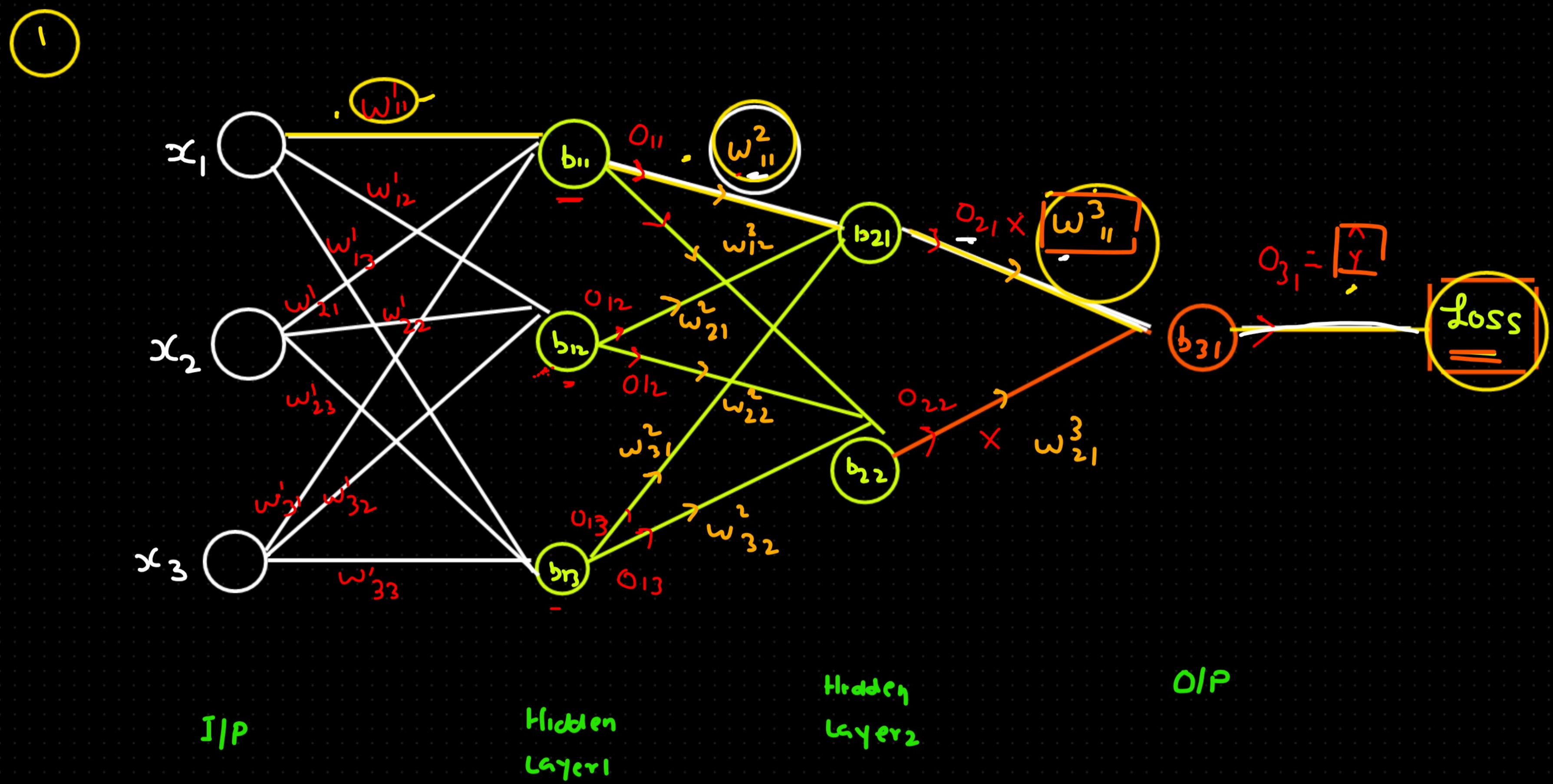


Trainable Parameter $3\times3+3 = 12$ $(3\times2)+2 = 8$ $(2\times1)+1 = 3$ $= 23$

Now let's revise the BP $\Rightarrow$ to update the (trainable Parameter) (weight + Biases)

Optimizer = (Gradient Decent) =

$$\boxed{W_{new} = W_{old} - \eta \frac{\partial L}{\partial w}}$$

$$\frac{\partial L}{\partial w} \Rightarrow \frac{\partial L}{\partial w_{11}^3} \qquad \left(\text{Loss} \rightarrow \underset{O_{31}}{\hat{y}} \rightarrow w_{11}^3\right)$$

$$\frac{\partial L}{\partial w_{11}^3} = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial w_{11}^3} \Rightarrow ?$$

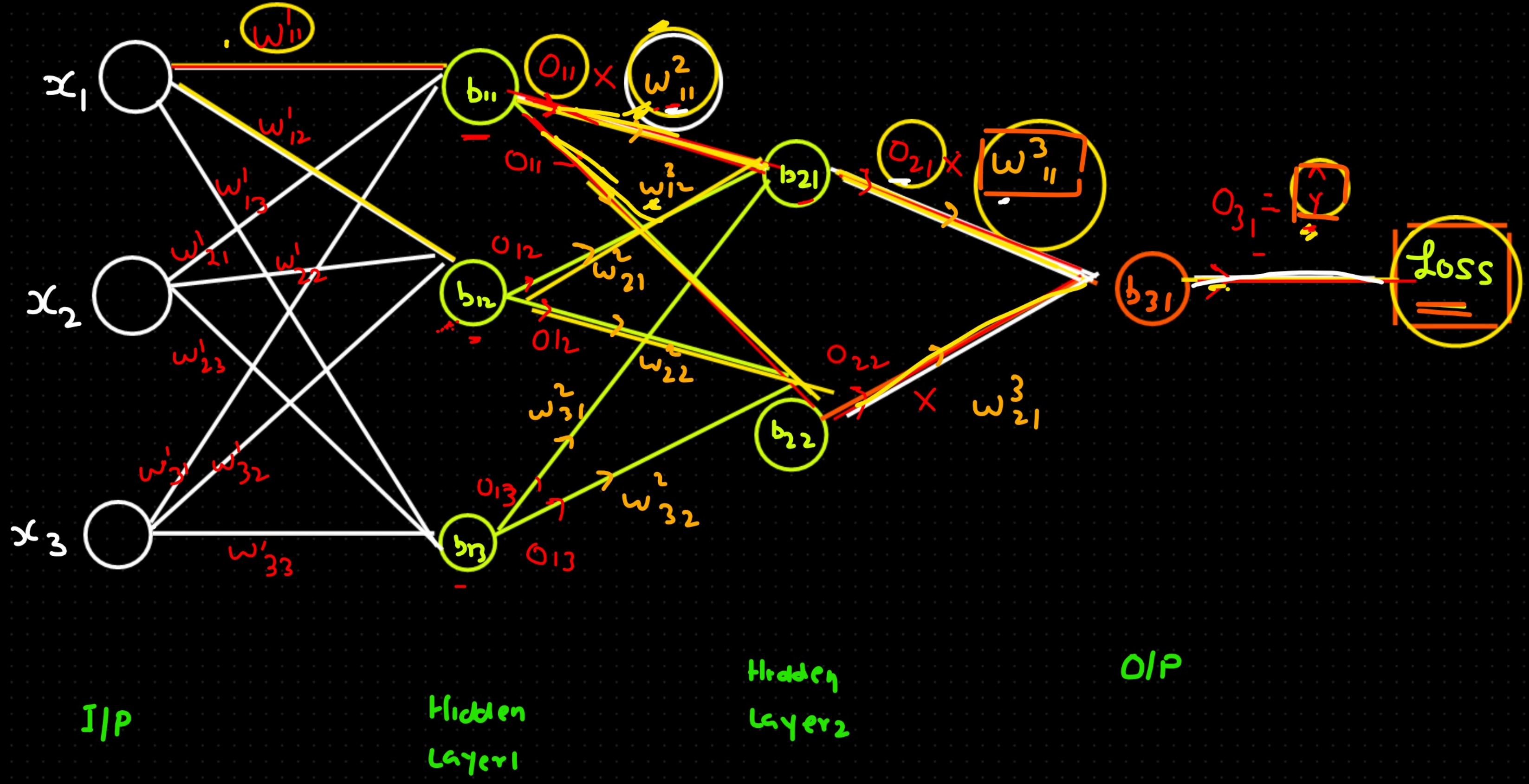
$$\text{Loss} \rightarrow \hat{y}(O_{31}) \rightarrow w_{11}^3 \rightarrow O_{21} \rightarrow w_{11}^2$$

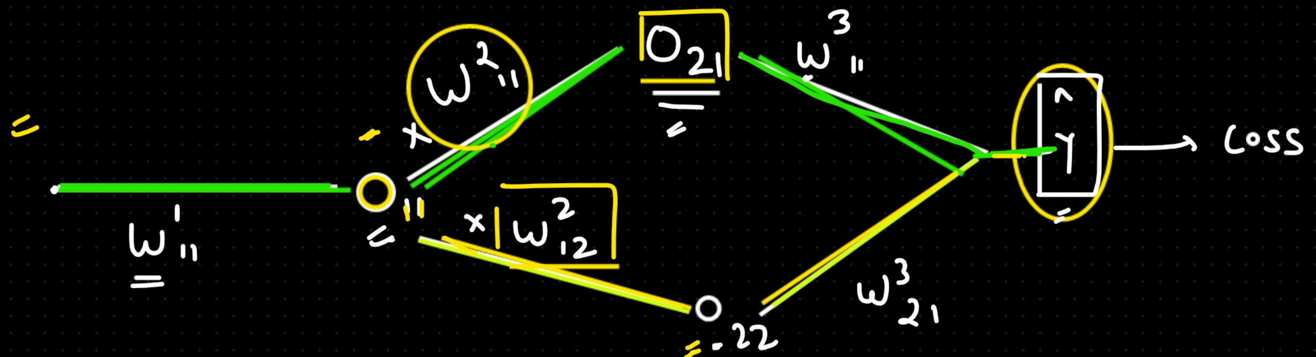
$$\frac{\partial L}{\partial w_{11}^2} = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial O_{21}} \times \frac{\partial O_{21}}{\partial w_{11}^2}$$

(chain rule of differensiation)

$$\text{Loss} \rightarrow \hat{y} \rightarrow w_{11}^3 \rightarrow O_{21} \rightarrow w^2_{11} \rightarrow O_{11} \rightarrow w^1_{11}$$

$$\frac{\partial L}{\partial w^1_{11}} = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial O_{21}} \times \frac{\partial O_{21}}{\partial w^2_{11}} \times \frac{\partial w^2_{11}}{\partial O_{11}} \times \frac{\partial O_{11}}{\partial w^1_{11}}$$





$\boxed{\dfrac{\partial L}{\partial w_{11}^{1}}}$ = will take both path

and for final out will add on both path

$$= \frac{\partial L}{\partial w'_{11}} = \frac{\partial L}{\partial \hat{y}} \left[\frac{\partial O_{31}}{\partial O_{21}} \times \frac{\partial O_{21}}{\partial O_{11}} \times \frac{\partial O_{11}}{\partial w'_{11}} + \frac{\partial O_{31}}{\partial O_{22}} \times \frac{\partial O_{22}}{\partial O_{11}} \times \frac{\partial O_{11}}{\partial w'_{11}} \right]$$

$$\left\{ \begin{array}{l} \text{Loss f/n} \\ \text{activation f/n} \\ \text{optimizer} \end{array} \right\}$$

Loss fn ⇒ Regression ⇒ MSE, MAE, huberloss

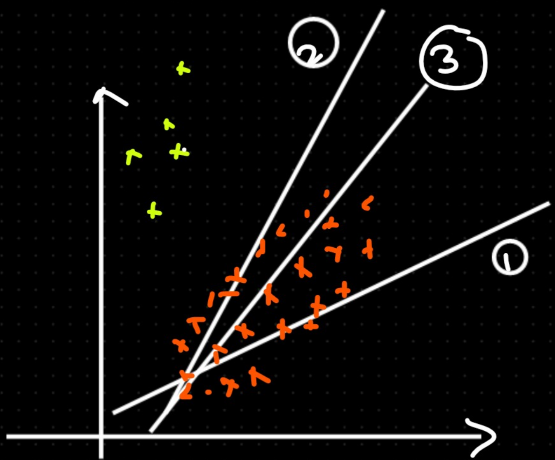
Classification ⇒ Binary, categorical, sparse cross entropy

MSE+MAE

Huber loss

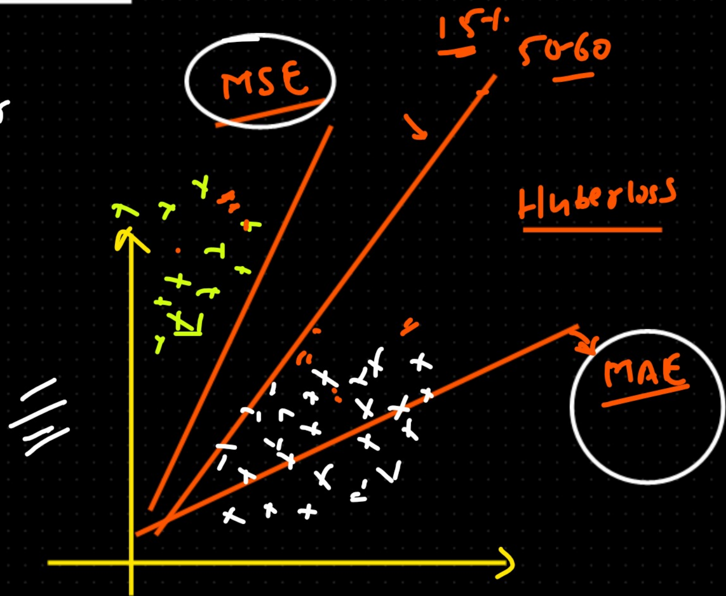
$$= \begin{cases} \frac{1}{2}(y-\hat{y})^2 & \leftarrow \text{this part in presence of outlier} \\ \delta\,|y-\hat{y}| - \frac{1}{2}\delta^2 & \leftarrow \text{if D.P. is outlier} \end{cases}$$

= δ → hyperparameter



⇒ MSE ⇒ Sensitive

MAE ⇒ not Sensitive



defines the loss function piecewise by

$$L_\delta(a) = \begin{cases} \frac{1}{2}a^2 & \text{for } |a| \le \delta, \\ \delta \cdot \left(|a| - \frac{1}{2}\delta\right), & \text{otherwise.} \end{cases}$$

This function is quadratic for small values of a, and linear for large values, with equal values and slopes of the different sections at the two points where $|a| = \delta$. The variable a often refers to the residuals, that is to the difference between the observed and predicted values $a = y - f(x)$, so the former can be expanded to[2]

$$L_\delta(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2 & \text{for } |y - f(x)| \le \delta, \\ \delta \cdot \left(|y - f(x)| - \frac{1}{2}\delta\right), & \text{otherwise.} \end{cases}$$

The Huber loss is the convolution of the absolute value function with the rectangular function, scaled and translated. Thus it "smoothens out" the former's corner at the origin.

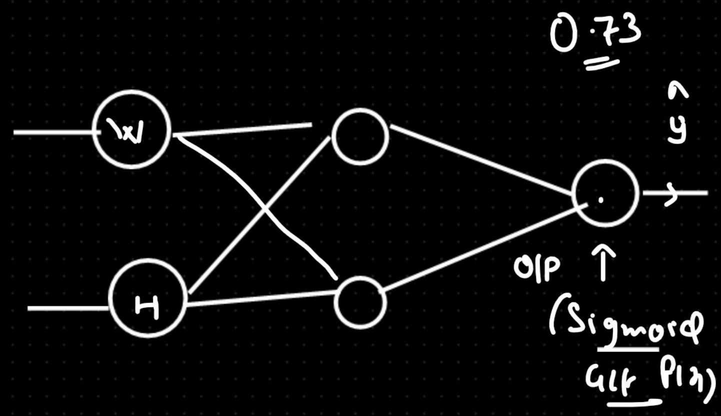
Loss fn for the classification

1. **Binary Cross entropy** ⇒ 2 classes Log Loss ⇒ logistic regression

weight	height	ob/Not ob
55	170	1
60	180	0
65	175	0

0.73

0.25



Single value Loss = $\boxed{-y \log(\hat{y}) - (1-y)\log(1-\hat{y})}$

For the whole data ⇒ Cost fn = $\frac{1}{n}\sum_{i=1}^{n}\left[-y_i \log(\hat{y}_i) - (1-y_i)\log(1-\hat{y}_i)\right]$

Assignment
Perform the derivation (derive) of log loss

$-1\log(0.73) - (1-1)\log(1-0.73)$

$-1 \times \log(0.73) - 0 \times \log(0.27)$

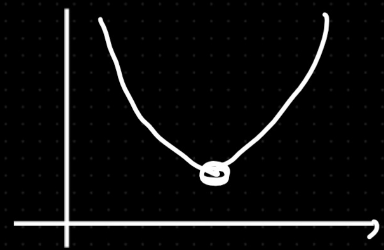
$\boxed{-\log(0.73)} = +(0.13667714)$

$$-0 \times \log(0.25) - (1-0) \times \log(1-0.25)$$

$$0 \qquad -1 \times \log(0.75) = -\log(0.75) = 0.12$$

Advantage

- We can use G.D. for this loss

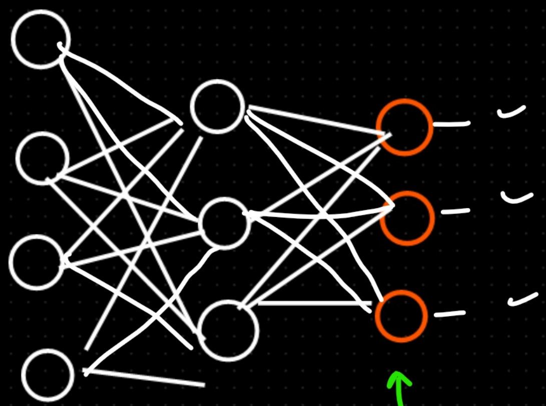


Diadvantase

Categorical cross entropy ⇒ Multiclass classification ⇒ $\boxed{\text{More than 2 class}}$

one hot encoded value for the target class

S_L	S_W	P_L	P_W	class
				Setosa
				verginica
				versicolor



↑

Softmax ⇒ $\dfrac{e^{z_i}}{\sum e^{z_i}}$ ⇒ $\dfrac{e^{z_1}}{e^{z_1}+e^{z_2}+e^{z_3}}$ $\quad \dfrac{e^{z_2}}{\cdots} \quad \dfrac{e^{z_3}}{\cdots}$

$\hat{y} = \underbrace{\text{act}(\text{Summation})}_{\text{Softmax}}$ ⇒ $(y - \hat{y})$ ⇒ Categorical cross entropy

$$\boxed{Loss = -\sum_{i=1}^{K} y_i \log(\hat{y}_i)}$$

$$\boxed{Loss = -\sum_{i=1}^{3} y_i \log(\hat{y}_i)}$$

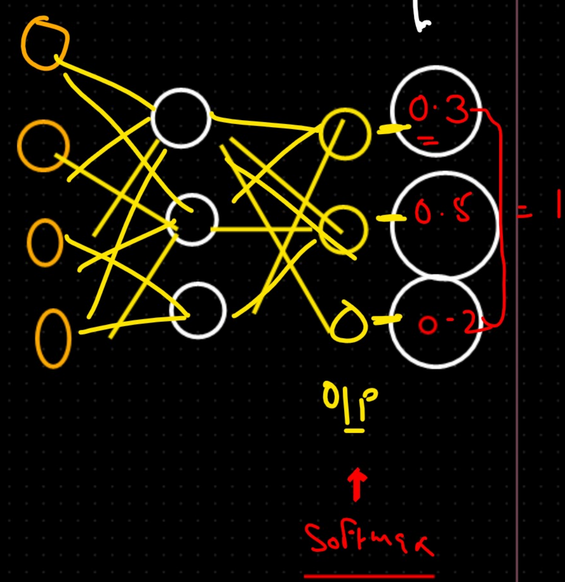
$$\boxed{-y_1 \log(\hat{y}_1) - y_2 \log(\hat{y}_2) - y_3 \log(\hat{y}_3)}$$

one hot encoding

S_L	S_W	P_L	P_W	(target)
-	-	-	-	verginica
-	-	-	-	setosa
-	-	-	-	versicolor

Verginica	setosa	versicolor
1	0	0
0	1	0
0	0	1

this is assumption



$$[0.3 \quad 0.5 \quad 0.2]$$

$$[1 \quad 0 \quad 0]$$

$$[0 \quad 1 \quad 0]$$

$$\boxed{-y_1 \log(\hat{y}_1) - y_2 \log(\hat{y}_2) - y_3 \log(\hat{y}_3)} = -1 \times \log(0.3) - \underbrace{0 \times \log(0.5)}_{0} - \underbrace{0 \times \log(0.2)}_{0}$$

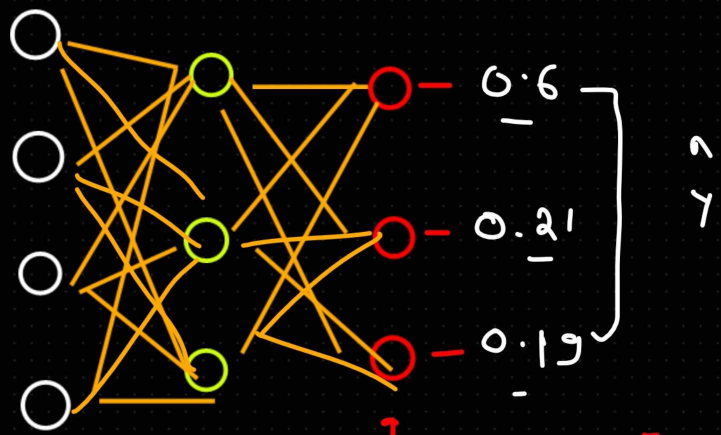
$$= -\log(0.3) = 0.52$$

$$\Rightarrow -0 \times \log(0.3) - 1 \times \log(0.5) - 0 \times \log(0.2)$$

$$= -1 \log(0.5) \Rightarrow -\log(0.5) = 0.30$$

Sparse categorical crossentropy $\Rightarrow$ Fast

S_L	S_W	P_L	P_W	Classes	Label / Integer
-	-	-	-	Setosa	1
-	-	-	-	Versinica	2
-	-	-	-	Versicolor	3



softmax $\Rightarrow \dfrac{e^{z_i}}{\sum e^{z_i}}$

1 st row

$$\begin{bmatrix} 1 & 0 & 0 \\ 0.6 & 0.21 & 0.19 \end{bmatrix} \Rightarrow \text{Cat cross entropy ?}$$

$$\begin{bmatrix} \boxed{1} & 2 & 3 \\ \boxed{0.6} & 0.21 & 0.19 \end{bmatrix}$$

$$-y_1 \log(\hat{y}_1) - \cancel{y_2 \log(\hat{y}_2)} - \cancel{y_3 \log(\hat{y}_3)}$$

$$\Rightarrow -1 \times \log(0.6) \Rightarrow -\log(0.6)$$

$$\begin{bmatrix} 1 & \boxed{2} & 3 \\ \times & & \times \\ 0.6 & \boxed{0.21} & 0.19 \end{bmatrix}$$

Height	Weight	Body size	
170	66	M	1
175	60	S	2
180	75	L	3
160	70	M	1
155	50	L	3

One hot encoding

CCA

$\Rightarrow$ CCA

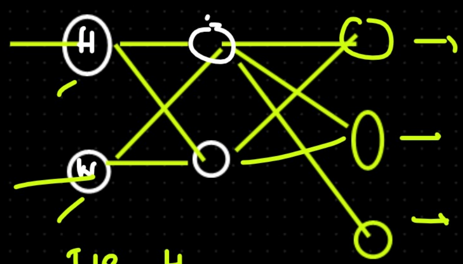
$\Rightarrow$ SCCA

$$- \sum_{i=1}^{3} y_i \log(\hat{y}_i)$$

$$= -y_1 \log(\hat{y}_1) - y_2 \log(\hat{y}_2) - y_3 \log(\hat{y}_3)$$

$$= [1 \quad 0 \quad 0]$$

$$[0.3 \quad 0.5 \quad 0.2]$$



→ 0.3

→ 0.5

→ 0.2

Softmax

$$= -1 \times \log(0.3) - \underbrace{0 \times \log(0.5)}_{0} - \underbrace{0 \times \log(0.2)}_{0}$$

$$= -\log(0.3) \Rightarrow 0.52$$

SCCA

labelling

$\Rightarrow$ $= [\cancel{1} \quad 2 \quad \cancel{3}]$

$\Rightarrow [0.3 \quad 0.5 \quad 0.2]$

SCCA v/s CCA

when the classes is low use CCA

when the classes is high use SCCA

CCA reduce the performance

SCCA increase the performance

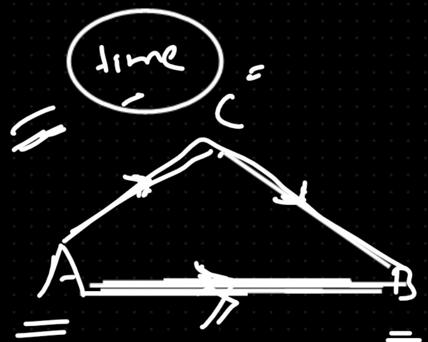
$-1 \times \log(0.5)$

$= -\log(0.5)$

$= 0.69$

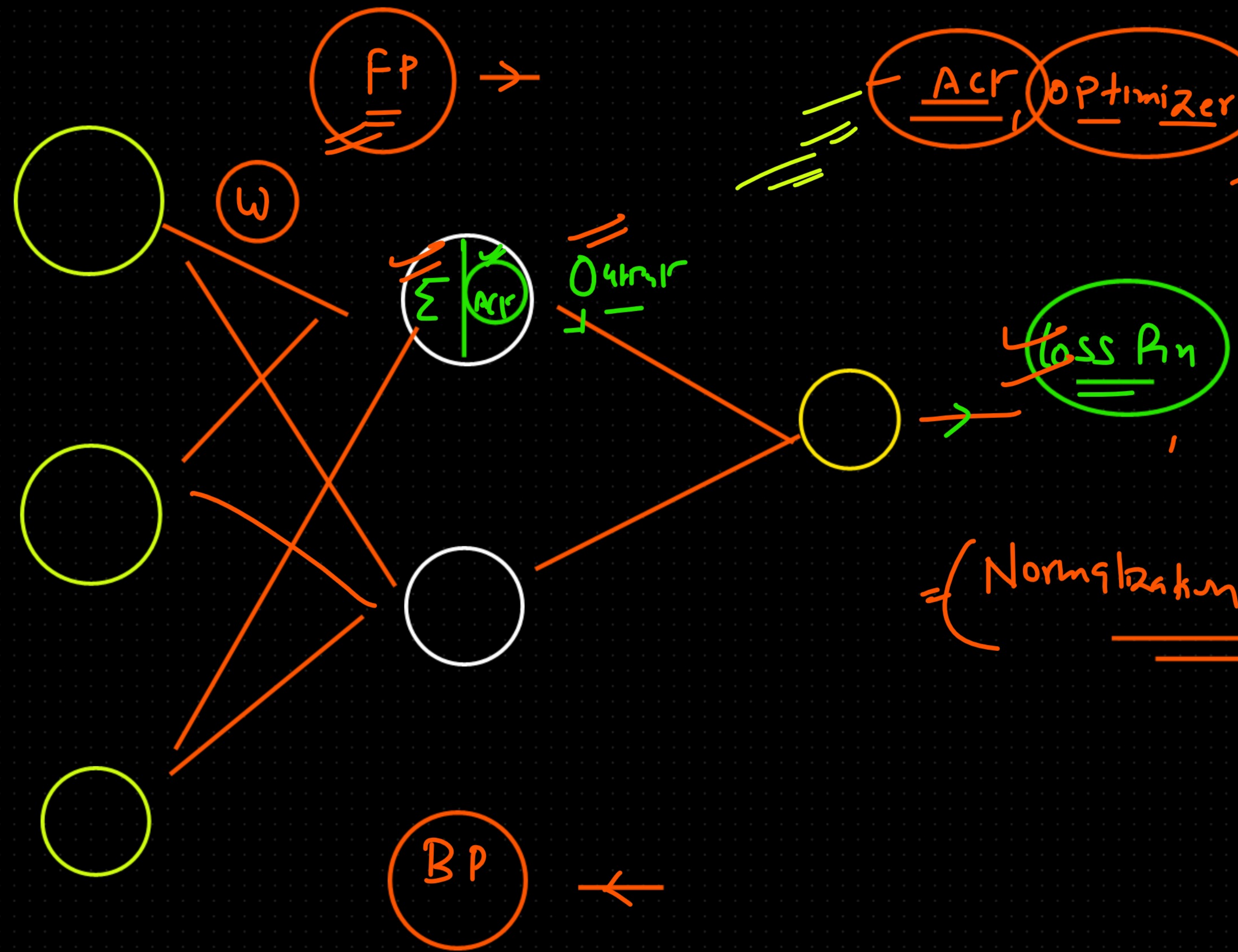
low data + less class ⇒ CCA is fine

huge data + high class ⇒ SSCA is fine



mapping

Simplified



Act, optimizer, weight initialization

(Normalization, Regularization)

Act Fn

Linear activation fn

1) Binary step Activation

$$y = f(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0 \end{cases}$$

2) Linear activation fn

$$y = f(x) = x$$

$$\left.\begin{aligned} f(1) &= 1 \\ f(2) &= 2 \\ f(100) &= 100 \end{aligned}\right\}$$

Non Linear activation fn

1) sigmoid fn = $y = f(x) = \dfrac{1}{1+e^{-x}}$

Logistic fn

2) tanh $\Rightarrow$ $y = f(x) = \dfrac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$

3) Relu $\Rightarrow$ $\boxed{y = f(x) = \max(0, x)}$

4) Prelu or Parametric relu $\Rightarrow$ $f(x) = \max(ax, x)$

5) Softmax $\Rightarrow$ $\text{softmax}(z_i) = \dfrac{\exp(z_i) \text{ or } e^{z_i}}{\sum \exp(z_i) \text{ or } e^{z_i}}$

Regression $\Rightarrow$ Linear A/c f/n or Sigmoid fn

Binary classification $\Rightarrow$ Sigmoid, tanh, (Relu)

multiclass $\Rightarrow$ Softmax

Why, where, how

Optimizer