MSCI-641 Assignment4

Results

Activations	Sigmoid			ReLU			Tanh		
Data Splits	Training	Validation	Testing	Training	Validation	Testing	Training	Validation	Testing
No Drop out and Reguralization	0.939	0.795	0.794	0.9225	0.7937	0.794	0.9117	0.796	0.796
L2-Reguralization(0.01)	0.8283	0.8093	0.808	0.8355	0.8113	0.807	0.8357	0.806	0.801
L2-Reguralization(0.1)	0.8124	0.809	0.8	0.8183	0.806	0.804	0.818	0.8067	0.804
L2-Reguralization(10)	0.7871	0.7863	0.785	0.7966	0.7964	0.794	0.8	0.7989	0.798
Dropout(0.3)	0.8715	0.8155	0.813	0.869	0.814	0.813	0.8633	0.8094	0.807
Dropout(0.5)	0.849	0.816	0.812	0.8434	0.8133	0.811	0.8485	0.8112	0.814
Dropout(0.7)	0.8277	0.8136	0.81	0.8264	0.8109	0.808	0.8311	0.813	0.812
L2- Reguralization(0.01) + Dropout(0.5)	0.805	0.8099	0.806	0.8124	0.8101	0.808	0.8119	0.808	0.809
L2- Reguralization(0.01) + Dropout(0.7)	0.7949	0.812	0.809	0.8048	0.8103	0.807	0.8106	0.8088	0.805
L2- Reguralization(0.01) + Dropout(0.3)	0.8106	0.8089	0.808	0.8168	0.8087	0.806	0.8164	0.808	0.806

Explanation of Results

The above table summarizes my results for most of the possible cases. I was able to get a test set accuracy of 81.4% with the use of Tanh activation function and dropout of 0.50. Activation functions are the backbone of neural networks without which a neural network would be a mere linear classifier. In my analysis out of the three activation functions ={sigmoid,ReLU,Tanh} ,Tanh proves to be the best. This is because as Tanh is bounded between -1 and 1 unlike sigmoid which is bounded between 0 and 1 hence its gradient is much more *steeper*(bounded between 0 and 1) than sigmoid's gradient(bounded between 0 and 0.25) which results in better update and learning of weights during back propagation. In addition,Tanh is also better than ReLU because ReLU has a *dying ReLU* problem. As ReLU = max(0,x), the negative side is zero which results in neurons output being always equal to zero in the case when they get a negative input. The real strength of ReLU lies in solving vanishing gradient problem but as the network is shallow here therefore this advantage is also not obvious. Hence, in shallow networks Tanh should and does work well than the other two activation functions. In addition, dropout also improves the results. It can be seen from the results that in the case of no regularization and dropout the training accuracy is much higher than test accuracy which is a clear sign of overfitting. Now, in order to combat *overfitting* both regularization and dropout are used but in our case even tough both reduce the effect of overfitting but dropout results in a much better accuracy. This is because regularization does not combat *co-adaptation problem* – as all the weights are learned together therefore weights having more predictive capability would be learnt more. As dropout simply omits random nodes without any bias to predictive capability therefore it is suited to deal with co-adaption problem. However a dropout of .70 is way too large and 0.30 to less for a shallow network therefore a dropout rate in between the two i.e 0.50 is optimal for a shallow network. Finally, a combination of dropout and regularization might cause underfitting and reduce the capability of the model as nodes are being dropped with a decrease in weights as well. Hence, an activation function of Tanh and dropout of 0.5 giving optimal result also gives an intuitive understanding.