CSE567

Programming Assignment #1

Handwritten Digits Classification Using Neural Networks

Date: March 4th 2016

Group #32

Gian Pietro Farina (50176742)

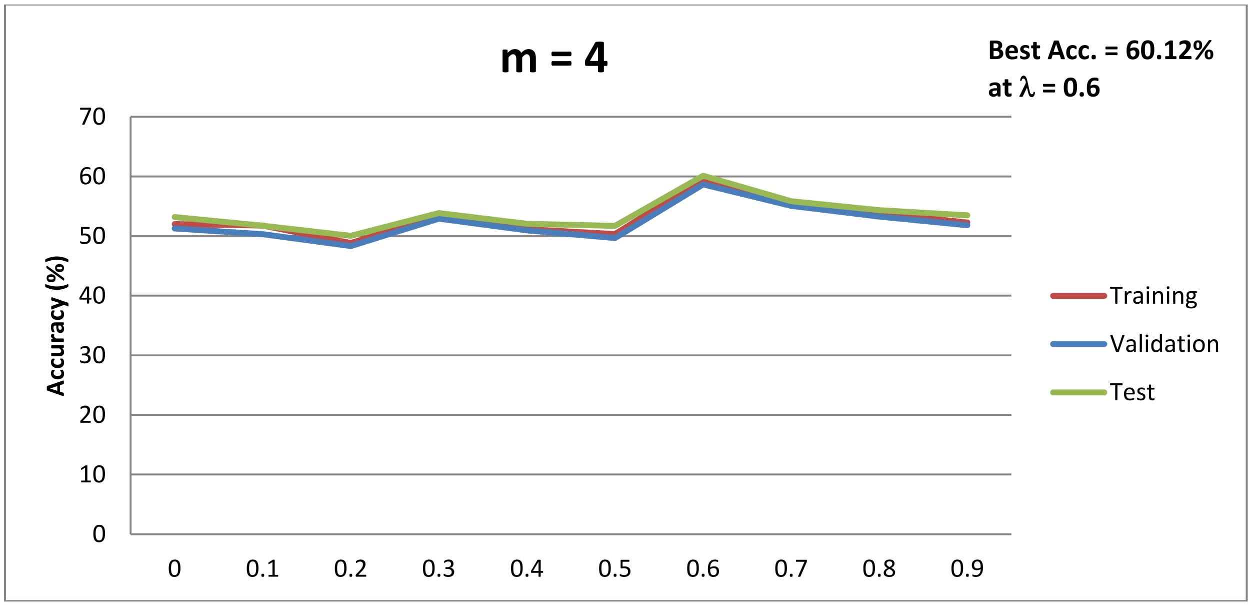
Meghana Ananth Gad (50182335)

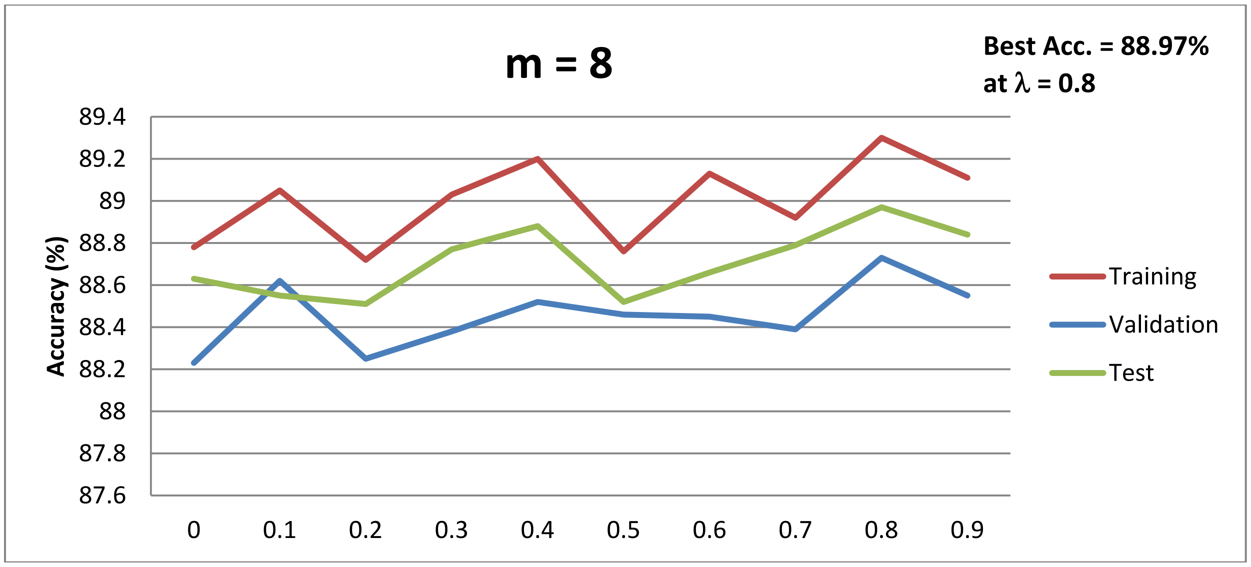
Ashwin Mittal (50168798)

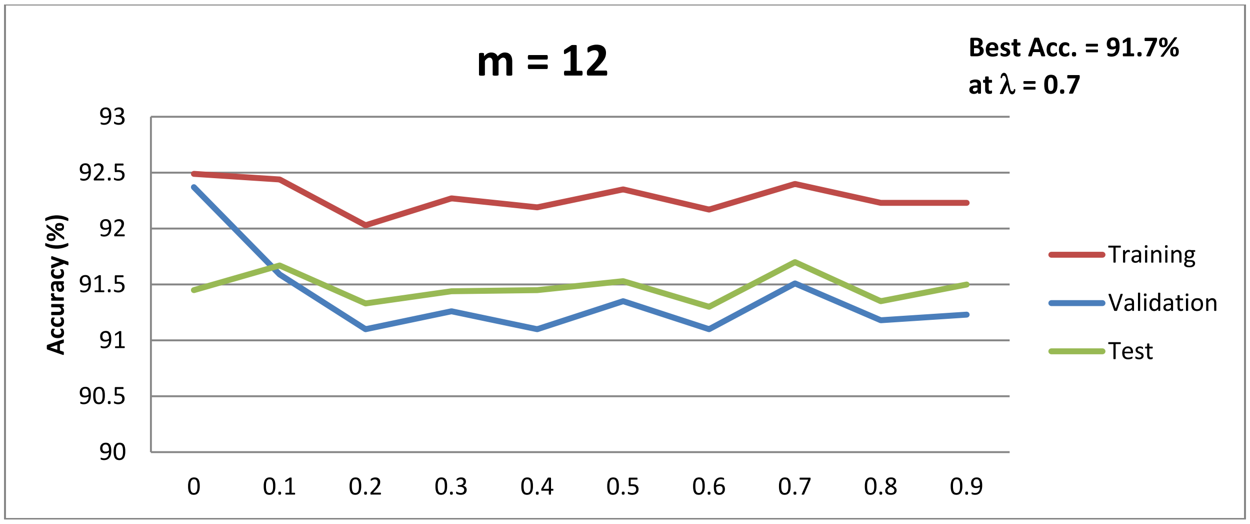
HYPER-PARAMETER SELECTION

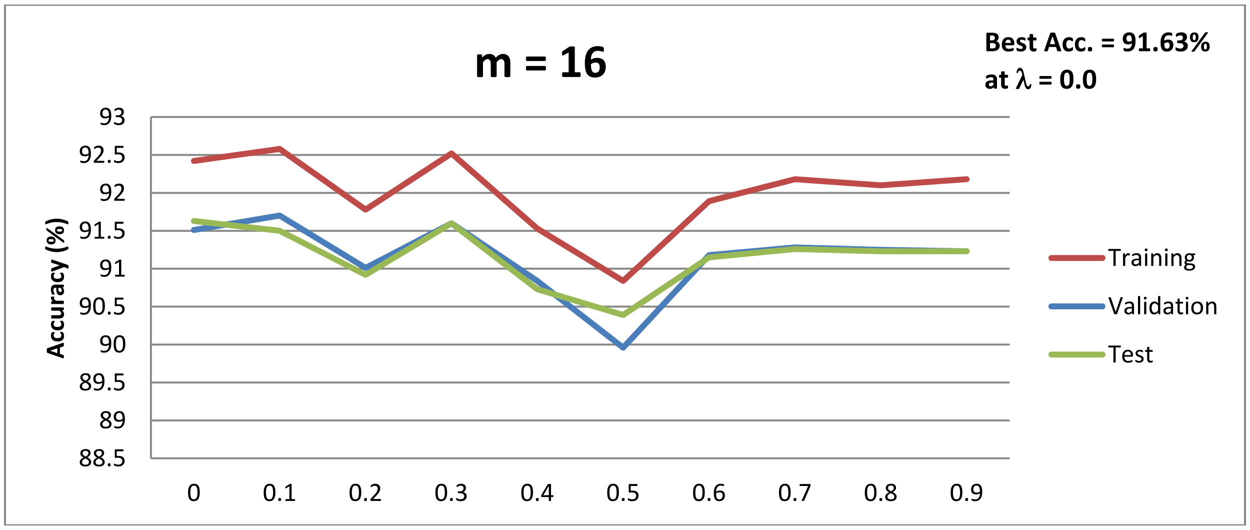
We ran experiments to find satisfactory values for the number of nodes (m) in the hidden layer and the normalization value (λ) using the grid search technique. For the number for nodes varying from 4 to 20 at intervals of 4, we calculated the accuracy on Training, Validation and Test data by setting values of λ between 0 and 1 with step size 0.1. This experiment took around 18 hours. Although it would have taken much longer time to search the entire hyper-parameter space, and it was possible to find relatively satisfactory hyper-parameters.

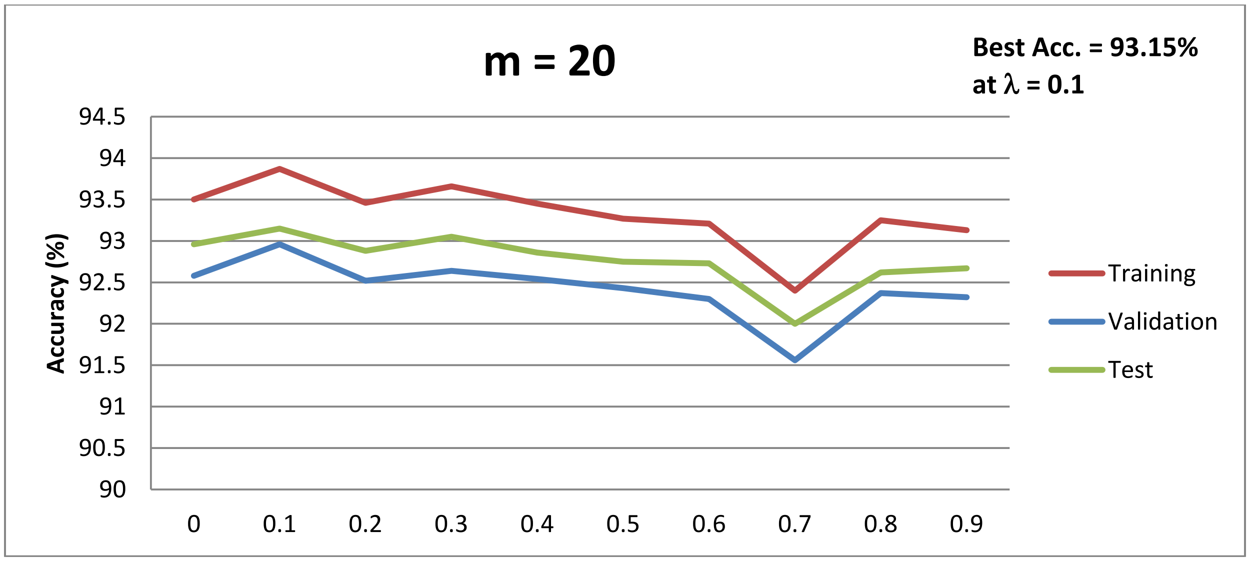
The results of the experiments are shown in the charts below, where m stands for number of hidden nodes in the network and λ is regularization co-efficient:






m = 12
Best Acc. = 91.7%
at λ = 0.7
Accuracy (%)
93
92.5
92
91.5
91
90.5
90
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Training
Validation
Test

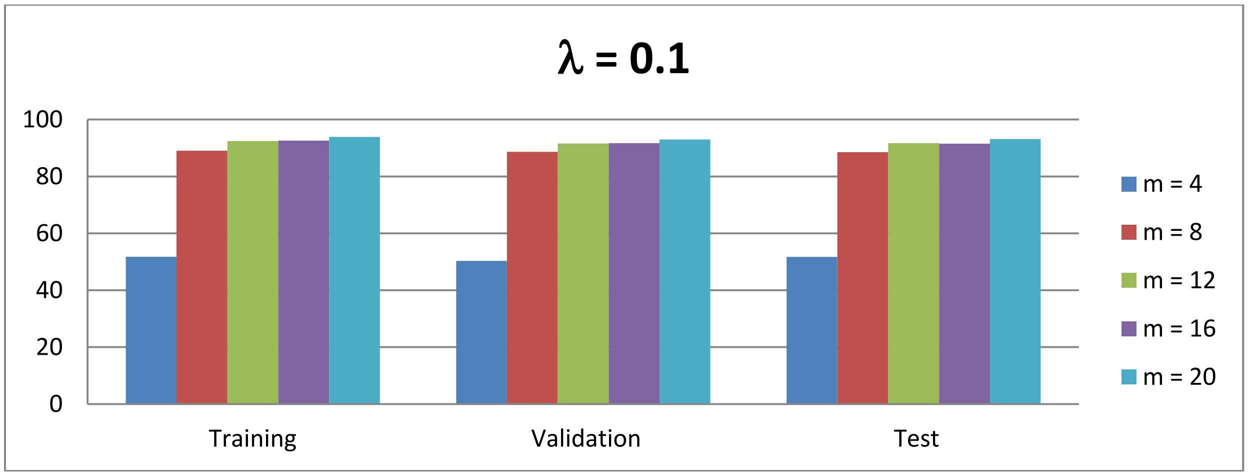

m = 16
Best Acc. = 91.63%
at λ = 0.0
Accuracy (%)
93
92.5
92
91.5
91
90.5
90
89.5
89
88.5
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Training
Validation
Test


m = 20
Best Acc. = 93.15%
at λ = 0.1
Accuracy (%)
94.5
94
93.5
93
92.5
92
91.5
91
90.5
90
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Training
Validation
Test

We made the following observations about the above plots:

- As the number of hidden nodes was increased, the best training, validation and testing accuracies increased.
- We also noticed that for cases where number of hidden nodes (m) was set to 4,8,12 best accuracies for test, training and validation sets were noted for higher values of λ(greater than 0.5).
- When the number of hidden nodes was set 16 and 20 the best accuracies were observed for smaller values of λ(less than 0.2).

We got the best accuracy for λ = 0.1 with 20 nodes. So we fixed λ to 0.1 and compared the accuracies obtained when the number of hidden nodes was set to 4, 8, 12, 16 and 20. The comparison for λ = 0.1 for all nodes is represented in following graph:



CONCLUSION

If the number of hidden nodes are set to a very small value (m=4 in our experiment) the performance of the network is not optimal. The value of generalization co-efficient for which we get optimal results varies depending on the number of hidden nodes in the network. This is due to the specific error function E(w1,w2,lambda)=y that we are minimizing. If we fix y then we are going to see a lower lambda when we have an increased number of hidden nodes with respect to when the number of hidden nodes is small.

When the value of hidden nodes increases the best accuracies are observed for smaller values of 'λ'. Hence, smaller values of 'λ' must be chosen when the number of hidden nodes are greater.

Please note:

- Random initial weights have been computed at the beginning, once and for all. So all the iterations start from the same initial random weights. This should not affect the results much because with enough data we usually converge independently by the initial starting point.

- We conducted the random permutation of the data only once during the pre processing stage and used the same data for all experiments. The final results are highly dependent of this permutation and we might have obtained different results if the permutation were different. For more precise results we should implement k-cross validation.