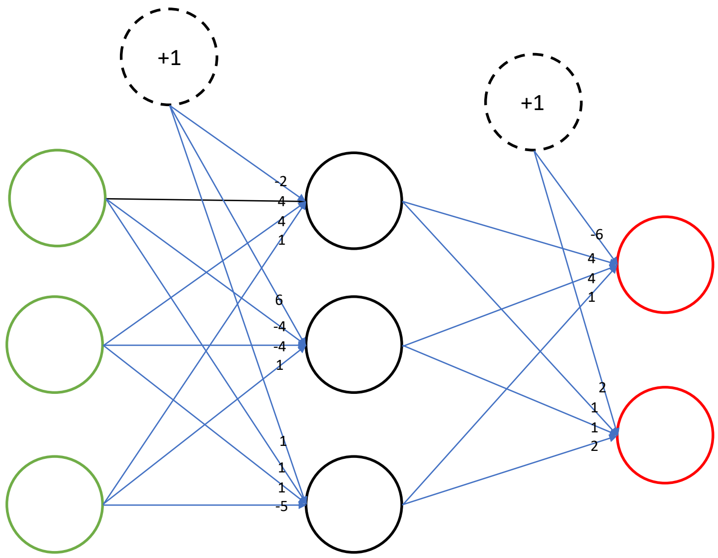
Theodore Kim
CS-UY 4563 – Introduction to Machine Learning
HW #6

1. Given the neural network:

$$\boldsymbol{W}^{(1)} = \begin{pmatrix} \mathbf{4} & \mathbf{4} & \mathbf{1} \\ \mathbf{-4} & \mathbf{-4} & \mathbf{1} \\ \mathbf{1} & \mathbf{1} & \mathbf{5} \end{pmatrix}, \boldsymbol{W}^{(2)} = \begin{pmatrix} \mathbf{4} & \mathbf{4} & \mathbf{1} \\ \mathbf{1} & \mathbf{1} & \mathbf{2} \end{pmatrix}, \boldsymbol{b}^{(1)} = \begin{pmatrix} \mathbf{-2} \\ \mathbf{6} \\ \mathbf{1} \end{pmatrix}, \boldsymbol{b}^{(2)} = \begin{pmatrix} \mathbf{-6} \\ \mathbf{2} \end{pmatrix}$$

(a) Draw Neural Network



(b) Determine $\boldsymbol{h_{W,b}(x)}$ when $\boldsymbol{x^T = (1, 2, 3)}$

$$h_{W,b}(x) = \begin{bmatrix} f(z_1^{(3)}) \\ f(z_2^{(3)}) \end{bmatrix} = \begin{bmatrix} f(W_1^{(2)} a^{(2)} + b_1^{(2)}) \\ f(W_2^{(2)} a^{(2)} + b_2^{(2)}) \end{bmatrix} = \begin{bmatrix} f(W_1^{(2)} \left(f\left(z^{(2)}\right)\right) + b_1^{(2)}) \\ f(W_2^{(2)} \left(f\left(z^{(2)}\right)\right) + b_2^{(2)}) \end{bmatrix}$$

$$h_{W,b}(x) = \begin{bmatrix} f\left(W_1^{(2)} \begin{bmatrix} f\left(W_1^{(1)} a^{(1)} + b_1^{(1)}\right) \\ f\left(W_2^{(1)} a^{(1)} + b_2^{(1)}\right) \\ f\left(W_3^{(1)} a^{(1)} + b_3^{(1)}\right) \end{bmatrix} + b_1^{(2)} \right) \\ f\left(W_2^{(2)} \begin{bmatrix} f\left(W_1^{(1)} a^{(1)} + b_1^{(1)}\right) \\ f\left(W_2^{(1)} a^{(1)} + b_2^{(1)}\right) \\ f\left(W_3^{(1)} a^{(1)} + b_3^{(1)}\right) \end{bmatrix} + b_2^{(2)} \right) \end{bmatrix}$$

$$h_{W,b}(x) = \begin{bmatrix} f\left(\begin{bmatrix}4 & 4 & 1\end{bmatrix}\begin{bmatrix} f\left(\begin{bmatrix}4 & 4 & 1\end{bmatrix}\begin{bmatrix}1\\2\\3\end{bmatrix} - 2\right) \\ f\left(\begin{bmatrix}-4 & -4 & 1\end{bmatrix}\begin{bmatrix}1\\2\\3\end{bmatrix} + 6\right) \\ f\left(\begin{bmatrix}1 & 1 & 5\end{bmatrix}\begin{bmatrix}1\\2\\3\end{bmatrix} + 1\right)\end{bmatrix} - 6\right) \\ f\left(\begin{bmatrix}1 & 1 & 2\end{bmatrix}\begin{bmatrix} f\left(\begin{bmatrix}4 & 4 & 1\end{bmatrix}\begin{bmatrix}1\\2\\3\end{bmatrix} - 2\right) \\ f\left(\begin{bmatrix}-4 & -4 & 1\end{bmatrix}\begin{bmatrix}1\\2\\3\end{bmatrix} + 6\right) \\ f\left(\begin{bmatrix}1 & 1 & 5\end{bmatrix}\begin{bmatrix}1\\2\\3\end{bmatrix} + 1\right)\end{bmatrix} + 2\right)\end{bmatrix}$$

$$h_{W,b}(x) = \begin{bmatrix} f\left(\begin{bmatrix}4 & 4 & 1\end{bmatrix}\begin{bmatrix}f(13)\\f(-3)\\f(19)\end{bmatrix} - 6\right) \\ f\left(\begin{bmatrix}1 & 1 & 2\end{bmatrix}\begin{bmatrix}f(13)\\f(-3)\\f(19)\end{bmatrix} + 2\right)\end{bmatrix}$$

$$h_{W,b}(x) = \begin{bmatrix} f(4f(13) + 4f(-3) + f(19) - 6) \\ f(f(13) + f(-3) + 2f(19) + 2)\end{bmatrix}$$

$$h_{W,b}(x) = \begin{bmatrix} f(3.999996 + 0.189703 + 0.999999 - 6) \\ f(3.999996 + 0.04743 + 1.999996 + 2)\end{bmatrix}$$

$$h_{W,b}(x) = \begin{bmatrix} f(-0.810302) \\ f(8.04722)\end{bmatrix}$$

$$h_{W,b}(x) = \begin{bmatrix} 0.3078 \\ 0.9997\end{bmatrix}$$

(c) For $x^T = (1, 2, 3)$ and $y = (1, 0)$, find $\delta_1^{(3)}$, $\frac{\partial J(W,b;x,y)}{\partial JW_{11}^{(2)}}$, $\delta_1^{(2)}$, and $\frac{\partial J(W,b;x,y)}{\partial JW_{11}^{(1)}}$

$$\delta_1^{(3)} = \frac{dJ}{dz_1^{(3)}} = -\left(y - f\left(z_1^{(3)}\right)\right) f'\left(z_1^{(3)}\right)$$

$$\delta_1^{(3)} = -\left(1 - f(W_1^{(2)}a^{(2)} + b_1^{(2)})\right) f'(W_1^{(2)}a^{(2)} + b_1^{(2)})$$

$$\delta_1^{(3)} = -(1 - 0.3078) f'(-0.810302)$$

$$\delta_1^{(3)} = -(1 - 0.3078)(0.3078 \cdot (1 - 0.3078))$$

$$\delta_1^{(3)} = -0.14748$$

$$\frac{\partial J(W, b; x, y)}{\partial J W_{11}^{(2)}} = a_1^{(2)} \delta_1^{(3)}$$

$$\frac{\partial J(W, b; x, y)}{\partial J W_{11}^{(2)}} = f(13)(-0.14748)$$

$$\frac{\partial J(W, b; x, y)}{\partial J W_{11}^{(2)}} = 0.9999(-0.14748)$$

$$\frac{\partial J(W, b; x, y)}{\partial J W_{11}^{(2)}} = -0.14748$$

$$\delta_1^{(2)} = \sum_{i=1}^{s_{(3)}} \left(\delta_1^{(3)} W_{i1}^{(2)} f'\left(z_1^{(2)}\right) \right)$$

$$\delta_1^{(2)} = \left(\delta_1^{(3)} f'\left(z_1^{(2)}\right) \right) \sum_{i=1}^{2} \left(W_{i1}^{(2)} \right)$$

$$\delta_1^{(2)} = \left(\delta_1^{(3)} f'\left(z_1^{(2)}\right) \right) \left(W_{11}^{(2)} + W_{21}^{(2)} \right)$$

$$\delta_1^{(2)} = (-0.14748)\big(0.99999 \cdot (1 - 0.99999)\big)(4 + 1)$$

$$\delta_1^{(2)} = -0.000007374$$

$$\frac{\partial J(W, b; x, y)}{\partial J W_{11}^{(1)}} = a_1^{(1)} \delta_1^{(2)}$$

$$\frac{\partial J(W,b;x,y)}{\partial J W_{11}^{(1)}} = 1(-0.000007374)$$

$$\frac{\partial J(W,b;x,y)}{\partial J W_{11}^{(1)}} = -0.000007374$$

2. Given the neural network: 3 layers, 2 neurons (input), 2 neurons (hidden), and 1 neuron (output):

$$\boldsymbol{W}^{(1)} = \begin{bmatrix} \mathbf{1} & \mathbf{2} \\ \mathbf{3} & \mathbf{4} \end{bmatrix}, \boldsymbol{W}^{(2)} = \begin{bmatrix} \mathbf{1} \\ \mathbf{2} \end{bmatrix}, \boldsymbol{b}^{(1)} = [\mathbf{1} \quad -\mathbf{1}], \boldsymbol{b}^{(2)} = [\mathbf{1}]$$

Training set: ((1, 0), 1), ((0, 1), 0). Perform 1 step of gradient descent where learning rate = 0.2.

<u>**Gradient Descent Steps**</u>	<u>**Updating *W***</u>	<u>**Updating *b***</u>
Updating the first layer (1) weights and intercept values	$W^{(1)} = W^{(1)} - \frac{a}{N}(\Delta W^{(1)})$	$b^{(1)} = b^{(1)} - \frac{a}{N}(\Delta b^{(1)})$
Determine delta values for the first training sample	$\Delta W^{(1)} = \Delta W^{(1)} + \delta^{(2)}(a^{(1)})^T$	$\Delta b^{(1)} = \Delta b^{(1)} + \delta^{(2)}$
Determine small delta value for layer 2 ($\delta^{(2)}$) (back propagation)	$\delta^{(2)} = \left(\left(W^{(2)}\right)^T \delta^{(3)}\right) \cdot \left(f\left(z^{(2)}\right)\left(1 - f\left(z^{(2)}\right)\right)\right)$ $\delta^{(2)} = \left(\begin{bmatrix}1\\2\end{bmatrix}(-0.00112)\right) \cdot \left(\begin{bmatrix}0.881\\0.731\end{bmatrix}\left(1 - \begin{bmatrix}0.881\\0.731\end{bmatrix}\right)\right)$ $\delta^{(2)} = \left(\begin{bmatrix}-0.00112\\-0.00224\end{bmatrix}\right) \cdot \left(\begin{bmatrix}0.881\\0.731\end{bmatrix}\left(\begin{bmatrix}0.119\\0.269\end{bmatrix}\right)\right)$ $\delta^{(2)} = \left(\begin{bmatrix}-0.00112\\-0.00224\end{bmatrix}\right) \cdot \left(\begin{bmatrix}0.105\\0.197\end{bmatrix}\right)$ $\delta^{(2)} = \begin{bmatrix}-0.0001176\\-0.0004413\end{bmatrix}$	
Determine input values from layer 1 ($a^{(1)}$) (forwards propagation)	$a^{(1)} = x$ $a^{(1)} = \begin{bmatrix}1\\0\end{bmatrix}$	**N/A**
Update delta values using $\delta^{(2)}$ and $a^{(1)}$	$\Delta W^{(1)}$ $= 0 + \begin{bmatrix}-0.0001176\\-0.0004413\end{bmatrix}[1 \quad 0]$ $\Delta W^{(1)} = \begin{bmatrix}-0.0001176 & 0\\-0.0004413 & 0\end{bmatrix}$	$\Delta b^{(1)} = \Delta b^{(1)} + \delta^{(2)}$ $\Delta b^{(1)} = 0 + \begin{bmatrix}-0.0001176\\-0.0004413\end{bmatrix}$ $\Delta b^{(1)} = \begin{bmatrix}-0.0001176\\-0.0004413\end{bmatrix}$

Repeat previous steps for next training example	$\Delta W^{(1)} = \begin{bmatrix} -0.0001176 & 0 \\ -0.0004413 & 0 \end{bmatrix} + \delta^{(2)}\left(a^{(1)}\right)^T$	$\Delta b^{(1)} = \begin{bmatrix} -0.0001176 \\ -0.0004413 \end{bmatrix} + \delta^{(2)}$
Determine small delta value for layer 2 ($\delta^{(2)}$) (back propagation)	$\delta^{(2)} = \left(\left(W^{(2)}\right)^T \delta^{(3)}\right) \cdot \left(f\left(z^{(2)}\right)\left(1 - f\left(z^{(2)}\right)\right)\right)$ $\delta^{(2)} = \left(\begin{bmatrix} 1 \\ 2 \end{bmatrix}(0.093)\right) \cdot \left(\begin{bmatrix} 0.953 \\ 0.018 \end{bmatrix}\left(1 - \begin{bmatrix} 0.953 \\ 0.018 \end{bmatrix}\right)\right)$ $\delta^{(2)} = \left(\begin{bmatrix} 0.093 \\ 0.186 \end{bmatrix}\right) \cdot \left(\begin{bmatrix} 0.953 \\ 0.018 \end{bmatrix}\left(\begin{bmatrix} 0.047 \\ 0.982 \end{bmatrix}\right)\right)$ $\delta^{(2)} = \left(\begin{bmatrix} 0.093 \\ 0.186 \end{bmatrix}\right) \cdot \left(\begin{bmatrix} 0.044791 \\ 0.017676 \end{bmatrix}\right)$ $\delta^{(2)} = \begin{bmatrix} 0.00417 \\ 0.003288 \end{bmatrix}$	
Determine input values from layer 1 ($a^{(1)}$) (forwards propagation)	$a^{(1)} = x$ $a^{(1)} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$	**N/A**
Update delta values using $\delta^{(2)}$ and $a^{(1)}$	$\Delta W^{(1)} = \begin{bmatrix} -0.0001176 & 0 \\ -0.0004413 & 0 \end{bmatrix} + \begin{bmatrix} 0.00417 \\ 0.003288 \end{bmatrix}\begin{bmatrix} 0 & 1 \end{bmatrix}$ $\Delta W^{(1)} = \begin{bmatrix} -0.0001176 & 0 \\ -0.0004413 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0.00417 \\ 0 & 0.003288 \end{bmatrix}$ $\Delta W^{(1)} = \begin{bmatrix} -0.0001176 & 0.00417 \\ -0.0004413 & 0.003288 \end{bmatrix}$	$\Delta b^{(1)} = \Delta b^{(1)} + \delta^{(2)}$ $\Delta b^{(1)} = \begin{bmatrix} -0.0001176 \\ -0.0004413 \end{bmatrix} - \begin{bmatrix} 0.00417 \\ 0.003288 \end{bmatrix}$ $\Delta b^{(1)} = \begin{bmatrix} -0.00429 \\ -0.00373 \end{bmatrix}$
Perform gradient descent on layer 1 weights using delta values	$W^{(1)} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} - 0.1\left(\begin{bmatrix} -0.0001176 & 0.00417 \\ -0.0004413 & 0.003288 \end{bmatrix}\right)$ $W^{(1)} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} - \begin{bmatrix} -0.00001176 & 0.000417 \\ -0.00004413 & 0.0003288 \end{bmatrix}$ $W^{(1)} = \begin{bmatrix} 1.00001176 & 1.999583 \\ 3.00004413 & 3.9996712 \end{bmatrix}$	$b^{(1)} = \begin{bmatrix} 1 & -1 \end{bmatrix} - 0.1\left(\begin{bmatrix} -0.00429 \\ -0.00373 \end{bmatrix}\right)$ $b^{(1)} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} - \left(\begin{bmatrix} -0.000429 \\ -0.000373 \end{bmatrix}\right)$ $b^{(1)} = \begin{bmatrix} 1.000429 \\ 0.999571 \end{bmatrix}$
Repeat above steps for the next layer (layer 2) of weights	$W^{(2)} = W^{(2)} - \frac{a}{N}\left(\Delta W^{(2)}\right)$	$b^{(2)} = b^{(2)} - \frac{a}{N}\left(\Delta b^{(2)}\right)$

Determine delta values for the first training sample	$\Delta W^{(2)} = \Delta W^{(2)} + \delta^{(3)}\left(a^{(2)}\right)^T$	$\Delta b^{(2)} = \Delta b^{(2)} + \delta^{(3)}$
Determine small delta value for layer 3 ($\delta^{(3)}$) (back propagation)	$\delta^{(3)} = -\left(y - f\left(z^{(3)}\right)\right) f'\left(z^{(3)}\right)$ $\delta^{(3)} = -\left(y - f\left(W^{(2)}a^{(2)} + b^{(2)}\right)\right) f'\left(z^{(3)}\right)$ $\delta^{(3)} = -\left(1 - f\left(\begin{bmatrix}1 & 2\end{bmatrix}\begin{bmatrix}0.881\\0.731\end{bmatrix} + [1]\right)\right) f'\left(z^{(3)}\right)$ $\delta^{(3)} = -(1 - 0.966)\left(0.966(1 - 0.966)\right)$ $\delta^{(3)} = -0.00112$	
Determine input values from layer 2 ($a^{(2)}$) (forwards propagation)	$a^{(2)} = f\left(z^{(2)}\right)$ $a^{(2)} = f\left(W^{(1)}a^{(1)} + b^{(1)}\right)$ $a^{(2)} = f\left(\begin{bmatrix}1 & 3\\2 & 4\end{bmatrix}\begin{bmatrix}1\\0\end{bmatrix} + \begin{bmatrix}1\\-1\end{bmatrix}\right)$ $a^{(2)} = f\left(\begin{bmatrix}1\\2\end{bmatrix} + \begin{bmatrix}1\\-1\end{bmatrix}\right)$ $a^{(2)} = f\left(\begin{bmatrix}2\\1\end{bmatrix}\right)$ $a^{(2)} = \begin{bmatrix}0.881\\0.731\end{bmatrix}$	**N/A**
Update delta values using $\delta^{(3)}$ and $a^{(2)}$	$\Delta W^{(2)} = 0 + -0.00112\begin{bmatrix}0.881 & 0.731\end{bmatrix}$ $\Delta W^{(2)} = \begin{bmatrix}-0.000987 & -0.000819\end{bmatrix}$	$\Delta b^{(2)} = -0.034$
Repeat previous steps for next training example	$\Delta W^{(2)} = \begin{bmatrix}-0.000987 & -0.000819\end{bmatrix} + \delta^{(3)}\left(a^{(2)}\right)^T$	$\Delta b^{(2)} = -0.034 + \delta^{(3)}$
Determine small delta value for layer 3 ($\delta^{(3)}$) (back propagation)	$\delta^{(3)} = -\left(y - f\left(z^{(3)}\right)\right) f'\left(z^{(3)}\right)$ $\delta^{(3)} = -\left(y - f\left(W^{(2)}a^{(2)} + b^{(2)}\right)\right) f'\left(z^{(3)}\right)$ $\delta^{(3)} = -\left(0 - f\left(\begin{bmatrix}1 & 2\end{bmatrix}\begin{bmatrix}0.953\\0.018\end{bmatrix} + 1\right)\right) f'\left(z^{(3)}\right)$ $\delta^{(3)} = -\left(0 - f(1.989)\right)\left(f(1.989)(1 - f(1.989))\right)$ $\delta^{(3)} = -(0 - 0.880)\left(0.880(1 - 0.880)\right)$ $\delta^{(3)} = 0.093$	
Determine input values from layer 2 ($a^{(2)}$) (forwards propagation)	$a^{(2)} = f\left(z^{(2)}\right)$ $a^{(2)} = f\left(W^{(1)}a^{(1)} + b^{(1)}\right)$ $a^{(2)} = f\left(\begin{bmatrix}1 & 3\\2 & 4\end{bmatrix}\begin{bmatrix}0\\1\end{bmatrix} + \begin{bmatrix}1\\-1\end{bmatrix}\right)$ $a^{(2)} = f\left(\begin{bmatrix}3\\4\end{bmatrix} + \begin{bmatrix}1\\-1\end{bmatrix}\right)$ $a^{(2)} = f\left(\begin{bmatrix}3\\-4\end{bmatrix}\right)$	**N/A**

	$a^{(2)} = \begin{bmatrix} 0.953 \\ 0.018 \end{bmatrix}$	
Update delta values using $\delta^{(3)}$ and $a^{(2)}$	$\Delta W^{(2)}$ $= [-0.000987 \quad -0.000819]$ $+ 0.093[0.953 \quad 0.018]$ $\Delta W^{(2)}$ $= [-0.000987 \quad -0.000819]$ $+ [0.0886 \quad 0.00167]$ $\Delta W^{(2)} = [0.0886 \quad 0.000851]$	$\Delta b^{(2)} = -0.034 + 0.093$ $\Delta b^{(2)} = 0.059$
Perform gradient descent on layer 2 weights using delta values	$W^{(2)}$ $= \begin{bmatrix} 1 \\ 2 \end{bmatrix} - 0.1([0.0886 \quad 0.000851])$ $= \begin{bmatrix} 1 \\ 2 \end{bmatrix} - ([0.00886 \quad 0.0000851])$ $W^{(2)} = \begin{bmatrix} 0.99114 \\ 1.9999149 \end{bmatrix}$	$b^{(2)} = [1] - \frac{0.2}{2}(0.059)$ $b^{(2)} = [1] - 0.1(0.059)$ $b^{(2)} = [1] - (0.0059)$ $b^{(2)} = [0.9941]$
Final weight values after 1 round of gradient descent	$W^{(1)}$ $= \begin{bmatrix} 1.00001176 & 1.999583 \\ 3.00004413 & 3.9996712 \end{bmatrix}$ $W^{(2)} = \begin{bmatrix} 0.99114 \\ 1.9999149 \end{bmatrix}$	$b^{(1)} = \begin{bmatrix} 1.000429 \\ 0.999571 \end{bmatrix}$ $b^{(2)} = [0.9941]$

3. Would overfitting be more an issue to large training sets or small training sets / with large or small number of parameters to learn?

Overfitting is more of an issue in small training sets because they are easier to converge upon. larger datasets may be more representative of the entire data population and it is harder to overfit a very complex function to a larger dataset. Overfitting is more of a problem with a large number of parameters to learn as the many parameters allow the neural network to emulate a more complex function, whereas a fewer number of parameters will not allow an overly fit function.

4. The regularization had the best increase of performance (approximately 92.1%)