# DEEP LEARNING – WORKSHEET 3

## Q1 to Q8 are MCQs with only one correct answer. Choose the correct option.

1. Which of the following is true about model capacity (where model capacity means the ability of neural network to approximate complex functions)?
   1. As dropout ratio increases, model capacity increases
   2. As number of hidden layers increase, model capacity increases
   3. As learning rate increases, model capacity increases
   4. None of the above

**Answer:** B) As number of hidden layers increase, model capacity increases

1. Batch Normalization is helpful because?
   1. It is a very efficient backpropagation technique
   2. It returns back the normalized mean and standard deviation of weights
   3. It normalizes (changes) all the input before sending it to the next layer
   4. None of the above

**Answer:** C) It normalizes (changes) all the input before sending it to the next layer

1. What if we use a learning rate that’s too large?
   1. Network will not converge B) Network will converge

C) either A or B D) None of the above

**Answer:** A) Network will not converge

1. What are the factors to select the depth of neural network?
2. Type of neural network (e.g. MLP, CNN etc.)
3. Input data
4. Computation power, i.e. Hardware capabilities and software capabilities
5. Learning Rate
6. The output function to map

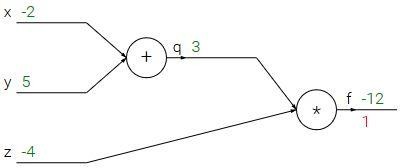
A) 1, 2, 4, 5 B) 2, 3, 4, 5

C) 1, 3, 4, 5 D) All of these

**Answer:** D) All of these

1. Suppose you have inputs as x, y, and z with values -2, 5, and -4 respectively. You have a neuron ‘q’ and neuron ‘f’ with functions:

q = x + y f = q \* z

Graphical representation of the functions is as follows:

What is the gradient of F with respect to x, y, and z? (use chain rule of derivatives to find the solution) A) (3, -4, -4) B) (-3, 4, 4)

C) (-4, -4, 3) D) (4, 4, 3)

Answer: C) (-4,-4,3)

1. Which of the following statement is the best description of early stopping?
   1. Train the network until a local minimum in the error function is reached
   2. Simulate the network on a test dataset after every epoch of training. Stop training when the generalization error starts to increase
   3. Add a momentum term to the weight update in the Generalized Delta Rule, so that training converges more quickly
   4. None of the above

**Answer:** B) Simulate the network on a test dataset after every epoch of training. Stop training when the generalization error starts to increase

1. Which gradient descent technique is more advantageous when the data is too big to handle in RAM simultaneously?
   1. Mini Batch Gradient Descent B) Stochastic Gradient Descent

C) Full Batch Gradient Descent D) either A or B

**Answer:** D) either A or B

1. Consider the scenario. The problem you are trying to solve has a small amount of data. Fortunately, you have a pre-trained neural network that was trained on a similar problem. Which of the following methodologies would you choose to make use of this pre-trained network?
   1. Freeze all the layers except the last, re-train the last layer
   2. Assess on every layer how the model performs and only select a few of them
   3. Fine tune the last couple of layers only
   4. Re-train the model for the new dataset

**Answer:** A) Freeze all the layers except the last, re-train the last layer

## Q9 and Q10 are MCQs with one or more correct answers. Choose all the correct options.

1. Which of the following neural network training challenge can be solved using batch normalization?
   1. Overfitting B) Training is too slow
2. Restrict activations to become too high or low
3. None of these

**Answer:** A) Overfitting, B) Training is too slow

1. For a binary classification problem, which of the following activations may be used in output layer?
   1. ReLU B) sigmoid

C) softmax D) Leaky ReLU

**Answer:** B) sigmoid, C) softmax

## Q11 to Q15 are subjective answer type question. Answer them briefly.

1. What will happen if we do not use activation function in artificial neural networks?

**Answer:** In ANNs, the output of a layer is given when weighted sum of inputs and bias is passed through an activation function. The weighted sum of inputs is given as follows:

*z*  *b* 

Where, b = bias, = ith input and = weight given to ith input.

Whereas, final output is given as follows:

*y*  *f* (*z*)

Where y is the final output and f (z) is activation function.

Now, if we do not use activation function, our model will give weighted sum of inputs and biases as final output which will be of linear form. This output is similar to the one we get in linear regression. In this case our ANN won’t be able to understand more complex functions which are non-linear and hence out ANN will have no special power. Therefore, use of activation function is necessary to build an ANN.

1. How does forward propagation and backpropagation work in deep learning?

**Answer:** **Forward propagation:** The inputs are provided with weights to the hidden layer. At each hidden layer, we calculate the output of the activation function at each node and this further propagates to the next layer till the final output layer is reached. Since we start from the inputs to the final output layer, we move forward and it is called forward propagation.

**Backpropagation:** We minimize the cost function by its understanding of how it changes with changing the weights and biases in a neural network. This change is obtained by calculating the gradient at each hidden layer (using the chain rule of derivatives). Since we start from the final cost function and go back each hidden layer, we move backward and thus it is called backward propagation

1. Explain briefly the following variant of Gradient Descent: Stochastic, Batch, and Mini-batch?

**Answer:** **A) Stochastic Gradient Descent:** Stochastic gradient descent is used to calculate the gradient and update the parameters by using only a single training example.

**B) Batch Gradient Descent:** Batch gradient descent is used to calculate the gradients for the whole dataset and perform just one update at each iteration.

**C) Mini-batch Gradient Descent:** Mini-batch gradient descent is a variation of stochastic gradient descent. Instead of a single training example, mini-batch of samples is used. Mini-batch gradient descent is one of the most popular optimization algorithms.

1. What are the main benefits of Mini-batch Gradient Descent?

**Answer:** Some of the main benefits of using Mini Batch Gradient Descent are as follows:

• It is computationally efficient compared to stochastic gradient descent.

• It improves generalization by finding flat minima.

• It improves convergence by using mini-batches. We can approximate the gradient of the entire training set, which might help to avoid local minima.

1. What is transfer learning?

**Answer:** Transfer learning is a machine learning method where a model developed for a task is reused as the starting point for a model on a second task. Transfer learning make use of the knowledge gained while solving one problem and applying it to a different but related problem. For example, knowledge gained while learning to recognize cars can be used to some extent to recognize trucks. Transfer learning is popular now days because many complex open source Deep learning architectures have been built for different type of tasks such as face recognition, image classification, NLP, etc. So rather than developing a new complex architecture from scratch, we use pre trained architectures and depending on the size of our dataset we may train a few last layers or the complete model again for our purpose.