Neural Networks and Deep Learning

Week 3: Shallow Neural Networks

1 Version 1: Which of the following are true? (Check all that apply)

Answer: $W_3^{[4]}$ is the column vector of parameters of the fourth layer and third neuron; $a^{[2]}$ denotes the activation vector of the second layer.

Comment: The vector $w_j^[i]$ is the column vector of parameters of the ith layer and jth neuron of that layer; $a^[j]$ denotes the activation function of the jth layer.

1 Version 2: Which of the following are true? (Check all that apply)

Answer: $a^{[2](12)}$ denotes the activation vector of the 2nd layer for the 12th training example; X is a matrix in which each column is one training example; $a^{[2]}$ denotes the activation vector of the 2nd layer; $a_4^{[2]}$ is the activation output by the 4th neuron of the 2nd layer.

2 Version 1: In which of the following cases is the linear (identity) activation function most likely used?

Answer: When working with regression problems.

Comment: In the problems such as predicting the price of a house it makes sense to use the linear activation function as output.

2 Version 2: The tanh activation usually works better than sigmoid activation function for hidden units because the mean of its output is closer to zero, and so it centers the data better for the next layer. True/False.

Answer: True.

Comment: The output of the tanh is between -1 and 1, centers the data which makes the learning simpler for the next layer.

2 Version 3: The tanh activation is not always better than sigmoid activation for hidden units because the mean of its output is closer to zero, and so it centers the data, making learning complex for the next layer. True/False.

Answer: True.

3 Version 1: Which of the following represents the activation output of the second neuron of the third layer applied to the fourth example?

Answer: $a_2^{[3](4)}$.

Comment: The superscript in brackets indicates the layer number, the superscript in parenthesis represents the number of examples, and the subscript the number of the neuron.

3 Version 2: Which of these is a correct vectorized implementation of forward propagation for layer l, where $1 \leq l \leq L$.

Answer: $Z^{[l]} = W^{[l]}A^{[l-1]} + b^{[l]}, A^{[l]} = g^{[l]}(Z^{[l]})$.

Comment: See lecture notes.

4 You are building a binary classifier for recognizing cucumbers (y=1) vs watermelons (y=0). Which one of these activation functions would you recommend using for the output layer?

Answer: sigmoid.

Comment: Sigmoid outputs a value between 0 and 1 which makes it a very good choice for binary classification. You can classify as 0 if the output is less than 0.5 and classify as 1 if the output is more than 0.5. It can be done with tanh as well but it is less convenient as the output is between -1 and 1.

5 Version 1: Consider the following code:

```
#+begin_src python
x = np.random.rand(3, 2)
y = np.sum(x, axis=0, keepdims=True)
#+end_src
```

Which will be y.shape?

Answer: (1,2).

Comment: By choosing the axis = 0 the sum is computed over each column of the array, thus the resulting array is a row vector with 2 entries. Since the option keepdims=True is used the first dimension is kept.

5 Version 2: Consider the following code:

```
A = np.random.randn(4,3)
B = np.sum(A, axis = 1, keepoims = True)
```

What will be B.shape? (If you are not sure, feel free to run this in python to find out).

Answer: (4,1).

Comment: We use keepdim = True to make sure that A.shape is (4,1) and not (4,).

6 Suppose you have built a neural network. You decide to initialize the weights and biases to be zero. Which of the following statements is true?

Answer: Each neuron in the first hidden layer will perform the same computation. So even after multiple iterations of gradient descent, each neuron in the layer will be computing the same thing as other neurons.

7 Version 1: Using linear activation functions in the hidden layers of a multilayer neural network is equivalent to using a single layer. True/False?

Answer: True.

Comment: When the identity or linear activation function g(c)=c is used the output of composition of layers is equivalent to the computations made by a single layer.

7 Version 2: Logistic regression's weights w should be initialized randomly ranther than to all zeros, because if you initialize to all zeros, then logistic regression will fail to learn a useful decision boundary because it will fail to break symmetry. True/False?

Answer: False.

Comment: Logistic regression does not have a hidden layer. If you initialize the weights to zeros, the first example x fed in the logistic regression will output zero but the derivatives of the logistic regression depend on the input x (because there is no hidden layer) which is not zero. So at the second iteration, the weights values follow x's distribution and are different from each other if x is not a constant vector.

8 Version 1: You have built a network using the tanh activation for all the hidden units. You initialize the weights to relative large values, using np.random.randn(...)*1000. What will happen?

Answer: This will cause the inputs of the thanh to also be very large, thus causing gradients to be close to zero. The optimization algorithm will thus become slow. Note: you may also want to take a look at the ReLU activation function.

Comment: tanh becomes flat for large values, this leads its gradient to be close to zero. This slows down the optimization algorithm.

8 Version 2: Which of the following are true about the tanh function?

Answer: For large values the slope is close to zero; The tanh is mathematically a shifted version of the sigmoid function.

Comment: [We can see in the graph of y = tanh(c) how as the values of c increase the curve becomes flatter; You can see the shape of both is very similar but tanh passes through the origin.]

9 Version 1: Consider the following 1 hidden layer neural network:

- x_1, x_2
- 4 neurons
- $a_1^{[2]}$
- $\hat{y}$

Which of the following statements are True? (Check all that apply)

Answer: $W^{[1]}$ will have shape (4,2); $b^{[1]}$ will have shape (4,1); $b^{[2]}$ will have shape (1,1); $W^{[2]}$ will have shape (1,4).

Comment: See lecture notes.

9 Version 2: Consider the following 1 hidden layer neural network:

- x_1, x_2, x_3, x_4
- 3 neurons
- $a_1^{[2]}$
- $\hat{y}$

Which of the following statements are True? (Check all that apply)

Answer: $b^{[1]}$ will have shape (3,1)

Comment: See lecture notes.

10 Version 1: Consider the following 1 hidden layer neural network:

- x_1, x_2, x_3, x_4
- 2 neurons
- $a_1^{[2]}$
- $\hat{y}$

What are the dimensions of $Z^{[1]}$ and $A^{[1]}$?

Answer: $Z^{[1]}$ and $A^{[1]}$ are (2,m).

Comment: The $Z^{[1]}$ and $A^{[1]}$ are calculated oevr a batch of training examples. The number of columns in $Z^{[1]}$ and $A^{[1]}$ is equal to the number of neurons in the first layer.

10 Version 2: In the same network as 9 Version 1, what are the dimensions of $Z^{[1]}$ and $A^{[1]}$?

Answer: $Z^{[1]}$ and $A^{[1]}$ are (4,m).