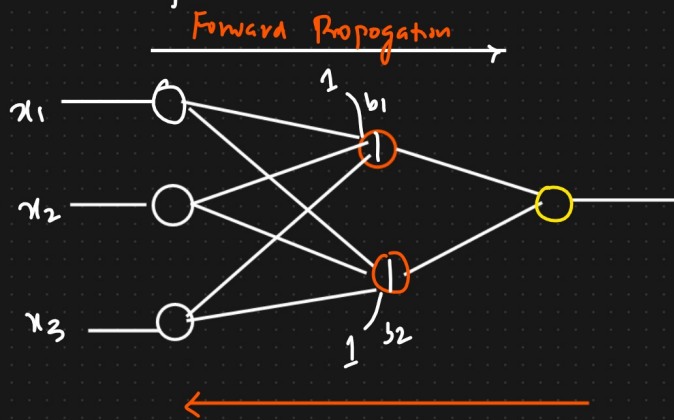


Loss function And Cost Function



$$\text{Cost fn} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

↑ Error

↓ ↓ ↓

Loss = $(y - \hat{y})^2$

Optimizers

Gradient Descent

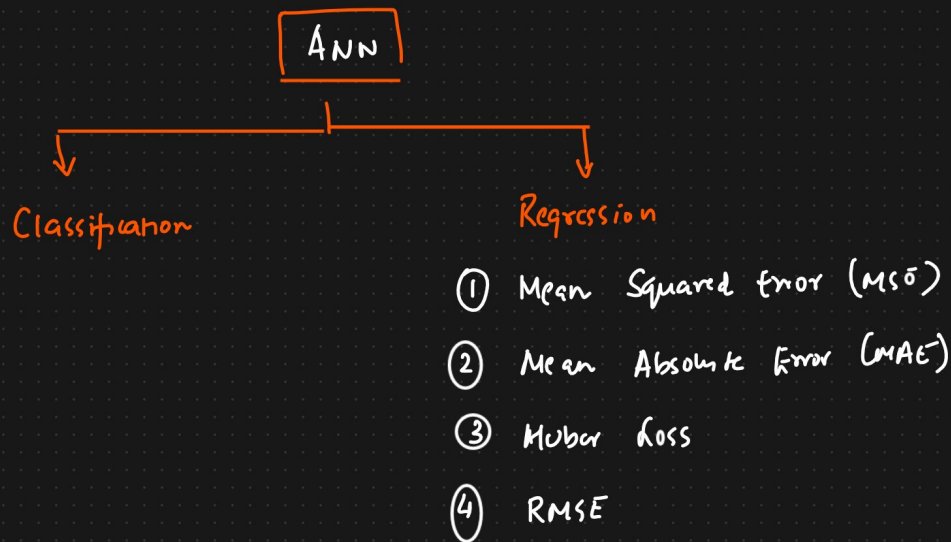
x_1	x_2	x_3	O/P
-	-	-	0
-	-	-	1
-	-	-	0
-	-	-	1

Loss function

$$\text{MSE} = (y - \hat{y})^2$$

Cost function

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$



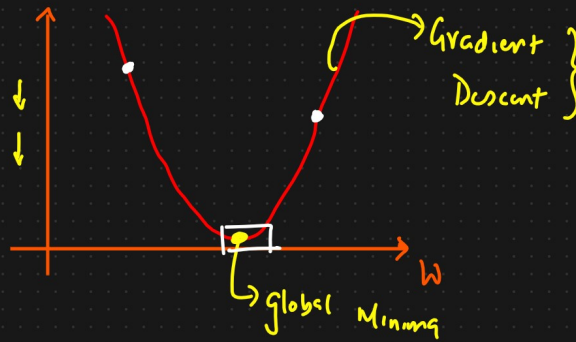
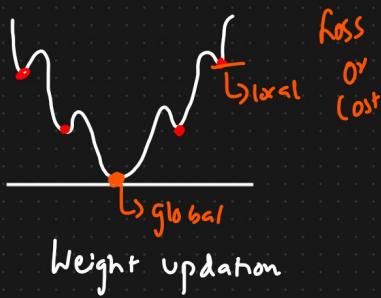
① Mean Squared Error (MSE)

$$\text{Loss function} = (y - \hat{y})^2$$

$$\text{Cost fn} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

↳ Quadratic Equations





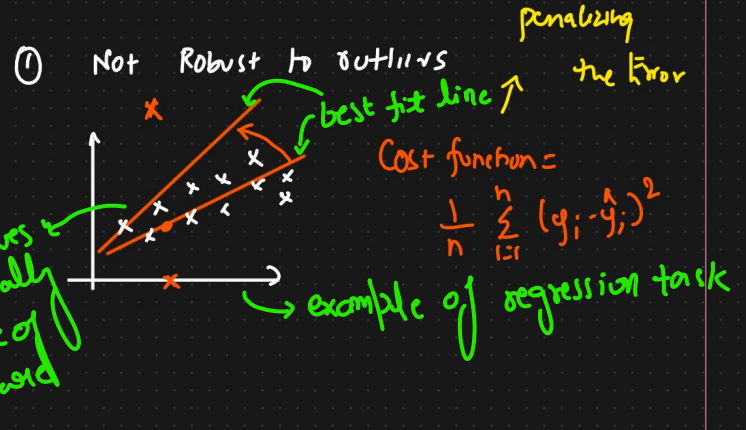
$$w_{\text{new}} = w_{\text{old}} - \eta \left[\frac{\partial L}{\partial w_{\text{old}}} \right]$$

Slope ↑

Advantages

Disadvantages

- ① MSE is Differentiable
- ② It has 1 local or global Minimum
- ③ It Converges faster



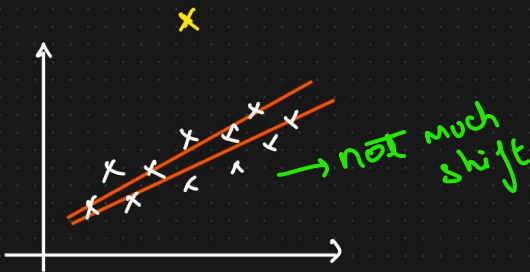
Mean Absolute Error (MAE)

$$\text{loss fn} = |y - \hat{y}|$$

$$\text{Cost fn} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

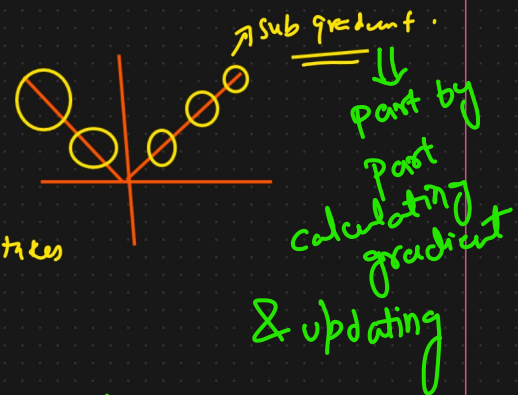
Advantages

- ① Robust to outliers



Disadvantages

- ① Converges usually takes time in MAE



go for this if you have outliers

Huber loss

- ① MSE
- ② MAE

$$\text{Cost fn} = \begin{cases} \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 & \text{if } |y_i - \hat{y}_i| \leq \delta \\ |y_i - \hat{y}_i| & \text{if } |y_i - \hat{y}_i| > \delta \end{cases}$$

MSD ↑ No outlier ↑ Hyperparameter ↑

$$\begin{cases} |y - \hat{y}| & -\frac{1}{2} \delta^2, \text{ otherwise} \end{cases}$$

↓
MAE

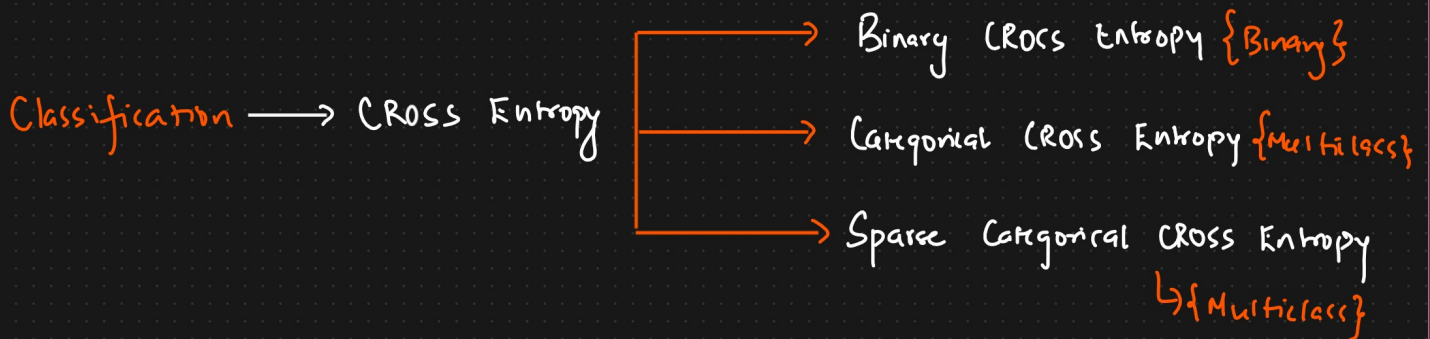
④ RMSE (Root Mean Squared Error)

$$\text{Cost function} = \sqrt{\frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{N}}$$

Advantages

Disadvantages

② Loss or Cost function For Classification Problems



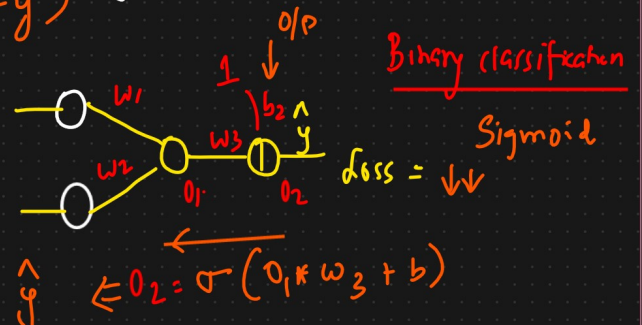
① Binary CROSS Entropy

log loss
↓

$$\text{loss} = -y * \log(\hat{y}) - (1-y) * \log(1-\hat{y})$$

y = Actual value
 \hat{y} = Predicted value

$$\text{loss} = \begin{cases} -\log(1-\hat{y}) & \text{if } y = 0 \\ -\log(\hat{y}) & \text{if } y = 1 \end{cases}$$



$$\hat{y} = \frac{1}{1 + e^{-z}} \Rightarrow \text{Sigmoid Activation function}$$

② Categorical Cross Entropy (Multi-class classification)

ONE → One Hot Encoding

	f_1	f_2	f_3	O/p	$j=1$ Good	$j=2$ Bad	$j=3$ Neutral
→ 2	3	4	Good	[1	0	0]	
→ 5	6	7	Bad	0	1	0	
→ 8	9	10	Neutral	0	0	1	

$C = \text{No. of categories}$
 $i = 1 \text{ to } n$

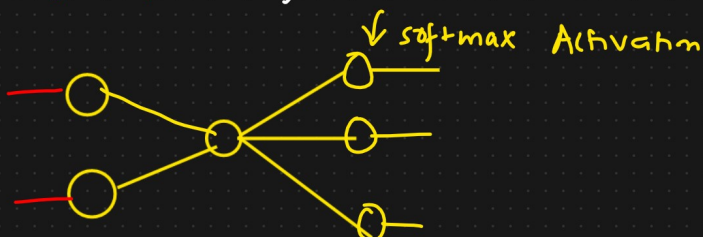
$$d(x_i, y_i) = - \sum_{j=1}^C y_{ij} * \ln(\hat{y}_{ij})$$

Actual value $\Leftarrow y_{ij} = \begin{bmatrix} y_{11} & y_{12} & y_{13} & \dots & y_{1c} \\ y_{21} & y_{22} & y_{23} & \dots & y_{2c} \\ \vdots & \vdots & \vdots & \ddots & \vdots \end{bmatrix}$

$$y_{ij} = \begin{cases} 1 & \text{if the element is in the class} \\ 0 & \text{Otherwise} \end{cases}$$

Prediction $\Leftarrow \hat{y}_{ij} \Rightarrow \text{Softmax Activation} = \text{Sof}(z) =$

$$\frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}}$$



O/p \hat{y}_{ij} = Probabilities

$[0.1, 0.2, 0.3, 0.2, 0.2] \leftarrow 1$
 \downarrow

Categorical $\Rightarrow [0.2, 0.3, 0.5]$

Cross Entropy

\downarrow
 [This also gives the probability of other categories]

② Sparse Categorical Cross Entropy

$\begin{matrix} 0 & 1^{st} & 2^{nd} \\ 0.2 & 0.3 & 0.5 \end{matrix} \rightarrow \text{Categories}$

\downarrow

$2^{nd} \text{ Index} \Rightarrow \text{O/p}$

$\begin{matrix} 0 & 1 & 2^{nd} & 3^{rd} & 4^{th} \\ 0.2 & 0.3 & 0.1 & 0.2 & 0.2 \end{matrix} \Rightarrow \begin{matrix} 1^{th} \text{ Index} \\ \downarrow \\ \text{Category} \Rightarrow \text{O/p} \end{matrix}$

Disadvantage

① Loosing info about the probability of other category.

* ③ Right Combination

Activation applied

\downarrow

Hidden Layers

O/p Layer

Problem Statement

Loss function

- | | | | | |
|---|----------------------|---------|-----------------------|----------------------------|
| ① | Relu or its Variants | Sigmoid | Binary Classification | Binary Cross Entropy |
| ② | Relu or its Variants | Softmax | Multi class | Categorical or Sparse CE |
| ③ | Relu or its Variants | Linear | Regression | MSE, MAE, Huber loss, RMSE |