Can sigmoid do Multiclass classification ?

$n = 3$

OHE

X_1	X_2	A	B	C
—	—	0	1	0
—	—	1	0	0

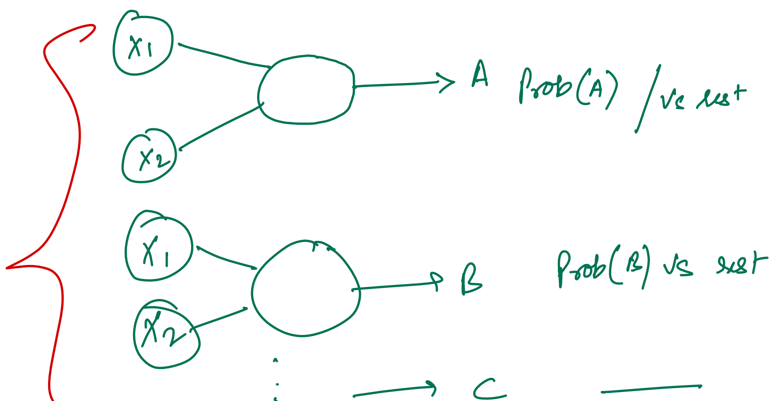
Target variable

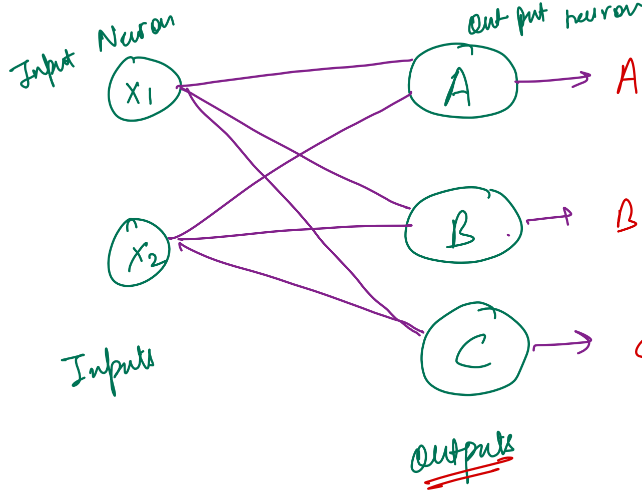
y
B
~~A~~

Prediction

	A'	B'	C'
X	1	0	0
✓	1	0	0

3 neural networks





A vs Rest → $P(A) \to 0.7$

B vs Rest → $P(B) \to 0.5$

C vs Rest → $P(C) \to 0.6$

$= 1$

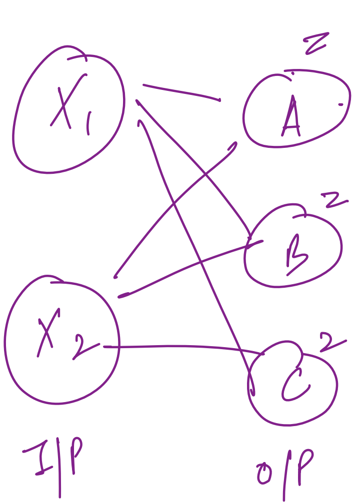
output layer 3 neurons

Input neuron → Taking an input.

Output neuron → $Z = (WX + b) *$ Activation → Sigmoid

Softmax → general way to introduce multiclass classification prob.

Total probability $= 1$



$$= \begin{bmatrix} 12 \\ 8 \\ 4 \end{bmatrix} \quad \begin{bmatrix} 0.8 \\ 0.5 \\ 0.4 \end{bmatrix} \quad \begin{bmatrix} 0.992 \\ 0.007 \\ 0.002 \end{bmatrix}$$

Sigmoid $\Rightarrow \frac{1}{1+e^{-z}}$ Softmax

Squeeze the o/p of least favourable outcomes

$$\text{Sigmoid} = \frac{1}{1+e^{-z}}$$

Softmax $\Rightarrow$ $\frac{z^i}{\Sigma z_i} \Rightarrow \frac{12}{12+8+4} \Rightarrow \frac{12}{24} \Rightarrow 0.5$

$$\frac{8}{8+12+4} = \frac{8}{24} \rightarrow 0.33$$

$$\frac{4}{8+12+4} \Rightarrow \frac{1}{6}\ \cancel{\frac{4}{24}} \rightarrow 0.16$$

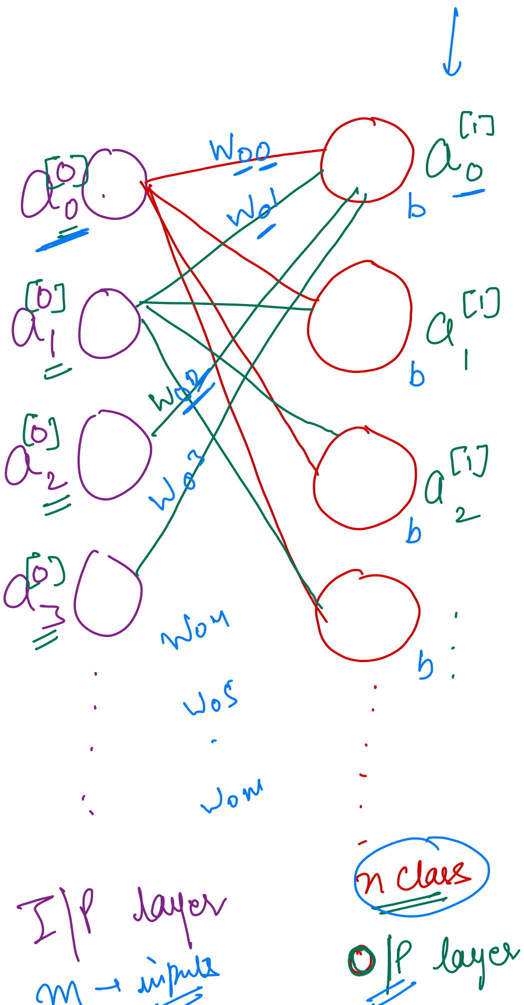
$$\frac{e^{z_i}}{\sum_{n=0}^{2} e^{z_i}}$$

1) e^{z_i} is easily differentiable

2) z is negative $\rightarrow \frac{e^{z_i}}{\sum e^{z_i}} \rightarrow$ No negative value

3) Squeezes the probability of less few outcome

log odds



$Z = WX + b$ → input

$a = Act(Z)$ → output

Act ↓ sigmoid

= Log Reg Unit (LRUs)

$$a_0^{[1]} = Act\left(W_{00}\, a_0^{[0]} + W_{01}\, a_1^{[0]} + W_{02}\, a_2^{[0]} \cdots + b\right)$$

Scale

$$\begin{bmatrix} W_{00} & W_{01} & W_{02} & \cdots & W_{0m} \\ W_{10} & W_{11} & W_{12} & \cdots & W_{1m} \\ \vdots & & & & \\ W_{n0} & W_{n1} & W_{n2} & \cdots & W_{nm} \end{bmatrix} * \begin{bmatrix} a_0^{[0]} \\ a_1^{[0]} \\ a_2^{[0]} \\ a_3^{[0]} \\ \vdots \\ a_m^{[0]} \end{bmatrix} + \begin{bmatrix} b_0 \\ \vdots \\ b_n \end{bmatrix}$$

Layer 1 Calculation of weights

Tight & neat impression → $a^{[l]} = \text{Activation}\left(\underline{W} \times a^{[l-1]} + \underline{B}\right)$

d features

X_1	X_2	X_3
2	3	4

x data points

$(x \times d)$

$(2, 3, 4)$

$\times \; [d \times n]$ weight $+$ bias → $\underline{x \times n}$

$1 \times n$

$n \times 1$

broadcasting

n → A B C

→ $(x \times n)$ output $[0.99, 0.002, 0.001]$

$(x \times d)$ → $(x \times n)$

Sigmoid is a special case of softmax.

2 classes → A & B

$$\frac{e^{z_i}}{\sum_{i=0}^{n} e^{z_i}} \Rightarrow \frac{e^{z_A}}{e^{z_A}+e^{z_B}} = \frac{\cancel{e^{z_A}}}{\cancel{e^{z_A}}}\left(\frac{1}{1+\frac{e^{z_B}}{e^{z_A}}}\right)$$

$$\Rightarrow \frac{1}{1+e^{z_B - z_A}} \Rightarrow \frac{1}{1+e^{-(z_A - z_B)}}$$

$$\Rightarrow \frac{1}{1+e^{-z}}$$

Cost function → Binary Classification

$$J(w,b) = -\frac{1}{N}\left(\sum_{i=1}^{N} y_i \log(a_i) + (1-y_i)\log(1-a_i)\right)$$

prediction

no. of data points

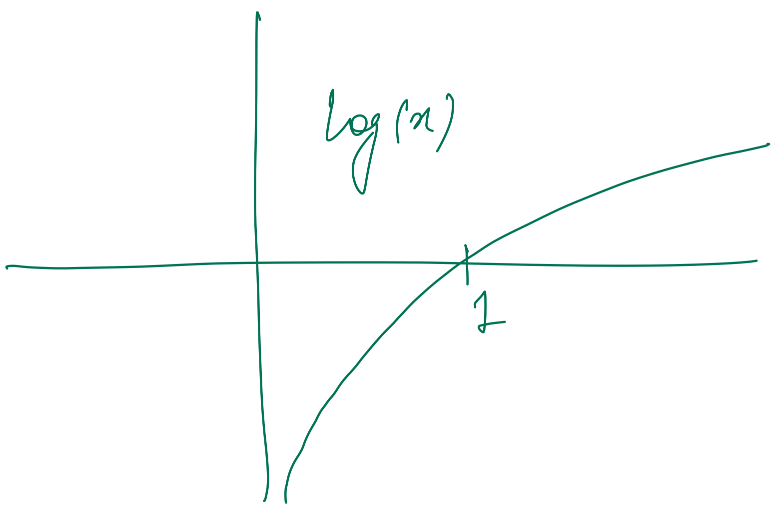
true label

1 data point

$y = 0$ $a = 0$

Loss $\Rightarrow -\frac{1}{1}\left[(0) \times \log(0) + (1-0)\log(1-0) \right]$

$\Rightarrow -\left(0 + 1 \log(1) \right) = 0$



0.01 0.99

totally opp $y=0 \longrightarrow a=1$

$$= -\frac{1}{1}\left(0 \log(\overset{0.99}{1}) + (1-0)\log(1-1)\right) \quad 0.01$$

$$\Rightarrow -\left(0 + \log(0)\right)$$

$$\Rightarrow -(-\infty) = \infty$$

Max loss

Categorical Cross entropy loss (CCE)

$$J(w,b) = -\frac{1}{N}\sum_{i=0}^{n}\sum_{j=0}^{k} y_{ij} \times \log(a_{ij})$$

no. of classes

OHE variable

y-true

i \ j	A	B	C
→	0	1	0
→	1	~~0~~ X	0 X

Predictions

	A'	B'	C'
	X	0	X
	1	~~0~~ X	0 X

$$-\frac{1}{1}\, y_{01} \log(a_{01})$$

$$\Rightarrow -1 \log(0)$$

$$\Rightarrow \infty$$

$$\Rightarrow -\frac{1}{1}\, y_{10} \log a_{10}$$

$$\Rightarrow -\frac{1}{1}\, 1 \log(1) \Rightarrow 0$$

→ X OHE true value / target → Sparse CCE

$CCE\left(y,\ act(XW+b)\right)$

actual

prediction

forward propogation

Cost function (minimize)

using gradient descent

Back propogation