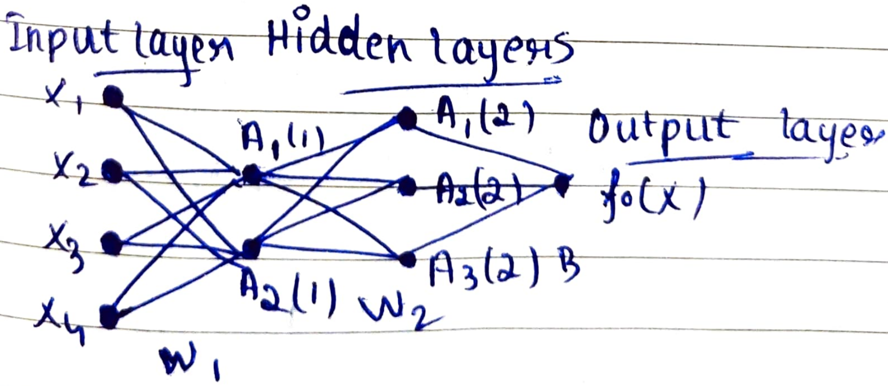
W-4

1. Approach A considers the situation as a kind of regression by prediction just one value, however the digit may not be identified correctly always.

Approach B's method of predicting the probabilities of each class makes more sense because this method tells us how much certain the model is in predicting the class.

Ex. If the written digit is 5, the model showing 50% 5, 40% 6, 10% ... is much better as we can understand that the model itself does not rule out the possibility of 6. In approach A, this is not possible.

2. (a)



Notes

Date / /

(b) Input x_1, x_2, x_3, x_4

1st hidden: $h_{k_1}^{(1)} = \underbrace{w_{11}^{(1)} x_1 + w_{12}^{(1)} x_2 + w_{13}^{(1)} x_3 + w_{14}^{(1)} x_4}_{\text{ReLU()}}$

~~2nd~~

$h_{k_2}^{(1)} = \text{ReLU}(w_{21}^{(1)} x_1 + w_{22}^{(1)} x_2 + w_{23}^{(1)} x_3 + w_{24}^{(1)} x_4)$

2nd hidden: $h_{l_1}^{(2)} = \text{ReLU}(w_{11}^{(2)} h_{k_1} + w_{12}^{(2)} h_{k_2} + b_2^{(2)})$

$h_{l_2}^{(2)} = \text{ReLU}(w_{21}^{(2)} h_{k_1} + w_{22}^{(2)} h_{k_2} + b_2^{(2)})$

$h_{l_3}^{(2)} = \text{ReLU}(w_{31}^{(2)} h_{k_1} + w_{32}^{(2)} h_{k_2} + b_2^{(2)})$

Output: $f(x) = w_1^{(3)} h_{l_1} + w_2^{(3)} h_{l_2} + w_3^{(3)} h_{l_3}$

(c) Considering $w = 1$, $b = 0$

Input: x_1, x_2, x_3, x_4

1st hidden: $h_{k_1}^{(1)} = \text{ReLU}(x_1 + x_2 + x_3 + x_4)$

$h_{k_2}^{(1)} = \text{ReLU}(x_1 + x_2 + x_3 + x_4)$

2nd hidden: $h_{l_1}^{(2)} = \text{ReLU}(h_{k_1} + h_{k_2})$

$h_{l_2}^{(2)} = \text{ReLU}(h_{k_1} + h_{k_2})$

$h_{l_3}^{(2)} = \text{ReLU}(h_{k_1} + h_{k_2})$

Notes

Date / /

output: $f(x) = h_{l_1} + h_{l_2} + h_{l_3}$

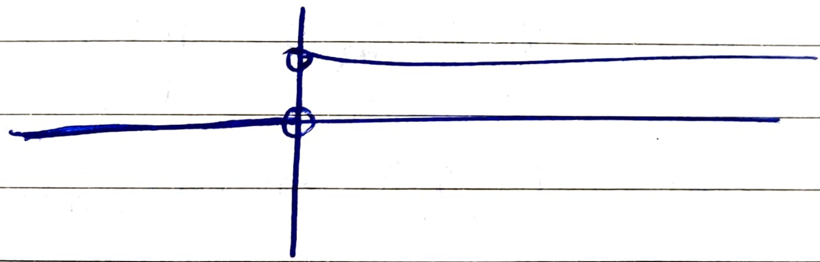
(d) Weights : $4\times 2 + 2\times 3 + 3\times 1$
$= 17$

Biases : $2 + 3 + 1 = 6$

$\underline{\overline{23}}$

3. $f(z) = \begin{cases} z, & z > 0 \\ 0, & z \leq 0 \end{cases} \Rightarrow$ ReLU function

(a) $f'(z) = \begin{cases} 1, & z > 0 \\ 0, & z < 0 \end{cases} \Rightarrow$ Undefined at $z = 0$



(b) $z < 0 \Rightarrow f(z) = 0 \quad f'(z) = 0$

$$\frac{\delta L}{\delta w} = \frac{\delta L}{\delta f} \times \frac{\delta f}{\delta z} \times \frac{\delta z}{\delta w} = 0$$

(c) Since gradient of loss is zero,

$w' = w - \eta \cancel{\frac{\delta L}{\delta w}} = w$, no change in

Notes Date / /

subsequent weights of neuron

4.

(a) $\frac{\delta J_{total}}{\delta w} = \frac{\delta J_{data}}{\delta w} + \frac{\delta}{\delta w}\left(\frac{\lambda}{2} w^2\right)$

the $\frac{\delta J_{total}}{\delta w} = \frac{\delta J_{data}}{\delta w} + \lambda w$

$w_{new} = w_{old} - \eta\left(\frac{\delta J_{total}}{\delta w}\right)$

$= w_{old} - \eta\left(\frac{\delta J_{data}}{\delta w} + \lambda w_{old}\right)$

(b) $w_{new} = w_{old} - \eta \frac{\delta J_{data}}{\delta w} - \eta \lambda w_{old}$

$= w_{old} \underbrace{(1-\eta\lambda)}_{\text{factor}} - \eta \frac{\delta J_{data}}{\delta w}$

$\eta, \lambda > 0$

$1 - \eta\lambda > 0$

(c) L2 regularization involves multiplying w_{old} by a small factor < 1, hence making it a decay.