Indian Institute of Technology Jammu

End Semester Examination 2023-01

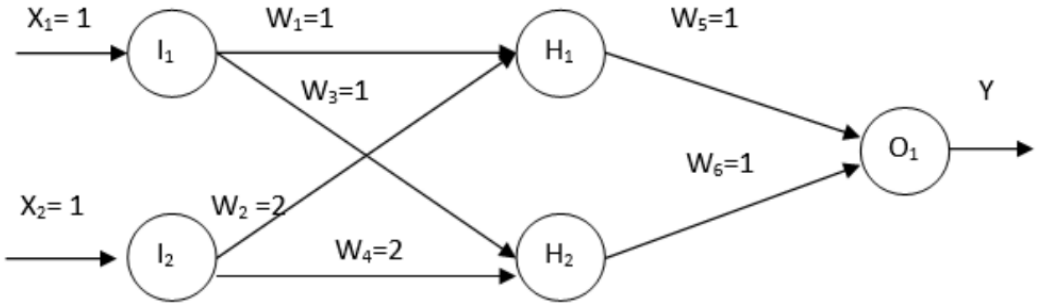
Deep Learning

[Duration: 150 mins] **[Max. Marks: 50]**

1. Consider the following Neural Network with momentum (alpha) = 0.5, learning rate (eta) = 0.24, desired output = 1 and sigmoid activation function.

   [**5 marks**]

   i. Perform one forward pass and calculate the error.
   ii. Calculate the updated weights for w5 and w6 using backpropagation.



2. You are designing a deep learning system to detect driver fatigue in cars. It is crucial that that your model detects fatigue, to prevent any accidents. Which of the following is the most appropriate evaluation metric: Accuracy, Precision, Recall, Loss Value. Explain your choice. [**2 marks**]

3. You want to solve a classification task. You first train your network on 20 samples. Training converges, but the training loss is very high. You then decide to train this network on 10,000 examples. Is your approach to fixing the problem correct? If yes, explain the most likely results of training with 10,000 examples. If not, give a solution to this problem. [**3 marks**]

```python
import numpy as np

def forward_prop(W, a_prev, b):
    z = W*a_prev + b
    a = 1/(1+np.exp(-z)) #sigmoid
    return a
```

4. You are given the following piece of code for forward propagation through a single hidden layer in a neural network. This layer uses the sigmoid activation. Identify and correct the error. [**5 marks**]

5. Why are scaling (γ) and shifting (β) often applied after the standard normalization in the batch normalization layer? [**2 marks**]

6. "Using L1 loss enforces sparsity on the weights of the network." Do you agree with this statement? Why/Why not? [**2 marks**]

7. Compare the number of parameters in a single-layer LSTM with 64 units against a single-layer simple RNN with 64 units, both having an input size of 32. Calculate and compare the total number of parameters in these networks.

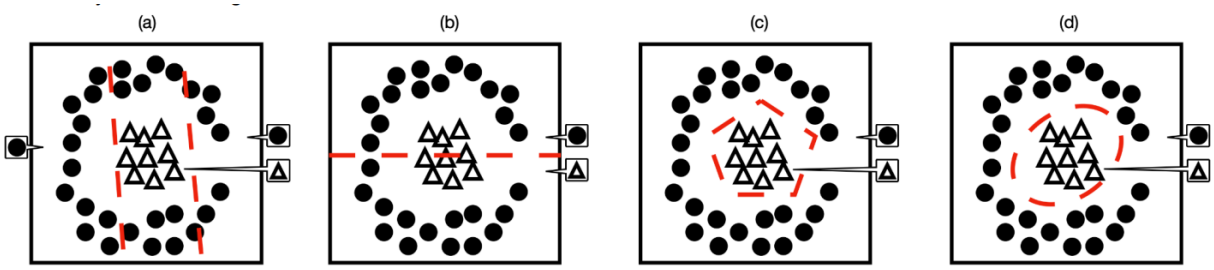
   [**5 marks**]

8. Given a dataset of two-dimensional points (right), each belonging to one of two classes, Judy wants to build a neural network to classify a new point. She will directly feed point coordinates into the network. She found an architecture online, written in Pytorch, but some parts are missing. Here is the code she found:

```python
import torch.nn.functional as F
import torch.nn as nn

class My_Neural_Network(torch.nn.Module):

  def __init__(self,input_dimension,output_dimension=2):

    super(My_Neural_Network, self).__init__()

    self.layer1 = torch.nn.Linear( ___ , 4)                # Input Linear layer  with _____ inputs and 4 outputs
    self.layer2 = torch.nn.Linear( ___ , 3)                # Hidden Linear layer with _____ inputs and 3 outputs
    self.layer3 = torch.nn.Linear( ___ ,output_dimension) # Output Linear layer with _____ inputs and 2 outputs


  def forward(self,x):
    x1 = self.layer1(x)              # Passing input through Linear Layer 1
    x2 = self.layer2(x1)             # Passing output from layer 1 to layer 2
    x3 = self.layer3(x2)             # Passing output from layer 2 to layer 3

    return x3                        # Return final output
```

a) Fill in the missing blanks in the __init__ method so that the forward executes correctly. [**2 marks**]

b) If she trains this neural network correctly and with sufficient data (from the distribution shown above, which of the following classification boundaries might she possibly observe? Choose all that apply, if any, and explain your reasoning. [**2 marks**]



9. Suppose you have an image dataset containing 1000 images, each with a resolution of 128×128 pixels. You are implementing a Vision Transformer (ViT) model for image recognition using multi-head self-attention. The ViT model employs 8 attention heads in its self-attention mechanism. Each image is divided into patches of 16×16 pixels. Calculate the total number of attention weights that will be computed for the entire dataset when processing through this ViT model. [**10 marks**]

10. Imagine employing a GAN architecture to enhance video resolution, utilizing paired high-resolution and low-resolution video data.

Could you outline your architecture, detailing the structures of both the Generator and Discriminator components? Specifically, elaborate on the chosen loss functions for both Generator and Discriminator, explaining the rationale behind their selection.

Suppose you're training the model using the RealBasic dataset and testing it on the IIT Jammu dataset. Regrettably, the performance of the trained model is subpar when applied to the IIT Jammu dataset. What strategies or solutions would you recommend to enhance the model's performance specifically on the IIT Jammu dataset? **[12 Marks]**