Homework 6

1. **Clearly describe each of these characteristics of a neural network:**

**a. Layered :** A neural network consist of multiple layers through which the data is passed through. That starts from input layer, hidden layer and output layer. Usually every layer is formed by multiple neurons also knowns as nodes that has its own computation formula that is composed of summation and an activation function which helps to calculate the estimated or predicated value for the results. This helps to get into further solving of the problem combined by the filtered values or output of previous layers, that is fed as input to the next layer.

**b. Feedforward –** A feedforward network is something that will restrict the network to follow a single direction of data or information flow and does not allow looping or cycling. The neural network is composed of two or more layers, most networks consist of three layers Input, hidden and an output layer that is composed of newurons or nodes at each layer. This helps to build a network that takes inputs at the input layer and process them forward to the next layers. Hence it is called as Feed forward network

**c. Completely connected –** A completely connected network is the one that every node at a given layer is connected to the node in the next layer, but not the nodes in the same layer as they are. Hence they are known as completely connected network. Each connection between nodes has a weight associated with it, that are assigned with value between 0 and 1 at random and improvised over the iteration of model improvement.

1. **What is the sole function of the nodes in the input layer?**

Every node in the input layer is composed of the multiple nodes which carry various attributes belonging to a particular dataset the number of input nodes depend on the number and type of attributes in the dataset. Every node in the input layer processes the data with its computational function composed of summation and the activation function which is used to process the data with defined weights for tat particular node

1. **Should we prefer a large hidden layer or a small one?Describe the benefits and drawbacks of each.**

More nodes in the hidden layer increase the power and flexibility of the network for identifying complex patterns hence they can opt for large number of nodes in the hidden layer. But overly large hidden layer will lead to overfitting of the model and tend to memorize the training set instead of generalized model. If the model tends to overfit, one can reduce the number of nodes in the hidden layer. Also if training accuracy is very low a better accuracy can be achieved by adding some more number of nodes in the hidden layer

1. **Explain why the updating term for the current weight includes the *negative* of the sign of**

**the derivative (slope).**

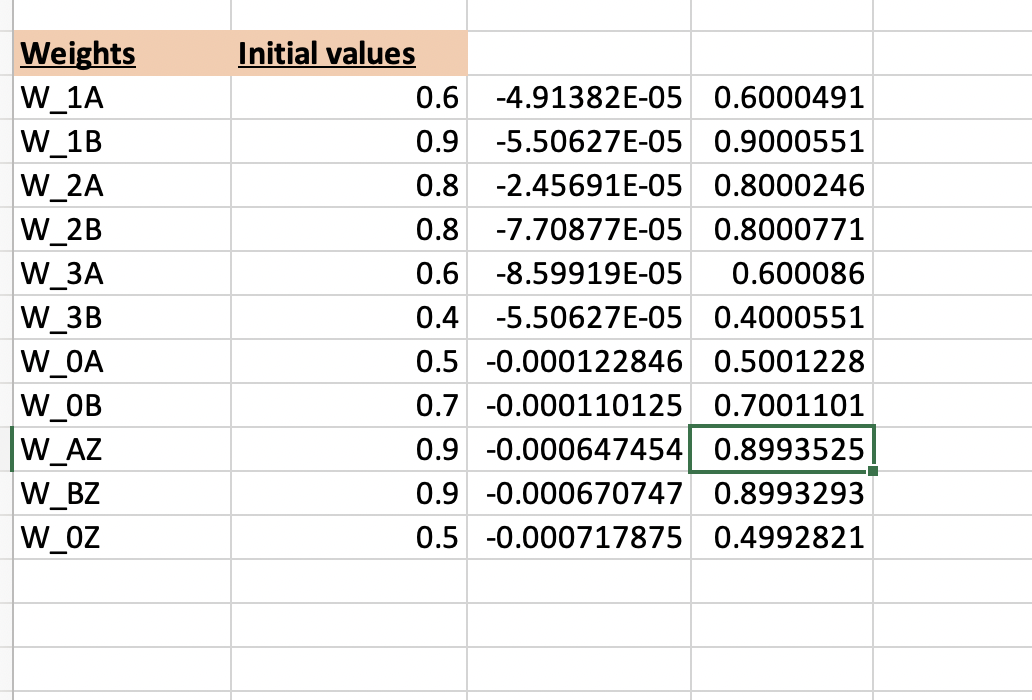
When we are adjusting the values of w1 close to the Left side of the derivative of SSE at Wcurrent, the slope is negative where we try to adjust it to the more closer optimal value that can be adjusted in current W1L defined. In order to adjust the resultant optimal value towards Wcurrent that is the negative sign of the derivative of SSE at Wcurrent, hence the sign will be negative that is - -0sgin (dSSE / dWcurrent)

1. **Adjust the weights *W*0*B*, *W*1*B*, *W*2*B*, and *W*3*B* from the example on back-propagation in the text.**

Adjusting weights *W*0*B*, *W*1*B*, *W*2*B*, and *W*3*B*.

The weights - *W*0*B*, *W*1*B*, *W*2*B*, and *W*3*B* is as follows –

*W*0*B* = 0.70  
 *W*1*B* = 0.9  
 *W*2*B* =0.8  
 *W*3*B =* 0.4



1. **Refer to Exercise 7. Show that the adjusted weights result in a smaller prediction error.**

Adjusted weights result in smaller prediction error – as seen from above example, the adjusted weights are result of reworking on the percolated error in the prediction throughout the network. The network calculated a predicted value for the target variable and compared to the actual expected target value this generates an error in prediction of actual value hence it reworks on the weights to reduce this error found. This helps to reduce the error in weights and result in smaller prediction error.

1. **True or false: Neural networks are valuable because of their capacity for always finding the global minimum of the SSE.**

**True**

1. **Describe the benefits and drawbacks of using large or small values for the learning rate.**

**Learning rate** – Usually the learning rate is in the range 0 to 1 that is a constant used to move the network weights towards global minimum of SSE. Hence it is critical to work on the optimal value of the learning rate.   
When learning rate is small the weight adjustments tend to be very small. If learning rate is small the network will take longer time to reach its optimal solution. But again if we take large value of 𝜂, it will tend to overshoot the optimal solution. Hence this will be too fast to reach the goal ans we might miss out ton the optuimal values of the learning rate to give us the best possible minimal SSE. Hence it is required to define a value that will be best optimal not too large or not too small to not to reach our goal. If we have a larger learning rate, on adding the value on other side will happen to have the new weight value to the opposite side of the W th next adjustment will overshoot W\*, that will cause oscillation between the two slopes of the gradient descent and never settle to a perfect value.

1. **Describe the benefits and drawbacks of using large or small values for the momentum term.**

**Momentum term - the back propagation algorithm uses addition of a momentum term called alpha** *𝛼*,. That is calculated as follows –

Δ*w*current = −*𝜂 𝜕*SSE + *𝛼*Δ*w*previous *𝜕w*current

where Δ*w*previous represents the previous weight adjustment, and 0 ≤ *𝛼 <* 1.

Thus, the new component *𝛼*Δ*w*previous represents a fraction of the previous weight adjustment for a given weight. Hence *𝛼k* term indicates large values of *𝛼* allow algorithm to remember more terms in adjustment history. Small values of *𝛼*  reduce the inertial effects and influence of previous adjustments.

Hence momentum component helps to dampen the oscillations around optimality by encouraging adjustments to stay in same direction. Early adjustment might be in same direction so that exponential average of adjustments will be in that direction. Momentum is helpful when the gradient of SSE wrt w is flat. If momentum is too large the wright adjustment may again overshoot the minimum. Due to cumulative influences of previous adjustments.