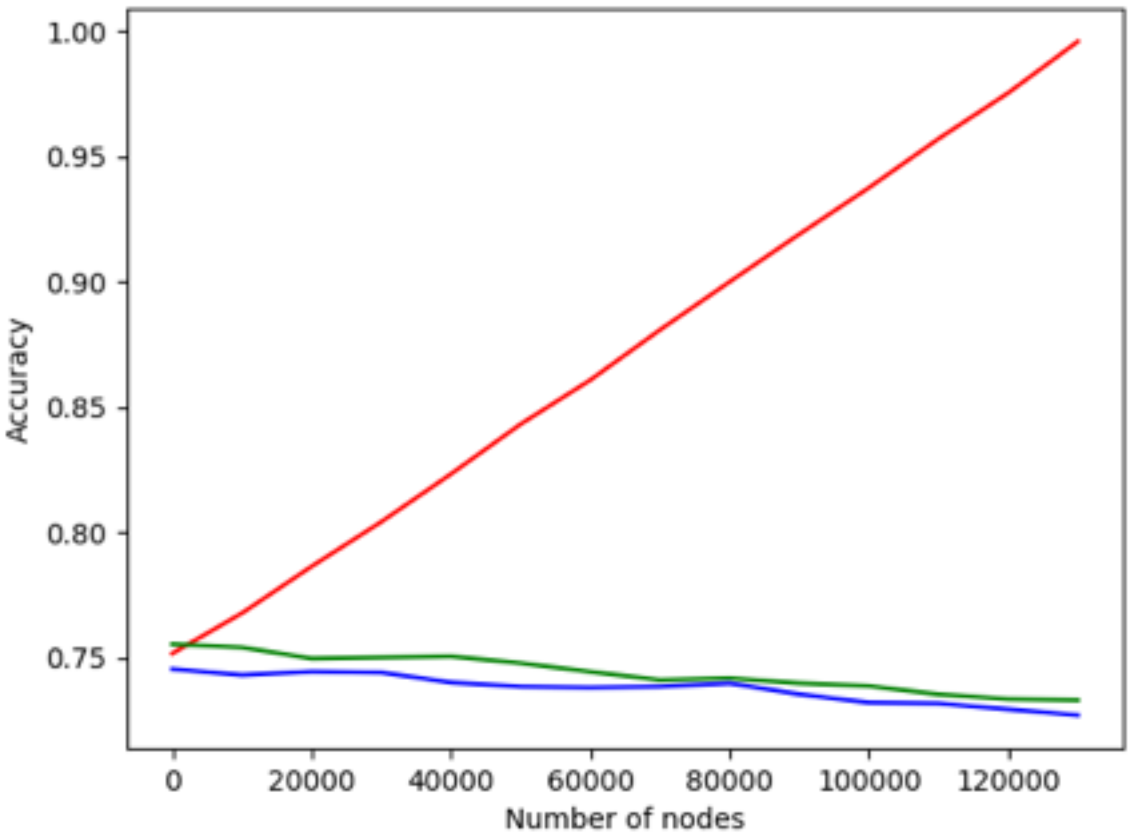
(b) Pruning and improving the accuracy:
Pruning the node and its subtree, which would help us give the best accuracy.
Initial validation set accuracy : 79.33%



Training accuracy : 84.4% at 17448 nodes -> 83.56% at 10280 nodes
Test accuracy: 81.18% at 17448 nodes -> 81.71% at 10280 nodes
Validation Set accuracy: 79.99% at 17448 nodes to 81.79% at 10280 nodes
(The plot is smooth because pruning was taking a lot of time and had to limit the number of iterations)

(c) Splitting on numerical attributes:

Splitting on numerical attributes : Fnlwgt, Education Number, Capital Gain, Capital Loss, Hour per Week.
Observations:
Training accuracy improves in a linear manner till around 100%. But this improvement in training accuracy is accompanied by decrease in testing accuracy and validation set accuracy.
Comparing to (a) part, it is better to have (a) part's decision tree as a model because it shows improvement in testing accuracy which is more important than the training accuracy improvement in (c) part.



Training accuracy : between 77%-85%
Test accuracy: between 77%-81%
Validation Set accuracy: between 73%-80%

(d) Validation set accuracy is best when max depth is allowed until 8, min samples leaf is the least among those tried i.e 1 so not setting min sample leaf would be better and min sample split is set to 3.
clf = DecisionTreeClassifier(random_state=0,min_samples_split=3,max_depth=8)

Validation set score= 92.2%
Training score=92.22%
Test score=92.21%

(e) Validation set accuracy is best when n_estimators is 15 or more, max features is 1 and bootstrap is True.
clf = RandomForestClassifier(random_state=0,bootstrap=True,n_estimators=15)

Validation set score= 91.03%
Training score=96.00%
Test score=91.17%

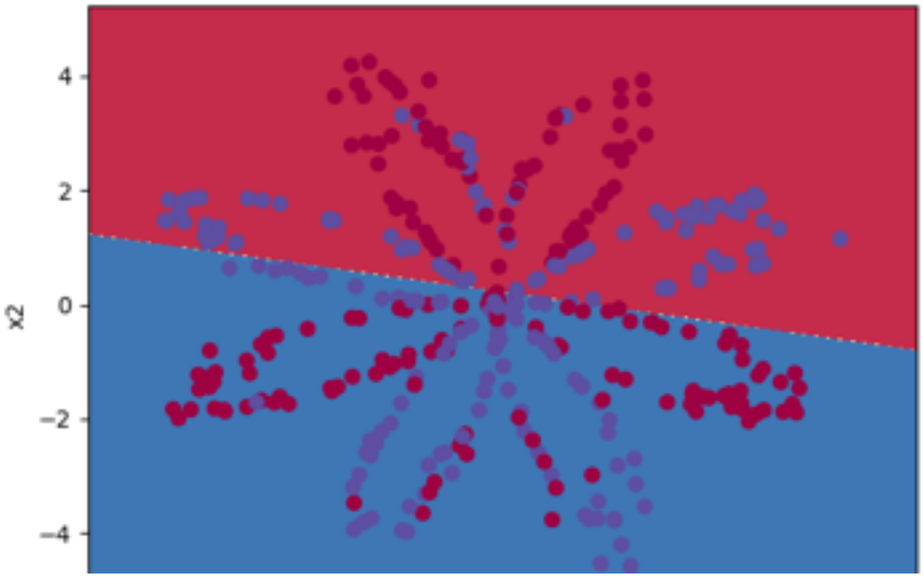
QUESTION 2:

(a) Program to implement general neural network architecture:

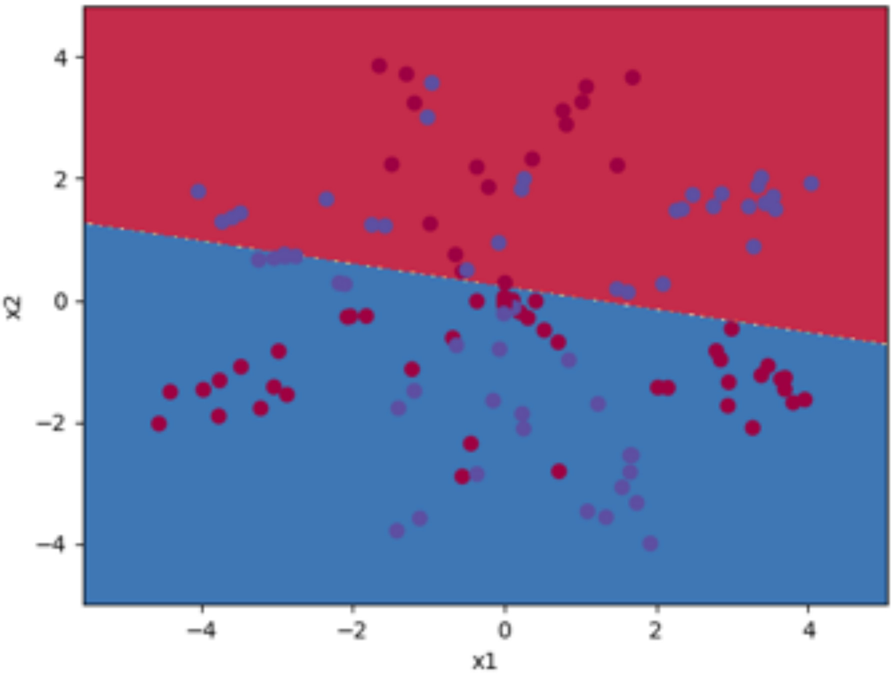
#Network trained by Stochastic Gradient Descent process
#input layer containing nodes = number of attributes/features in the X dataset
#hidden layers
#output layer containing one node which outputs between 0 to 1
if output<0.5 then the final output should be 0
and output>0.5 means final output should be 1

(b) (1) Network trained using logistic regression classifier:

Train accuracy:45.7%
Test accuracy: 38.33%



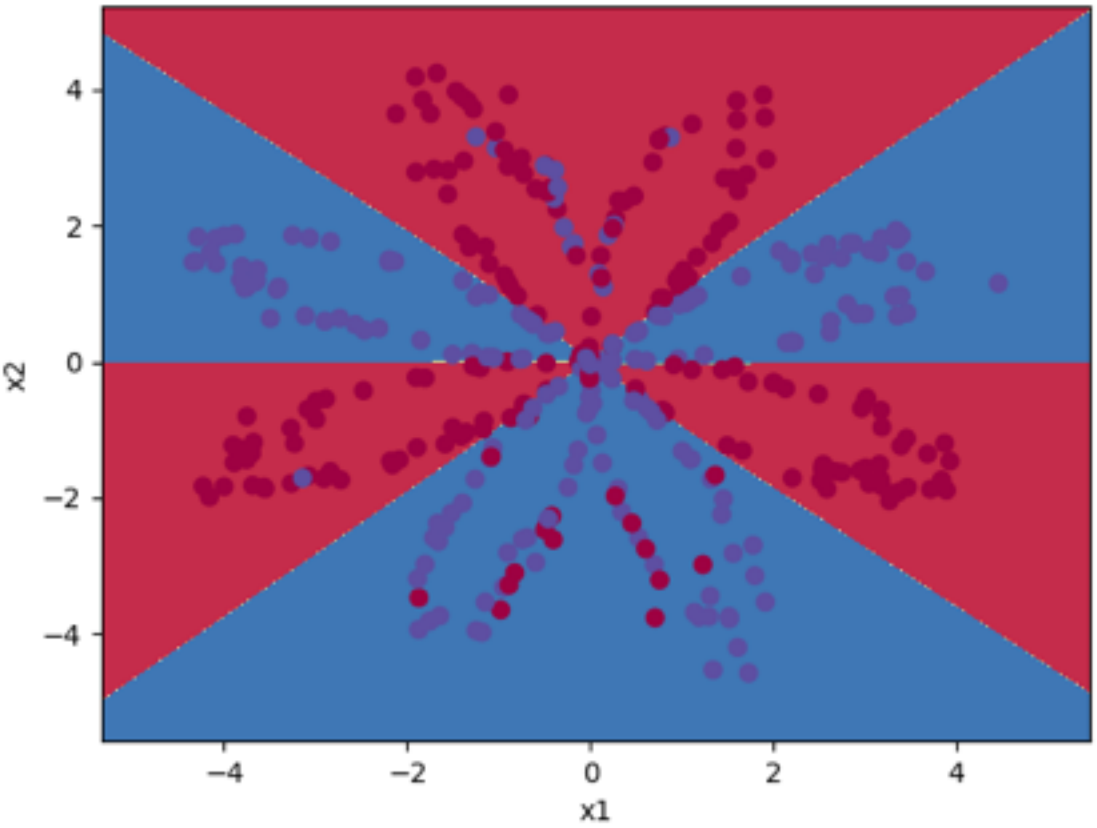
Training data



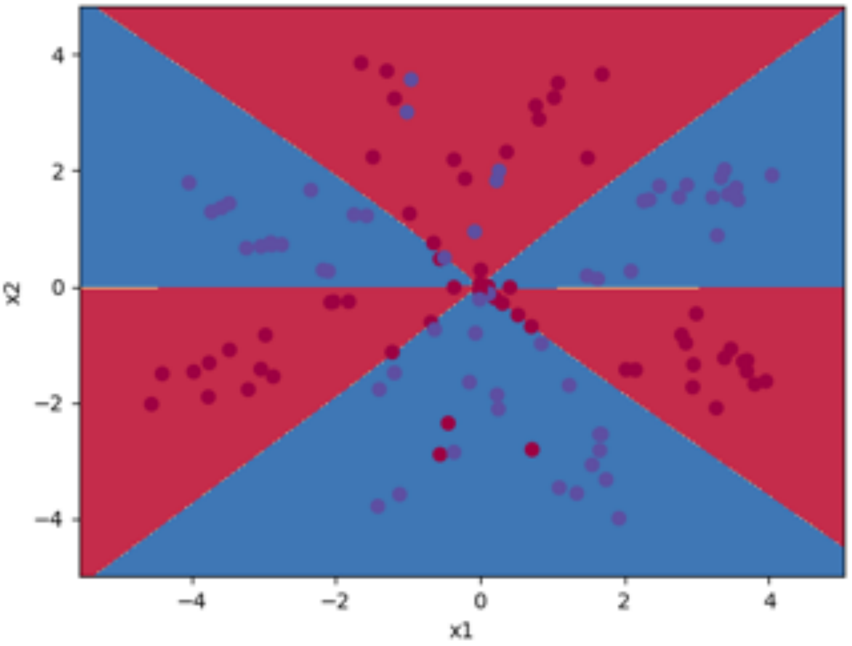
Test data

(2) Network with a single hidden layer having 5 units(2000 iterations,batch size=length of training examples):

Training score: 89.736%



Test score: 88.33%

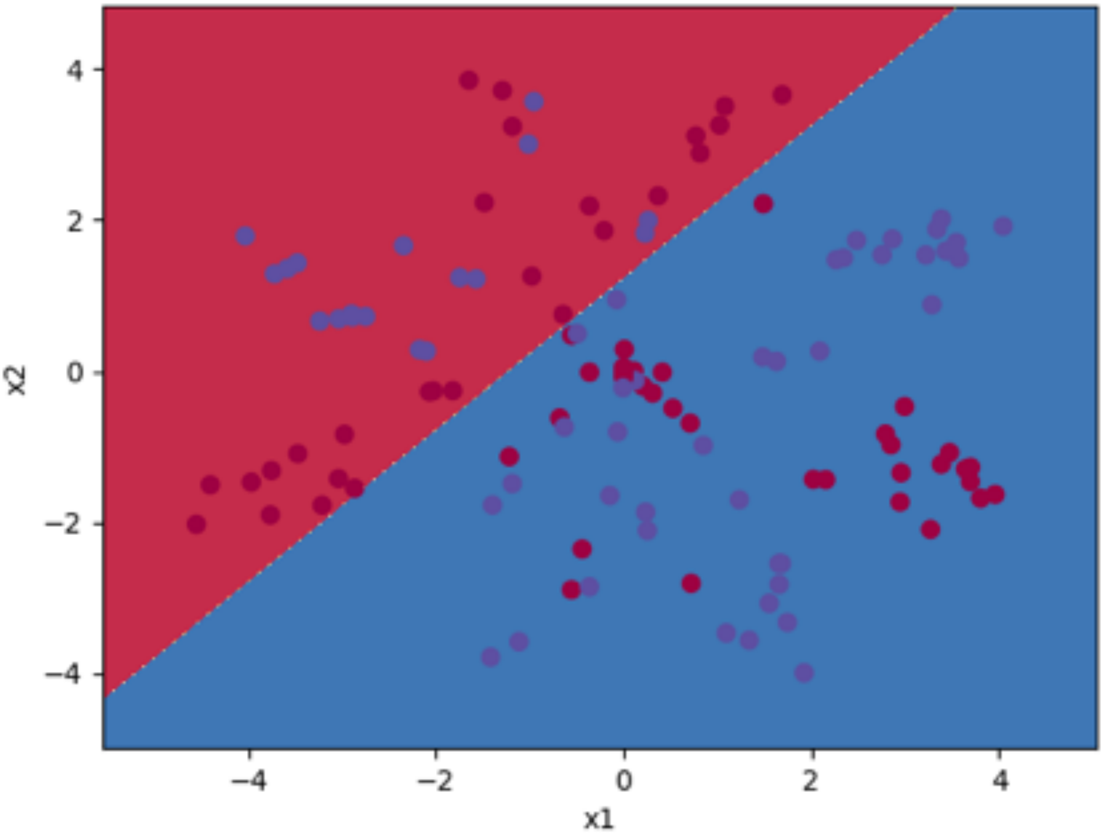


(3) Network with 1 hidden layer:

Number of units=1

Training accuracy: 51.84%

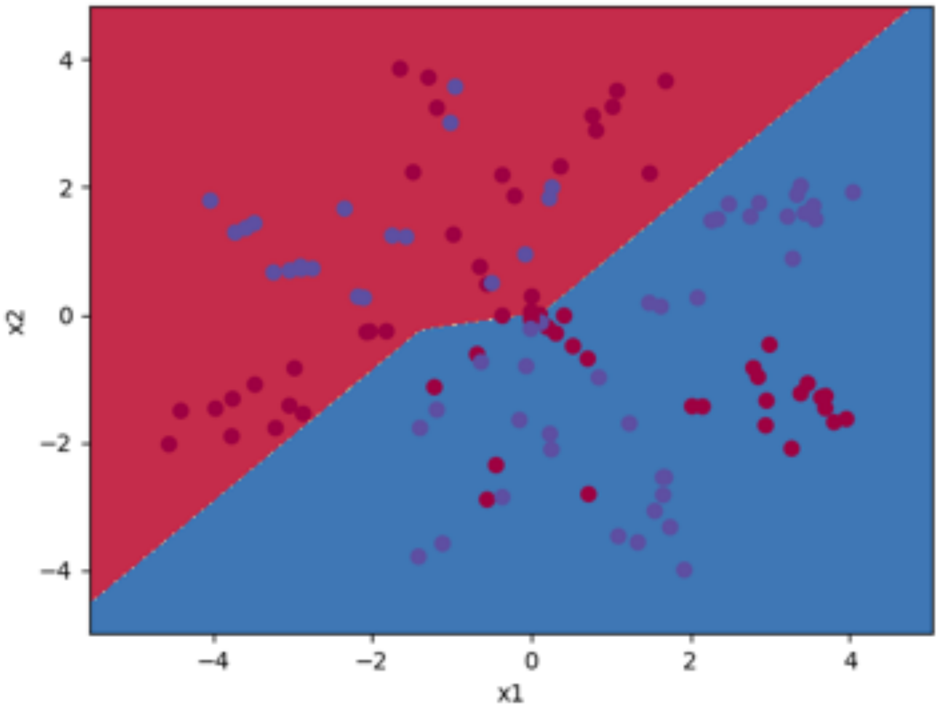
Testing accuracy:48.83%



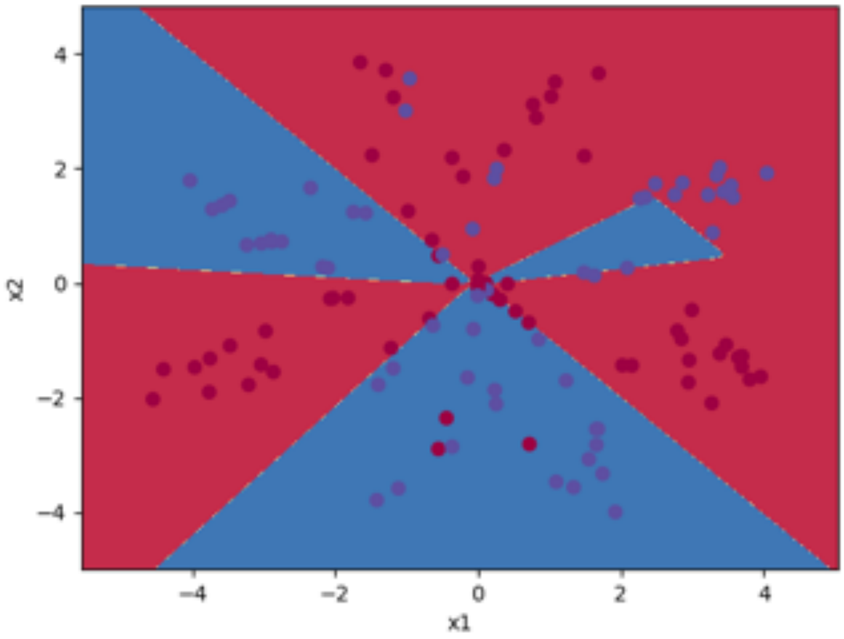
Number of units=2

Training accuracy:53.66%

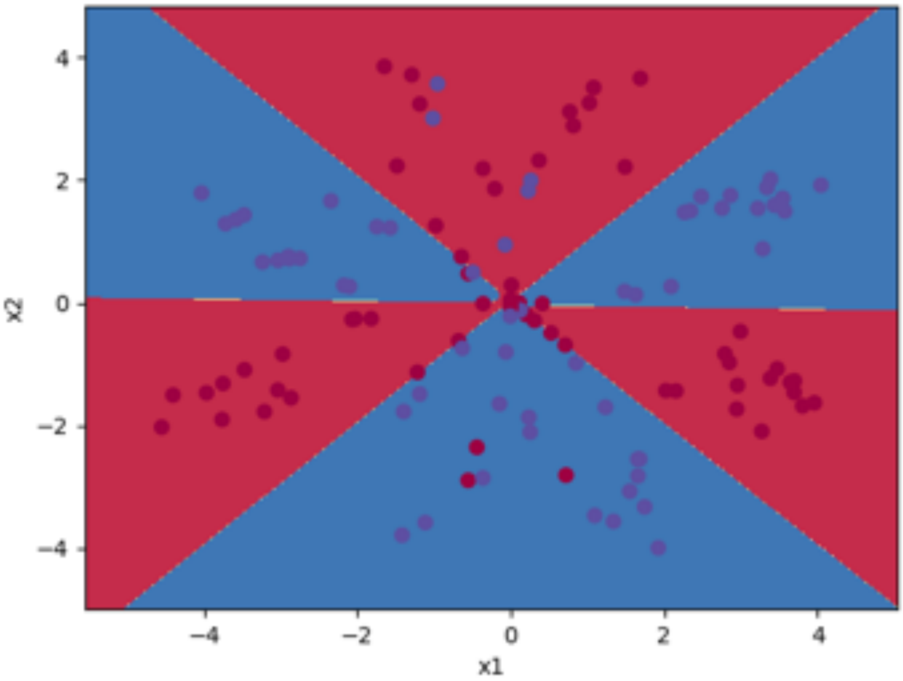
Testing accuracy:48.33%



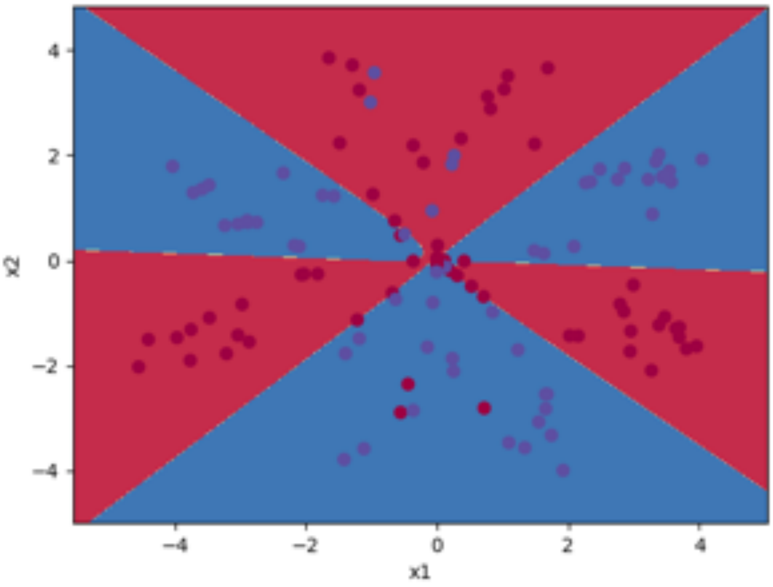
Number of units=3
Training accuracy: 79.4%
Testing accuracy: 75.16%



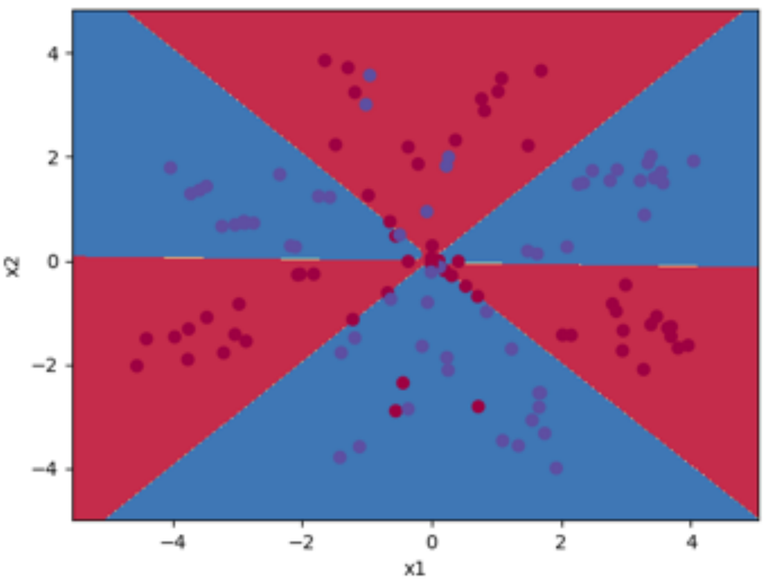
Number of units=10
Training accuracy: 89.7%
Testing accuracy: 85.8%



Number of units=20
Training accuracy: 90%
Testing accuracy: 86.667%

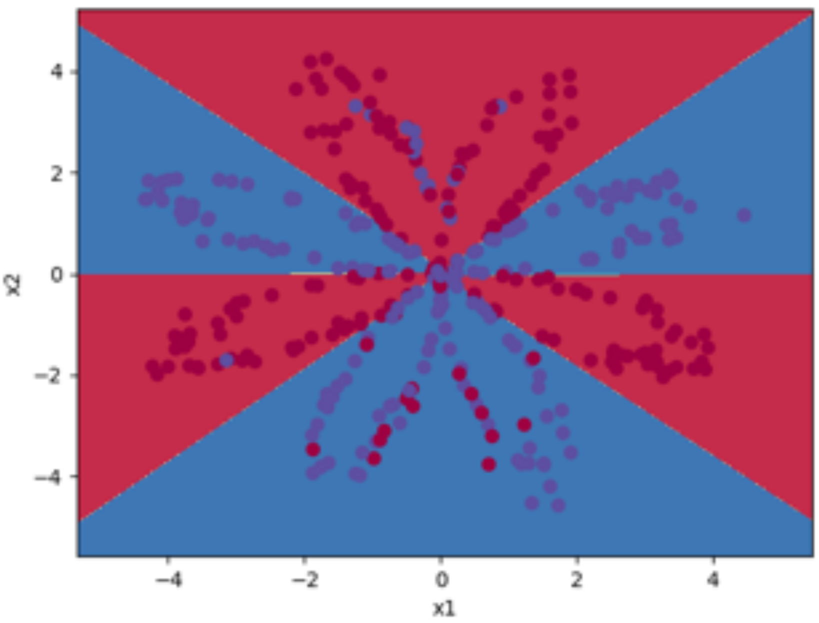


Number of units=40
Training accuracy: 88.26%
Testing accuracy:83.33%

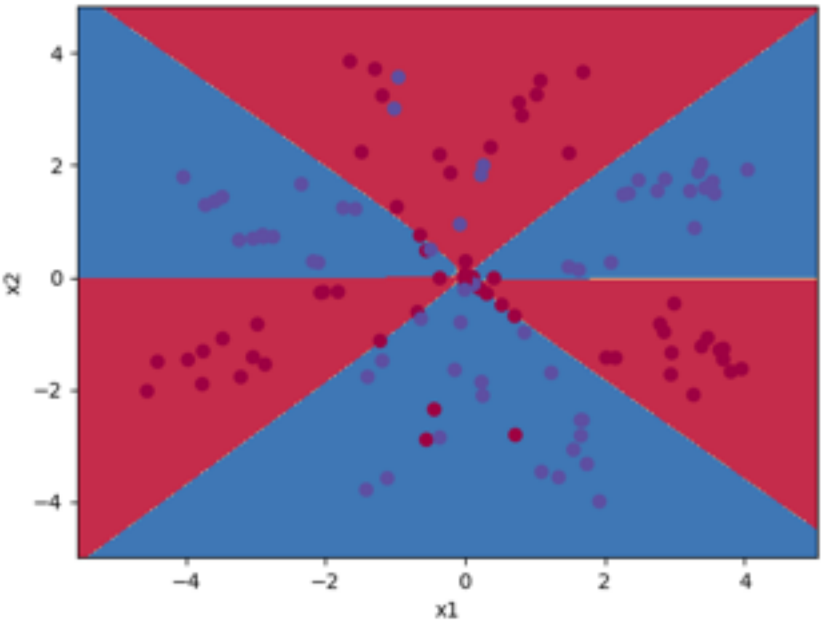


Based on the experimentation, number of optimal units can be either 10,20 or 40 since they all represent similar training and testing accuracies. Number of units =20 seems to be most optimal over test accuracies. A reason can be that as the connections increase, there are more options to adjust weights to give us the optimum values.

(4) Training accuracy = 90.26%



Testing accuracy = 85.833%



2 hidden layers with 5 units each is giving better testing accuracies than the previous part in which single hidden layer is implemented.

(c) Working with MNIST

(1) By LIBSVM, training accuracy= 100%, test accuracy = 97.66%