Birla Institute of Technology & Science - Pilani, Hyderabad Campus
Second Semester 2020-21
BITS F441: Selected Topics from Computer Science (Deep Learning)
Test 2

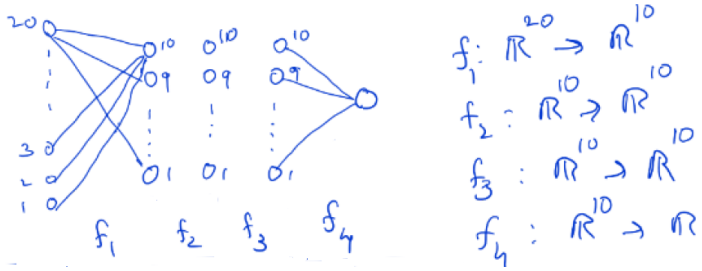
Type: Open (Online) **Time: 30 mins** **Max Marks: 24** **Date: 10.10.2020**
All parts of the same question should be answered together.

BITS Id No:

Name:

Code of Honor: I hereby agree to the fact that I will not give or receive any aid during the examination. This includes, but is not limited to, viewing the answering of others, sharing answers with others, and making unauthorized use of internet while taking the exam. If I violate any of these in any manner, I am ready to receive the consequences of punishment awarded for using unfair means in examination.

Problems Statement: Suppose you are asked to build a regression model with twenty features and one target attribute. You are given 700 training examples and 300 testing examples. You are not given freedom to choose the architecture of the network rather the following is the architecture that you are forced to adhere to.



The above network is a complete 4-partite graph (or) equivalently every node in the ith layer is connected to every node in the (i+1)th layer.
This problem statement should be used to answer first question to fifth question.

1. Suppose f1, f2, f3, f4 are taken to be linear transformations and let the corresponding matrices be W1, W2, W3, W4. Find out the number of parameters of the model. [2 marks]
Sol: (20*10) + (10*10) + (10*10) + 10 = 410
2. As part of design decisions, propose four activation functions that can be considered for the hidden units. [2 marks]
Sol: Sigmoid, tanh, Relu, softplus, hard tanh.
3. As discussed in class, what is the distribution followed by the target attribute given the feature vector. What is the expectation of that distribution? [3 marks]
Sol: Given the feature vector, the target attribute follows normal distribution and expectation of the distribution is the predicted target attribute.

4. As discussed in the class, the cost function is given by

$$J(\theta) = -\mathbb{E}_{x,y \sim \hat{p}_{data}} \log p_{model}(y|x)$$

For our problem, what do the following terms refer to? [5 Marks]

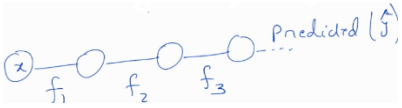
$$\mathbb{E}_{x,y \sim \hat{p}_{data}}$$

$$\log p_{model}(y|x)$$

Sol: The first term refers to average of log $p_{model}(y_i/x_i)$ where (x_1,y_1), (x_2,y_2), . . , (x_n,y_n) are 'n' training examples.
The second term refers to log likelihood of target variable given the feature vector.

5. With an example, explain how does saturation of cost functions impacts learning of the model parameters using gradient methods. [4 Marks]
Note: No need to derive anything here.
Sol: The typical sigmoid function in the output layer saturates when a positive example is misclassified and the predicted outcome is a value very close to 0 and vice versa. The delta changes in weights will not make much difference in the error and hence takes longer time to converge.

6. Imagine you are building a regression model for a problem with only one feature. The training examples are (1, 6), (2, 8) and (4, 12). The network is a very simple network as shown below: [8 Marks]



Though it is very uncommon, the activation function at each unit is taken to be the identity function.

f_1, f_2, f_3 are linear transformations i.e., $f_i(y) = (w_i)y$ for i = 1, 2, 3.

Suppose you are applying stochastic gradient descent algorithm to learn the parameters.

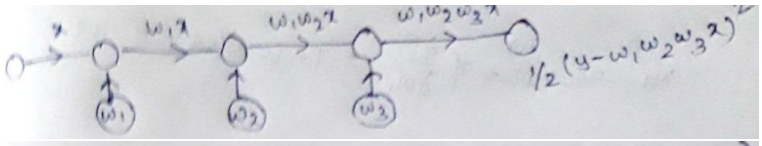
w_1, w_2, w_3 are initialized to be 1, 1, 1 at the start of the algorithm.

The learning parameter, eta, is taken to be 0.1.

The first training example considered by the algorithm is (1, 6).

Find out the weights of w_1, w_2, w_3 at the end of first iteration.

Sol:



Initial weights $(\omega_1^{(1)}, \omega_2^{(1)}, \omega_3^{(1)}) = (1,1,1)$

$$\frac{\partial E}{\partial \hat{y}} = \frac{1}{2}(y-\hat{y})^2 = \frac{1}{2}\times 2(y-\hat{y})(-1) = -(y-\hat{y})$$

$$= -(6-(1\times 1\times 1\times 1)) = -5$$

By back propagation algo;

$$\frac{\partial E}{\partial \omega_3} = \frac{\partial E}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial \omega_3} = (-5)(\omega_1\omega_2 x) = -5\times 1 = -5$$

$$\frac{\partial E}{\partial \omega_1\omega_2 x} = \frac{\partial E}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial \omega_1\omega_2 x} = (-5)(\omega_3) = -5\times 1 = -5$$

$$\frac{\partial E}{\partial \omega_2} = \frac{\partial E}{\partial \omega_1\omega_2 x} \times \frac{\partial \omega_1\omega_2 x}{\partial \omega_2} = (-5)(1) = -5$$

$$\frac{\partial E}{\partial \omega_1 x} = \frac{\partial E}{\partial \omega_1\omega_2 x} \times \frac{\partial \omega_1\omega_2 x}{\partial \omega_1 x} = (-5)\times(1) = -5$$

$$\frac{\partial E}{\partial \omega_1} = \frac{\partial E}{\partial \omega_1 x} \times \frac{\partial \omega_1 x}{\partial \omega_1} = (-5)(1) = -5$$

$$\omega_1^{(2)} = \omega_1^{(1)} - \eta \frac{\partial E}{\partial \omega_1} = 1-(0.1)(-5) = 1+0.5 = 1.5$$

$$\omega_2^{(2)} = \omega_2^{(1)} - \eta \frac{\partial E}{\partial \omega_2} = 1-(0.1)(-5) = 1+0.5 = 1.5$$

$$\omega_3^{(2)} = \omega_3^{(1)} - \eta \frac{\partial E}{\partial \omega_3} = 1-(0.1)(-5) = 1+0.5 = 1.5$$