# DEEP LEARNING – WORKSHEET 3 with Answers

## 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: ( A)

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 )

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: ( B )

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

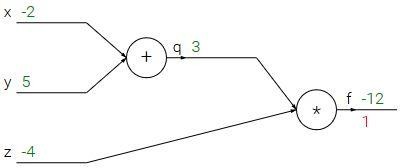
A) 1, 2, 4, 5 B) 2, 3, 4, 5

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

Answer: ( C )

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: ( A )

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: ( C )

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: ( A )

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: ( C )

## 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 )

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 )

## 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: The purpose of the activation function is to introduce non-linearity into the output of a neuron. We know, neural network has neurons that work in correspondence of weight, bias and their respective activation function.

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

Answer: orward propagation (or forward pass) refers to the calculation and storage of intermediate variables (including outputs) for a neural network in order from the input layer to the output layer. We now work step-by-step through the mechanics of a neural network with one hidden layer. This may seem tedious but in the eternal words of funk virtuoso James Brown, you must “pay the cost to be the boss”.

For the sake of simplicity, let us assume that the input example is x∈Rd and that our hidden layer does not include a bias term. Here the intermediate variable is:

(4.7.1)

z=W(1)x,

where W(1)∈Rh×d is the weight parameter of the hidden layer. After running the intermediate variable z∈Rh through the activation function ϕ we obtain our hidden activation vector of length h ,

(4.7.2)

h=ϕ(z).

The hidden variable h is also an intermediate variable. Assuming that the parameters of the output layer only possess a weight of W(2)∈Rq×h , we can obtain an output layer variable with a vector of length q :

(4.7.3)

o=W(2)h.

Assuming that the loss function is l and the example label is y , we can then calculate the loss term for a single data example,

(4.7.4)

L=l(o,y).

According to the definition of L2 regularization, given the hyperparameter λ , the regularization term is

(4.7.5)

s=λ2(∥W(1)∥2F+∥W(2)∥2F),

where the Frobenius norm of the matrix is simply the L2 norm applied after flattening the matrix into a vector. Finally, the model’s regularized loss on a given data example is:

(4.7.6)

J=L+s.

We refer to J as the objective function in the following discussion.

Backpropagation is about understanding how changing the weights and biases in a network changes the cost function. Ultimately, this means computing the partial derivatives ∂C/∂wljk and ∂C/∂blj. ... Backpropagation will give us a procedure to compute the error δlj, and then will relate δlj to ∂C/∂wljk and ∂C/∂blj.

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

Answer: Mini-batch gradient descent is a variation of the gradient descent algorithm that splits the training dataset into small batches that are used to calculate model error and update model coefficients. Implementations may choose to sum the gradient over the mini-batch which further reduces the variance of the gradient.

Stochastic gradient descent (often abbreviated SGD) is an iterative method for optimizing an objective function with suitable smoothness properties (e.g. differentiable or subdifferentiable).

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

Answer: Mini Batch Gradient Descent Batch : A Compromise

* Easily fits in the memory.
* It is computationally efficient.
* Benefit from vectorization.
* If stuck in local minimums, some noisy steps can lead the way out of them.
* Average of the training samples produces stable error gradients and convergence.

1. What is transfer learning?

Answer: Transfer learning is the reuse of a pre-trained model on a new problem. It's currently very popular in deep learning because it can train deep neural networks with comparatively little data.