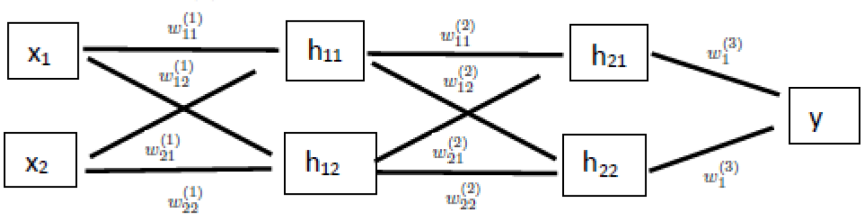
THE HONG KONG UNIVERSITY OF SCIENCE & TECHNOLOGY

MSBD 5012: Machine Learning
Homework 3

Assigned: 20/10/18 **Due Date: 10/11/18**

To submit your work, hand it to the instructor on the due date.

Question 1 Consider the following feedforward neural network with one input layer, two hidden layers, and one output layer. The hidden neurons are ReLU units, while the output neuron is a sigmoid unit.



The weights of the network and their initial values are as follows:

Between input and first hidden: $$\begin{bmatrix} w_{11}^{(1)} & w_{12}^{(1)} \\ w_{21}^{(1)} & w_{22}^{(1)} \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

Between two hidden layers: $$\begin{bmatrix} w_{11}^{(2)} & w_{12}^{(2)} \\ w_{21}^{(2)} & w_{22}^{(2)} \end{bmatrix} = \begin{bmatrix} -1 & -1 \\ 1 & 1 \end{bmatrix}$$

Between second hidden and output: $$\begin{bmatrix} w_1^{(3)} \\ w_1^{(3)} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

For simplicity, assume the units do not have bias parameters. Let there be only one training example $(x_1, x_2, y) = (1, 2, 0)$.

(a) Consider feeding $(x_1, x_2) = (1, 2)$ to the network. What are the outputs of the hidden units? What is the logit $z = u_{21}w_1^{(3)} + u_{22}w_2^{(3)}$ calculated at the output unit? The output of the output unit is a probability distribution $p(y|x_1 = 1, x_2 = 2, \theta)$. What is the distribution?

(b) Next consider backpropagation. The loss function for the training example is $L = -\log p(y = 0|x_1 = 1, x_2 = 2, \theta)$. What is the error $\frac{\partial L}{\partial z}$ for the output unit? What are the errors for the hidden units? What are $\frac{\partial L}{\partial w_{22}^{(2)}}$ and $\frac{\partial L}{\partial w_{22}^{(1)}}$? If we want to reduce the loss on the example, should we increase or decrease the two parameters?

Question 2: Why is the sigmoid activation function not recommended for hidden units, but it is fine for an output unit.

Question 3: What is dropout used for in deep learning? How does it work? Why does it work? Answer briefly.

Question 4: What are the key ideas behind the Adam algorithm for training deep neural networks? Answer briefly.

Question 5: The input of a convolutional layer has shape 27 × 27 × 256 (width, height, depth). The layer uses 384 3 × 3 filters applied at stride 1 with no zero padding. What is the shape of the output of the layer? How many parameters are there? How many float multiplication operations it will take to compute the net inputs of the all the output units?

Question 6: (a) Is it a good idea to apply dropout to a convolutional neural network? If so, which part of the model should we apply dropout to? Answer briefly.

(b) Is it a good idea to apply dropout to a recurrent neural network? If so, which part of the model should we apply dropout to? Answer briefly.

Question 7 In an LSMT cell, $\mathbf{h}^{(t)}$ is computed from $\mathbf{h}^{(t-1)}$ and $\mathbf{x}^{(t)}$ using the following formulae:

$$\begin{aligned}
\mathbf{f}_t &= \sigma(\mathbf{W}_f\mathbf{x}^{(t)} + \mathbf{U}_f\mathbf{h}^{(t-1)} + \mathbf{b}_f) \\
\mathbf{i}_t &= \sigma(\mathbf{W}_i\mathbf{x}^{(t)} + \mathbf{U}_i\mathbf{h}^{(t-1)} + \mathbf{b}_i) \\
\mathbf{o}_t &= \sigma(\mathbf{W}_q\mathbf{x}^{(t)} + \mathbf{U}_q\mathbf{h}^{(t-1)} + \mathbf{b}_q) \\
\mathbf{c}_t &= \mathbf{f}_t \otimes \mathbf{c}_{t-1} + \mathbf{i}_t \otimes \tanh(\mathbf{U}\mathbf{x}^{(t)} + \mathbf{W}\mathbf{h}^{(t-1)} + \mathbf{b}) \\
\mathbf{h}^{(t)} &= \mathbf{o}_t \otimes \tanh(\mathbf{c}_t)
\end{aligned}$$

(a) Intuitively, what are the functions of the forget gate $\mathbf{f}_t$ and the input gate $\mathbf{i}_t$ do? Answer briefly.

(b) Why do we use the sigmoid function for $\mathbf{f}_t$ and $\mathbf{i}_t$, but tanh for the memory cell $\mathbf{c}_t$ and the output $\mathbf{h}^{(t)}$? Answer briefly.