BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI (RAJ.)
First Semester 2023-24
CS F425
Deep Learning
MIDSEM EXAMINATION [Closed Book]

Date: 14th October 2023 Weightage: 25 % Max. Marks: 50 Duration: 90 minutes

Instructions: *Answer each question on a fresh page and answer all parts of the same question together.*

1. Fill in the blanks: **[1*5 = 5]**
   i. The purpose of using an activation function in a deep neural network is to introduce ________________________________
   ii. If the regularization parameter, alpha, is set to zero, it implies ________________________.
   iii. The L2 regularization works by penalizing the sum of ________________________, while the L1 regularization works by penalizing the sum of ____________________________.
   iv. If there are 100 hidden units in the neural network and the keep probability while using Dropout is 0.4, then ______________ units are dropped.

2. If the input is of size 128x128x3 and the network structure is as indicated in the first column below, calculate the output feature map dimensions for each layer.
   The notation follows the convention:
   • CONV-K-N denotes a convolutional layer with N filters, each them of size KxK. Padding and stride parameters are always 0 and 1, respectively.
   • POOL-K indicates a KxK pooling layer with stride K and padding 0.
   • FC-N stands for a fully-connected layer with N neurons. **[3]**

Layer	**Feature map dimensions**
INPUT	128x128x3
CONV-9-32	
POOL-2	
CONV-5-64	
POOL-2	
CONV-5-64	
POOL-2	
FC-6	6

3. Answer the following questions concisely (write *to-the-point* answers): **[2+2+2+4 = 10]**
   i. While training a neural network, explain what will happen if the learning rate is set: (a) too high, (b) too low?
   ii. What do you understand by “hyperparameters” in a deep neural network and how are they different from the “parameters”? Name two “hyperparameters”.
   iii. You build a neural network classifier with five hidden layers and achieve a training error of 0.2%. You merrily report the “success” to your boss, who asks you about the validation error. To your disbelief, that turns out to be 25%. What could be the reason for such a result? Wasn’t your model “trained well”? Explain.
   iv. Name TWO approaches, along with a 1-2 line description of each, which you can use to obtain a better model in the scenario described in (iii) above. Assume you already have a good amount of data and more data is not required.

4. The figure below shows an input feature map of size 4 x 4 and filter of size 2 x 2 that will be used for computing convolution features. Calculate the output after the filter is applied to the input: (a) using a stride of 1 and “valid padding”, (b) using a stride of 1 and “same padding”. **[3+3]**

1	3	5	7
9	11	13	15
2	4	6	8
10	12	14	16

Input feature map

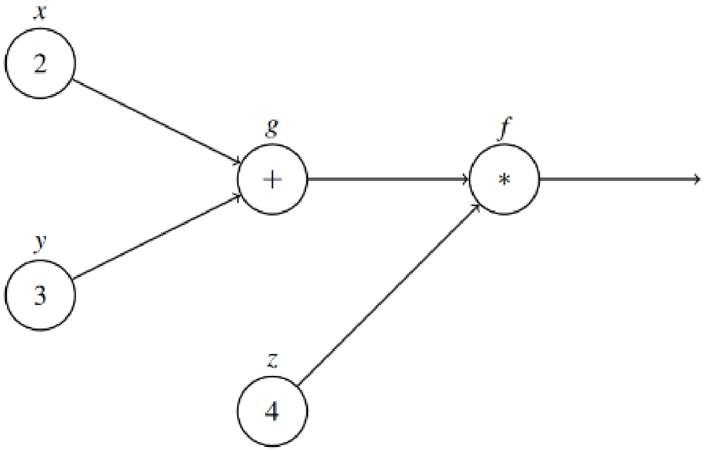
4	6
5	3

filter

5. a) Explain the perceptron training algorithm.
   b) Prove that the perceptron algorithm makes at most $(1/\gamma)^2$ mistakes, where γ is the margin.
   c) Describe the perceptron loss function. Discuss about its differentiability.
   d) Is it possible to represent a XOR function using a perceptron (i.e. 1 node)? If yes, how, if no why not?
   [3+3+1+1=8]

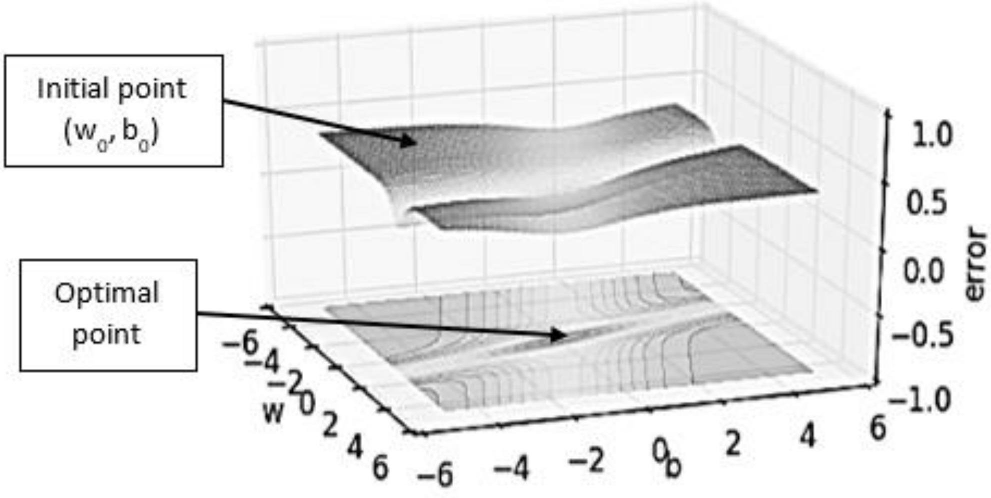
6. Implement the following logic gates using single or multilayer perceptron by setting appropriate weight and bias values. **[2+2+4=8]**
   a) AND gate
   b) OR gate
   c) XOR gate

7. Write the output, as well as the gradients of the function f w.r.t. the inputs at each node by computing a forward and a backward pass on the following computation graph on the right. Show the steps and calculations involved. **[5]**



8. a) Briefly explain: (i) Gradient decent, (ii) Momentum based, and (iii) Nesterov momentum based gradient descent algorithms.

   b) Which of these three gradient descent techniques is the most suitable for minimizing the following error function, when initialized with same initial values of **w** and **b**? Explain. **[3+2=5]**


Initial point
(w_0, b_0)
Optimal point
error
w
b