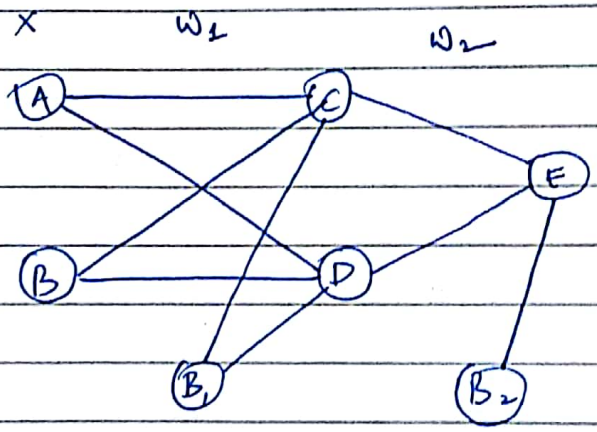
Evergreen
Page No.
Date : / /201

# THEORY ASSIGNMENT

## QUESTION 1

XOR f^n cannot be modelled using a neural network with linear activation function. As a neural network with linear activation fn is equal to a linear model.




Input at hidden layer 1.

$= w_1 X$

Output at hidden layer 1

$= w_1 X + B_1$

Input at output layer

$= w_2 ( w_1 X + B_1 )$

Scanned by CamScanner

Evergreen
Page No.
Date : / /201

Output at output layer

$= W_2(W_1 X + B_1) + B_2$

$= W_1 W_2 X + W_2 B_1 + B_2$

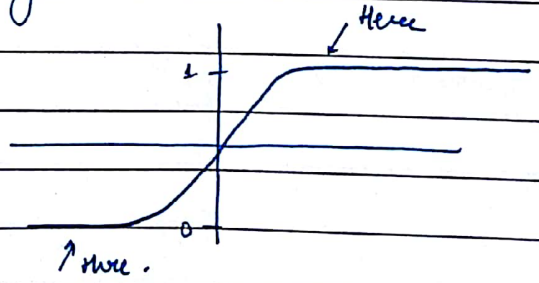
$= \underbrace{W_1 W_2}_{W^*} X + \underbrace{W_2 B_1 + B_2}_{B^*}$

$= W^* X + B^*$ which is equivalent to linear models.

## QUESTION 2

One of the possible reason that the model didn't get trained is because of 'vanishing gradient problem' of sigmoid function.

Vanishing Gradient Problem occurs when a sigmoid fn enters a region where slope difference is negligible. ie




If ReLU fn is used instead of sigmoid, vanishing gradient problem won't occur but data explosion can occur and also dead unit problem can occur.

Scanned by CamScanner

Evergreen
Page No.
Date : / /201

Solution

Possible solutions:.

- Scale down input by a factor for Vanishing Gradient Problem.
- using leaky Relu or normalising input data.

## QUESTION 3

Let

$y$ = target

$\hat{y}$ = predicted

$a = \sigma(z)$

$$\sigma(z) = \frac{1}{1+e^{-z}} \quad \text{(Sigmoid)}$$

$$\hat{y} = e(z) = \sigma(wx+b)$$

Now

$$MSE = \frac{1}{2}(y-\hat{y})$$

By diff.

$$\frac{d(MSE)}{dw} = -(y-\hat{y})\frac{\partial(\hat{y})}{\partial w} \quad \text{——4)}$$

$$\frac{\partial(\hat{y})}{\partial w} = \frac{\partial(\sigma(z))}{\partial w}.$$

$$= \frac{e^{-z}}{(1+e^{-z})^2} \cdot x$$

Scanned by CamScanner

$$= \sigma(z)(1-\sigma(z)).x \qquad \text{——— (2)}$$

Using (1) and (2)

$$\frac{\partial MSE}{\partial w} = (\hat{y}-y)\,\sigma(z)(1-\sigma(z)).x \qquad \text{——— (3)}$$

For Cross Entropy Cost Fn.

$$C = -[y \log \hat{y} + (1-y)\log(1-\hat{y})]$$

$$\frac{\partial C}{\partial w} = -\left[\frac{y}{\hat{y}}\frac{\partial(\hat{y})}{\partial w} + \left(\frac{(y-1)}{(1-\hat{y})}\frac{\partial(\hat{y})}{\partial w}\right)\right]$$

$$= -\left[\frac{y}{\hat{y}} + \left(\frac{-(1-y)}{(1-\hat{y})}\right)\right]\frac{\partial(\hat{y})}{\partial w}$$

$$= \frac{y}{\hat{y}} + \left[\frac{-(1-y)}{(1-\hat{y})}\right]\sigma(z).(1-\sigma(z))$$

$$= \frac{y-\hat{y}}{\hat{y}(1-\hat{y})}\,\sigma(z)(1-\sigma(z)).x.$$

$$= (\hat{y}-y).x. \qquad \text{——— (4)}$$

Comparing 3 and 4, we can say $\sigma(z)(1-\sigma(z))$ acts as slow down factor.

Scanned by CamScanner