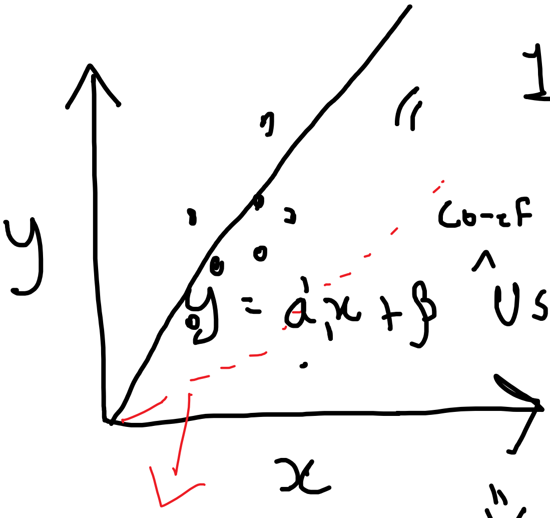
Recap

28 July 2025 20:42



1 feature

Co-ef

→ bias / y-intercept

$y = a_1 x + \beta$ Use x to predict y

Best fit line to capture the relation b/w x and y

$y = w_1 x + w_0$

Cannot be line of best fit, sum of variance is not max

Normalization is a type of {standardization}

↓ different types

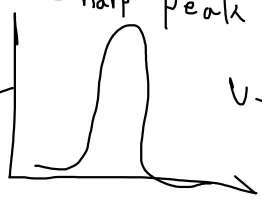
{Normal distribution} ⇐ {Normalize based on how many stdev away your data is from the mean}

{train vs test distributions should be similar}

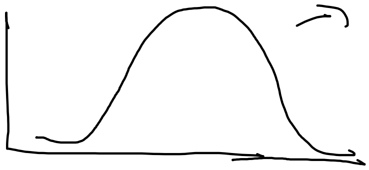
Validate beforehand

Sharp peak

high Kurtosis ←



vs



→ wide peak

$$\left\{ y = w_1 x_1 + w_2 x_2 + w_3 x_3 \cdots\cdots w_d x_d + w_0 \right\}$$

{ Multiple Linear Regression }

$y \Rightarrow$ lifespan

$x_1 \Rightarrow$ Current age

$x_2 \Rightarrow$ Weight

$-5 \Rightarrow$ +1 in age
lifespan -5 down
+2 in age
lifespan -10 down

$$\left\{ \overset{y}{\text{lifespan}} = w_1 \times \text{Current age} + w_2 \times \text{weight} + w_0 \right.$$

How many More years person will live

{ w_1 is negative } $\Rightarrow$ decrease in target

$$\overset{y}{\text{Lifespan}} = -5 \times \text{age}$$

age	lifespan
1	-5
2	-10
3	-15

{ for every 1 unit increase in current age, lifespan reduces by 5 years }

W is positive $\Rightarrow$ for 1 unit increase in my variable x_2, lifespan increase by W units

$$\text{Lifespan} = \underset{\text{positive}}{\underset{\Downarrow}{0.5}} \times \text{weight (Mass)}$$

Weight	lifespan
1	0.5
2	1

positive

$$\left\{ \begin{array}{c|c} 1 & 0.5 \\ 2 & 1 \\ 3 & 1.5 \end{array} \right\}$$

current
Age has negative impact but
weight has positive impact on target (lifespan)

positive weight => target will increase
negative weight => target will decrease

W is 0 => No impact on my target

Which is more important? -> x_1?

Age
Weight
=> Which is more important?

You might initially think, the variable with a higher weight or co-efficient has higher impact

$$\text{Lifespan} = -5 * \text{age} + 0.5 \times \text{mass} + 50$$

$$\left\{ \begin{array}{l} \text{age} \overset{\text{impact}}{\Rightarrow} 5 \\ \text{Mass} \overset{\text{impact}}{\Rightarrow} 0.5 \end{array} \right\}$$ initial thought is age has higher impact

---

$$LF = -5 \times \text{age} + 0.5 \times \text{mass} + 0.00001 \times \text{platelet count}$$

age => 0 to 100

Mass => 2 ...

mass ⇒ 3 to 150 kg

platelet count ⇒ 500 to Millions

Unsure !!

{Feature scaling} ⇒ When feature importance Matters !!

⇓

information is not lost !! } ⇒ interpretation way changes

age ⇒ −2 std, −1.5 std, −0.8 std, 0 std (Mean), 1 std, 2 std

Weight ⇒ Scale is now in std away from Mean

platelets ⇒ scale in std " " "

{you can always convert std back into original}

$$z = \frac{x_i - \bar{x}}{\sigma_x} \Rightarrow \{\sigma_x z + \bar{x} = x_i\}$$

z ⇓ std away from Mean

inverse Scaling

the Way you interpret will change but there is no neg/pos/human impact

Gradient Descent Revision

28 July 2025 22:07

loss function ? $\Big|\, y_{actual} - y_{pred} \,\Big|$

⇓

{function of the error} Minimize !!

⇓

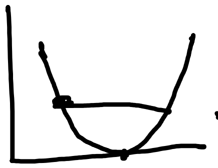
Best Fit !!

$$L = 3x^2 + 5$$

$$\frac{dL}{dx} = 6x = 0$$

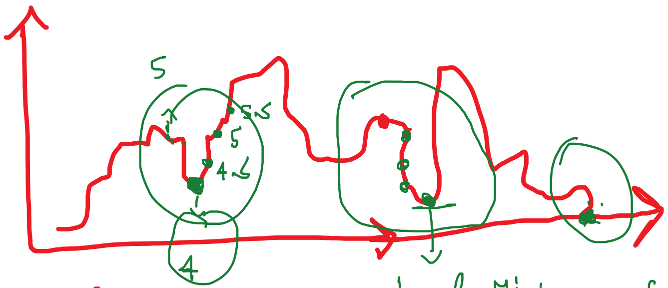
{Minima at $x = 0$} ⇒ not possible for all functions

But in real life situations



Convex

will be on or above the curve starting



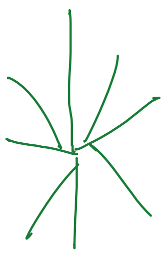
⇒ No global Minima

local Minima

{Non-convex functions}

⇓

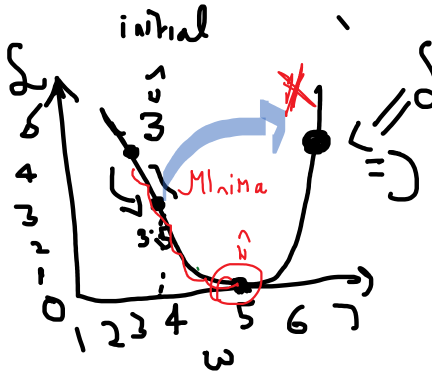
best minima possible



$$\prod_{i=0}^{i=n} \tanh\left(\mathrm{Relu}\left(w_1x_1 + w_2x_2\right)\right) \times \mathrm{sigmoid}\left(w_3x_3\right)$$

{Converting into an optimization problem}

Gradient Descent



$\Rightarrow L = w^2$

$\Rightarrow$ Where is minima

initialize a random $W = 3$

$$\left\{ w^{t+1} = w^t - \boxed{\lambda} \frac{dL}{dw} \right\} \text{ update weight}$$

λ = small number

λ = 0.1

$$w^{t+1} = 3 - 0.1 \times 2w$$

$$w^{t+1} = 3.5$$

$$= 3.6$$

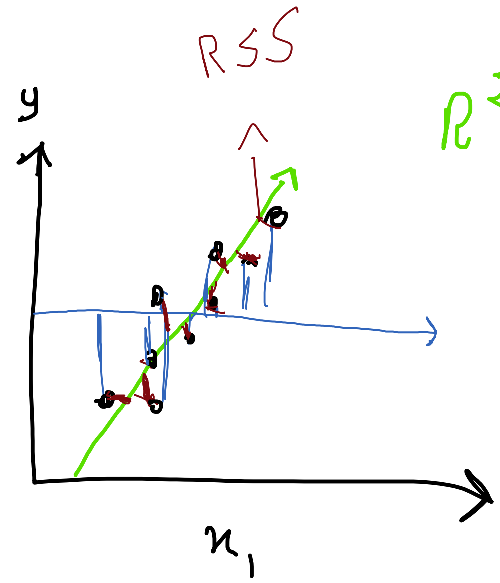
$$= 3.7$$

10^{th} iteration = $w = 5$ => Loss Function will be 0

Gradient descent will help to find a good result! We never know if it is the best possible result

R squared

28 July 2025 22:54



$$R^2 = 1 - \frac{RSS}{TSS}$$ (Total Variance)

TSS => Total Sum of Squares => distance b/w Mean value and the values

Sum of squared

RSS => Residual Sum of Squares — Sum of squared dist b/w Regression line and points

Error in Regression line

$$1 - \left\{\frac{\text{Error}}{\text{Total}}\right\}$$ Error percent

$$1 - \left\{\frac{RSS}{TSS}\right\}$$ Error percent

{ 1 − Error percent = Total info captured }

{ R^2 => How much variance was captured by the Regression line }

$$R^2 = 1 - \left\{\frac{RSS}{TSS}\right\}^{\text{error}}$$

28 July 2025 22:43

$$X \cdot W^T + b^{\checkmark \; w_0}$$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \cdot \begin{bmatrix} w_1 & w_2 & w_3 \end{bmatrix}$$

$$y = \{ x_1 w_1 + x_2 w_2 + x_3 w_3 + w_0 \}$$

28 July 2025 23:06

linear Regression

optimal $W^T = [w_1, w_2, w_3 \dots w_d]$

$$\hat{y} = W^T x + w_0$$

$$x, y \in R^d$$

Minimize $\frac{1}{m} \sum_{i=1}^{i=m} (\hat{y}_i - y_i)^2 \Rightarrow$ Minimizing sum of squared Residuals

$\Downarrow$

$\{$ Mean squared error $\}$

Why we are not Minimizing $|\hat{y}_i - y_i|$

$\Downarrow$

Mean absolute error

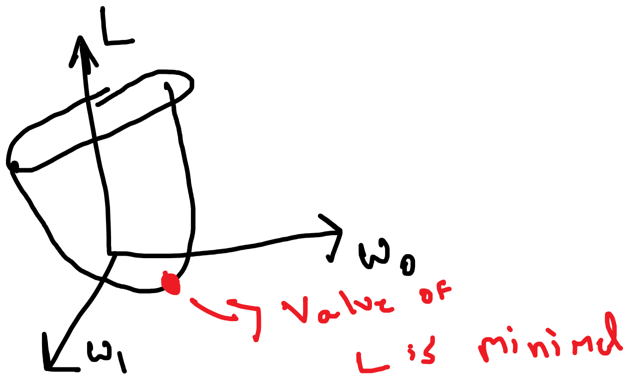
Gradient Calculation

28 July 2025 23:11

$$\text{Min} \ \frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)} - y_{pred}\right]^2$$

$$= \quad \left.y^i_{pred} = w_0 + w_1 x^{(i)}\right\} \Rightarrow \text{prediction function}$$

$$\underset{\text{loss fn}}{\underset{\Downarrow}{L}} = \frac{1}{m}\sum_{i=1}^{m}\left[\underset{\text{Actual}}{\underset{\Downarrow}{y^i}} - \underbrace{\left(w_0 + w_1 x^{(i)}\right)}_{\text{predicted}}\right]^2$$



Gradient descent in action

28 July 2025 23:16

$$\frac{\partial L}{\partial w_0} \text{ and } \frac{\partial L}{\partial w_1}$$ formula to update weights

$$\left\{ w_0^{t+1} = w_0^t - \lambda \frac{\partial L}{\partial w_0} \quad w_1^{t+1} = w_1^t - \lambda \frac{\partial L}{\partial w_1} \right\}$$

$$\frac{\partial L}{\partial w_0} = \frac{\partial \left(y - (w_1 x_1 + w_0)\right)^2}{\partial w_0}$$

$w_1 x_1 + w_0 = y_{pred}$

$(y - y_{pred})^2$

$$= 2(y - \hat{y}) \times \frac{\partial -w_0}{\partial w_0}$$

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

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

$$\frac{\partial L}{\partial w_1} = -2(y - \hat{y}) \cdot x_1$$