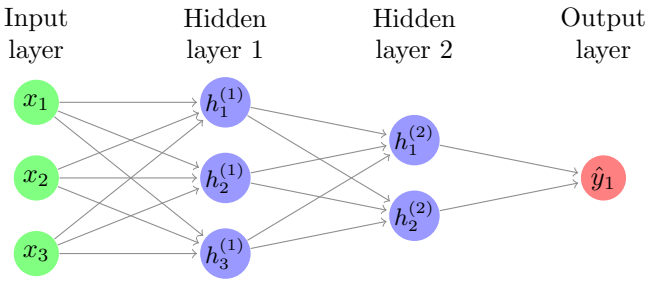
DEEP LEARNING WEEK 3

0. **Use the following data to answer the questions below** The following diagram represents a neural network containing two hidden layers and one output layer. The input to the network is a column vector $x \in R^3$. The activation function used in hidden layers is sigmoid. The output layer doesn't contain any activation function and the loss used is squared error loss $(pred_y - true_y)^2$.



The following network doesn't contain any biases and the weights of the network are given below:

W1=$\begin{bmatrix}1 & 1 & 2\\ 3 & 1 & 1\\ 1 & 2 & 3\end{bmatrix}$ **W2**=$\begin{bmatrix}1 & 1 & 2\\ 3 & 1 & 1\end{bmatrix}$ **W3**=$\begin{bmatrix}2 & 5\end{bmatrix}$

The input to the network is: **x**=$\begin{bmatrix}1\\ 1\\ 1\end{bmatrix}$

The target value y is: **y**=10

1. What is the total number of parameters in the following network?
   a) 15
   b)7
   c)9
   d)17
   Answer:d)
   Solution: Elements of weight and bias matrix represent the parameters of the network. Since the biases are not present in the network counting the elements of weight matrices gives the answer.

2. What is the predicted output for the given input x1 after doing the forward pass? (Choose the option closest to your answer)
   a)7.33
   b)6.92
   c)6.31
   d)8
   Answer:b)
   Solution: Doing the forward pass in the network we get
   h1=W1.x1=$\begin{bmatrix}1 & 1 & 2\\ 3 & 1 & 1\\ 1 & 2 & 3\end{bmatrix} . \begin{bmatrix}1\\ 1\\ 1\end{bmatrix} = \begin{bmatrix}4\\ 5\\ 6\end{bmatrix}$

a1= sigmoid(h1)= $\begin{bmatrix}0.982\\0.993\\0.997\end{bmatrix}$ a

h2=W1.a1= $\begin{bmatrix}1 & 1 & 2\\3 & 1 & 1\end{bmatrix} . \begin{bmatrix}0.982\\0.993\\0.997\end{bmatrix} = \begin{bmatrix}3.969\\4.936\end{bmatrix}$

a2=sigmoid(h2)= $\begin{bmatrix}0.981\\0.992\end{bmatrix}$

y= $\begin{bmatrix}2 & 5\end{bmatrix} . \begin{bmatrix}0.981\\0.992\end{bmatrix} = 6.922$

3. Compute and enter the loss between the output generated by input x and the true output y. (NAT)
Answer: Range(9.38,9.58)
Solution: Loss=$(6.922 - 10)^2 = 9.447$

4. If we call the predicted y as $\hat{y}$ then what is the gradient $dL/d\hat{y}$? (L is the loss function)

a)-5.17
b)-7.52
c)-6.15
d)-7.15
Answer:c)
Solution:$dL/d\hat{y} = 2x(\hat{y} - y)$=2x(6.922-10)=-6.15

5. What is the sum of elements of $\nabla w3$? (Choose the closest value to your answer)

a)-12.9
b)-11.6
c)-10.07
d)-12.14
Answer:d)
Solution:
$[\nabla w31, \nabla w32] = [a_{31} \times dL/d\hat{y}, a_{32} \times dL/d\hat{y}]$=[0.981x-6.156,0.992x-6.156]=[-6.039,-6.106]. The sum of elements of this vector should give the required answer.

6. What is the sum of elements of $\nabla w2$?
Answer: Range(-1.2,-1.4)
Solution: To find $\nabla w2$, find $\nabla h31, \nabla h32$ and then compute $[a21, a22] * [\nabla h31, \nabla h32]^T$.

7. What is the sum of elements of $\nabla w2$?
Answer: Range(-0.04,-0.08)
Solution:

8. The probability of all the events $x_1, x_2, x_2....x_n$ in a system is equal($n > 1$). What can you say about the entropy $H(X)$ of that system?(base of log is 2)

a)$H(X) \leq 1$
b)$H(X) = 1$
c)$H(X) \geq 1$

d)We can't say anything conclusive with the provided information.

Answer: c)
Solution: Since all elements are equal our entropy is of the form
$H(X) = \sum_{i=1}^{n} -p_i.log(p_i) = \sum_{i=1}^{n} -log(1/n) \geq 1$

9. Let p and q be two probability distributions. Under what conditions will the cross entropy between p and q be minimized?

   a) p=q

   b) All the values in p are lower than corresponding values in q

   c) All the values in p are lower than corresponding values in q

   d) $p = 0$ [0 is a vector]

   Answer: a

   Solution: Cross entropy is lowest when both distributions are the same.

10. Suppose we have a problem where data x and label y are related by $y = x^2 + 1$. Which of the following is not a good choice for the activation function in the hidden layer if the activation function at the output layer is linear?

    a)Linear
    b)Relu
    c)Sigmoid
    d)Tan^{-1}(x)

    Answer: a)
    Solution: If we chose the first activation function then the output of the neural network will be a linear function of the data since the network is just doing a combination of weight and biases at every layer, hence we won't be able to learn the non-linear relationship.