ECS170
Xie Zhou, 912143385

Problem Set 5

Problem 1)
The training data has to be linearly separable for the perceptron algorithm to make perfect predictions. If not, the algorithm will not converge. There must be a hyperplane that can separate the training data.

Problem 2)

a) The function can be differentiated and then be used in the gradient descend method. When the change in the weight vector is small enough, we know it converges.
b) The function is convex. This property guarantees that it will converge at its local/global minimum.

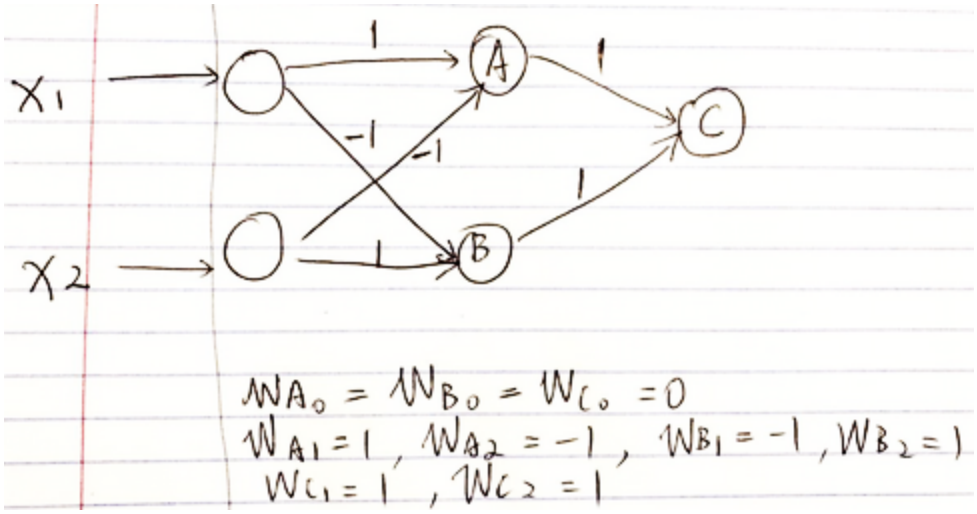
Problem 3)
Regularization can help avoid overfitting. It adds a loss function to the original objective function. This method eliminates less influential features/weights which can potentially over-train our model.

Problem 4)
Stochastic gradient descent update each w_i in the weight for each sample j whereas standard gradient descent update each w_i for all the samples (sums up all errors). Therefore, SGD updates weight vector more frequently. It converges much faster than standard GD, especially when given a large set of data.

Problem 5)

a) Perceptron units are less sensitive to small changes. Sigmoid functions are smooth and their outputs are in [0, 1]. Perceptron algorithm outputs either 0 or 1 (hard boundary w/ thresholds); it uses a step function. Therefore, on extreme ends of the two functions, their outputs are similar. However, say on small changes in weights, sigmoid units will react to the change whereas perceptron units won't.
b) The sigmoid function adds nonlinearity to the network. With only linear units, the network can only learn functions whose results are linear combinations of its inputs. In addition, the sigmoid function has its derivative sig(x)' = sig(x)*(1-sig(x)). Its relationship with its derivative is convenient for computation in back-propagation.

Problem 6)



g(n) = 0 if $w^T x <= 0$;

g(n) = 1 if $w^T x > 0$.