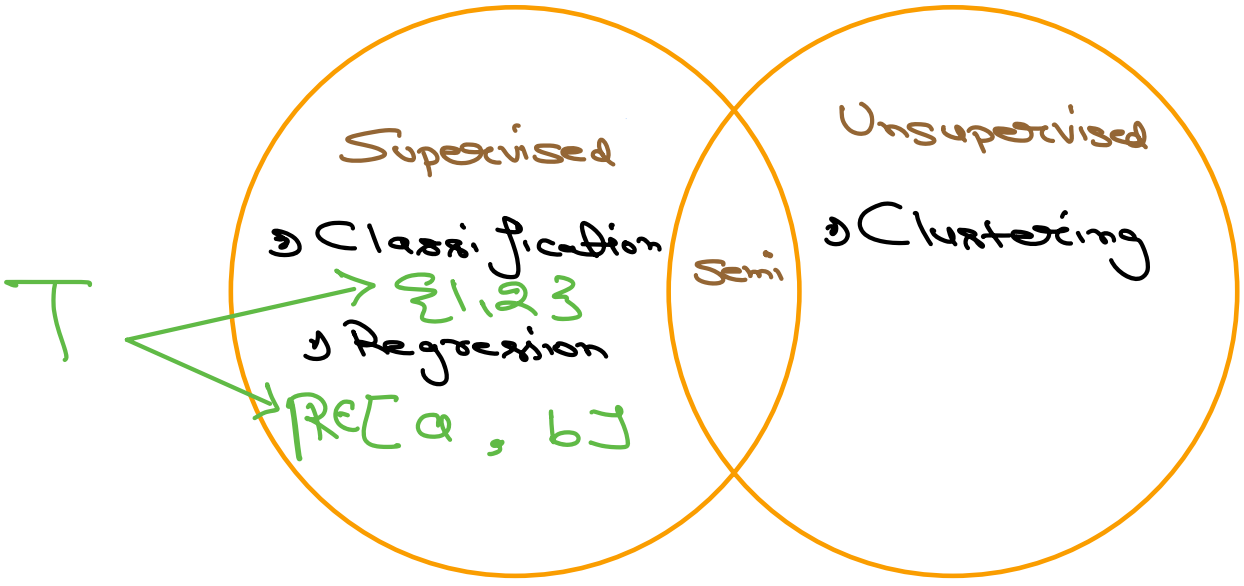
Linear Regression 2

Agenda

- Recap
- Model Interpretability
- Feature importance
- Scratch Implementation
  - Optimization
  - Calculate Gradient

Types of ML

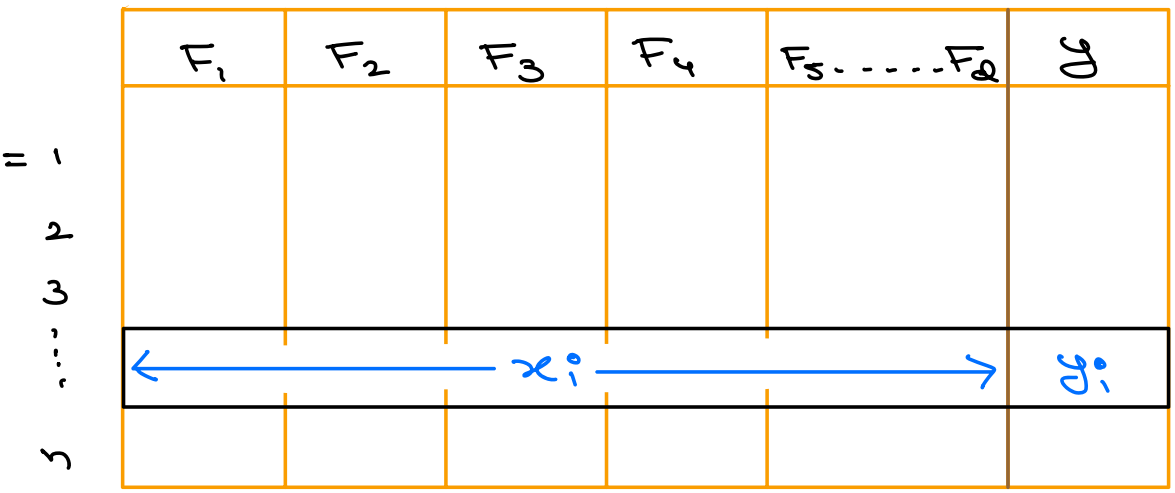


Regression

③ Deals with prediction of continous numeric values.

Ex :- Stock price prediction

Car Value prediction



n → no of rows / samples

d → no of features

ith sample → x_i

ith label → y_i

$y_i \in (0, \infty)$

Regression

* Linear Regression

Linear Hyperplane

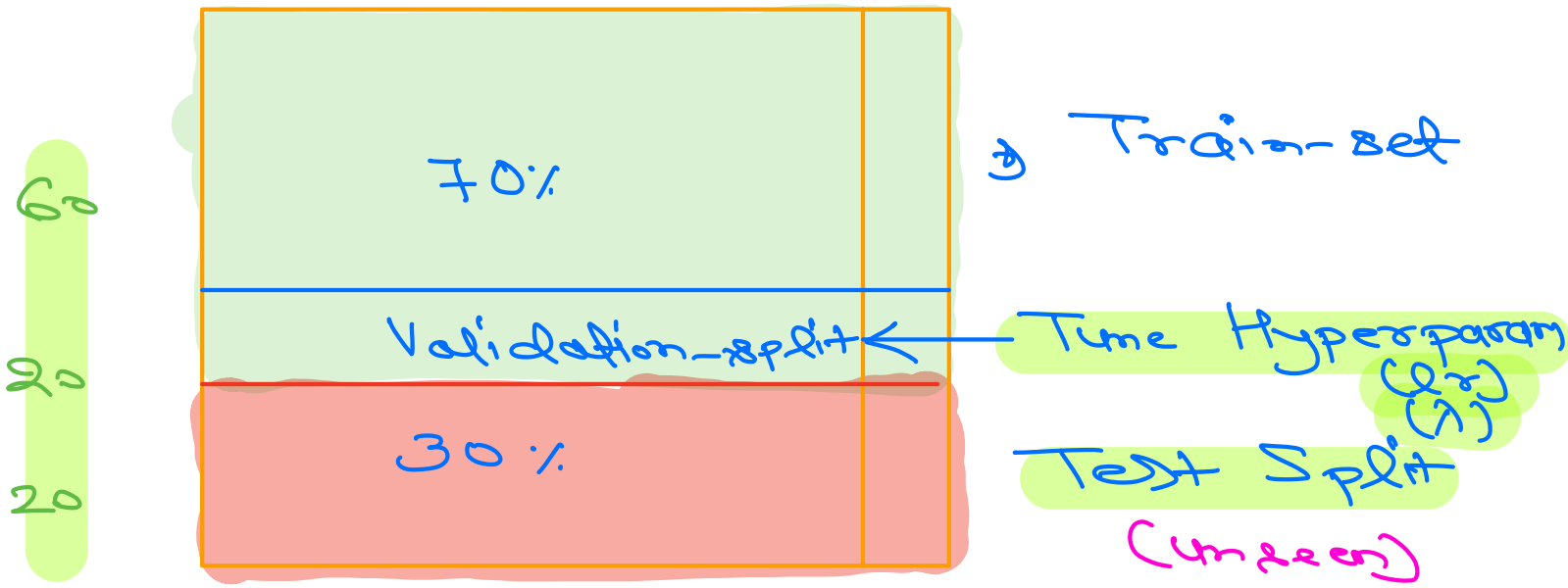
$$w^T \cdot x + w_0 \Rightarrow (\mathbb{R}^1)$$

① Missing value

② Convert Str to Numerical

③ Normalizing

Train Test Split



* Data Leakage

To avoid data leakage Evaluation must be done on Unseen.

Standardization & Mean Value Imputation

μ_g, σ_g

μ_g → full dataset

σ_g → full dataset

μ train-set → Imputing Test-set

$w^T x + w_0$

* current in

Linear Regression

Single Variable L.R.

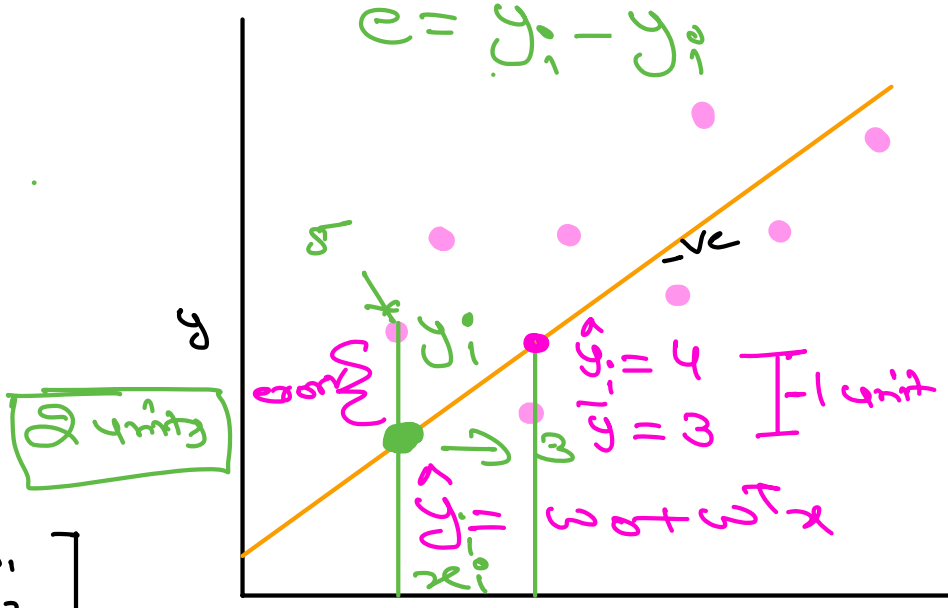
$$\hat{y} = w_0 + w_1(x)$$

Multi Variate L.R

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

Actual - Predi

$$e = y_i - \hat{y}_i$$



- w is a Vector of dims $\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}$ → num-feature x
- x is a Vector of dims $\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_d \end{bmatrix}$ → input Dataset

Goal: Find value of w_0 and w_1 for Best Fit line

minimizing the Errors

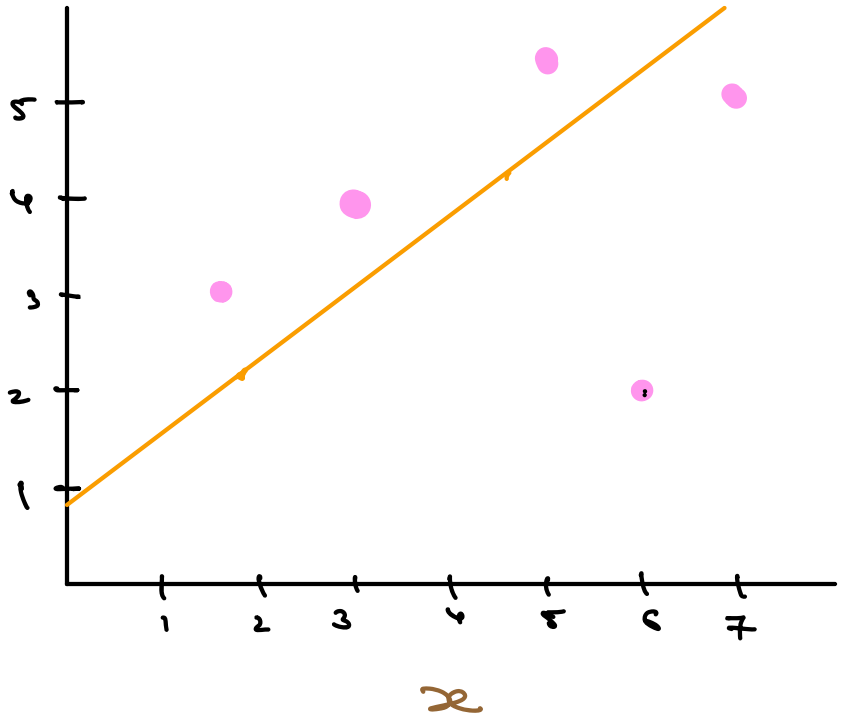
1. Sum of all Errors = $\sum_{i=1}^{n} e_i$ ✗
2. Sum of abs Errors = $\sum_{i=1}^{n} |e_i|$ ✓
3. Sum of Sqr Errors = $\sum_{i=1}^{n} (e_i)^2$ ✓

Error or Residual

i	y_i	$\hat{y}_i$	$y_i - \hat{y}_i$	
1	2	3	-1	
2	4	2	2	(4)
3	5	1	4	(16)

$\hat{y}$: Prediction

$w^T x + b$



(1) Mean Absolute Error

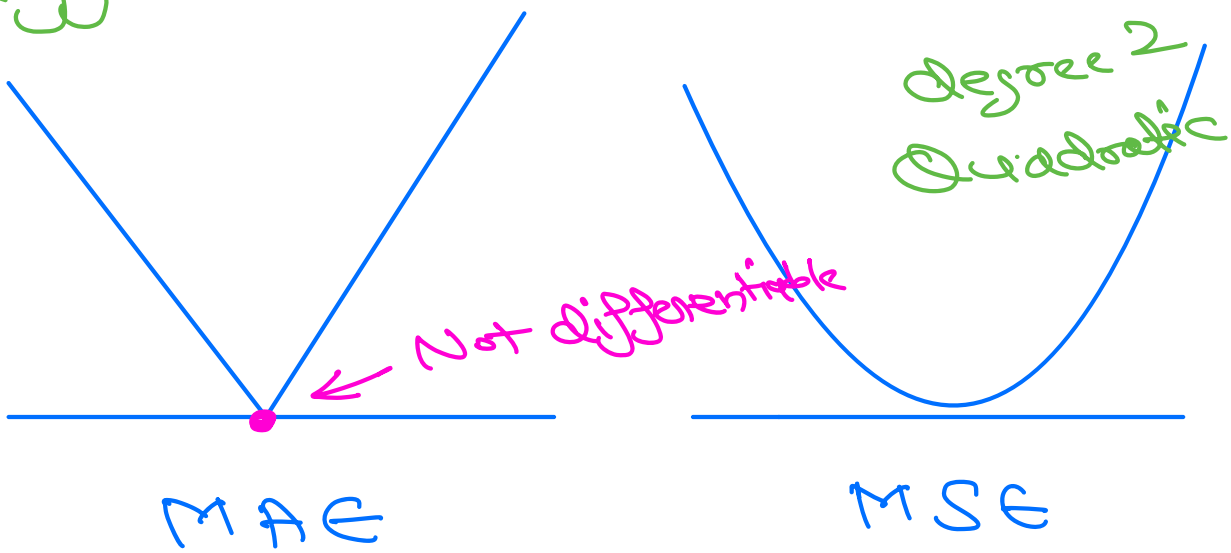
$$\frac{1}{n}\sum_{i=1}^{n} |e_i| \quad \boxed{2.33}$$

(2) Mean Squared Error

$$\frac{1}{n}\sum_{i=1}^{n} (e_i)^2 \quad (7)$$

$21,000 \pm \sqrt{7}$

Loss function must be differentiable



② Loss function ≠ Eval Metric

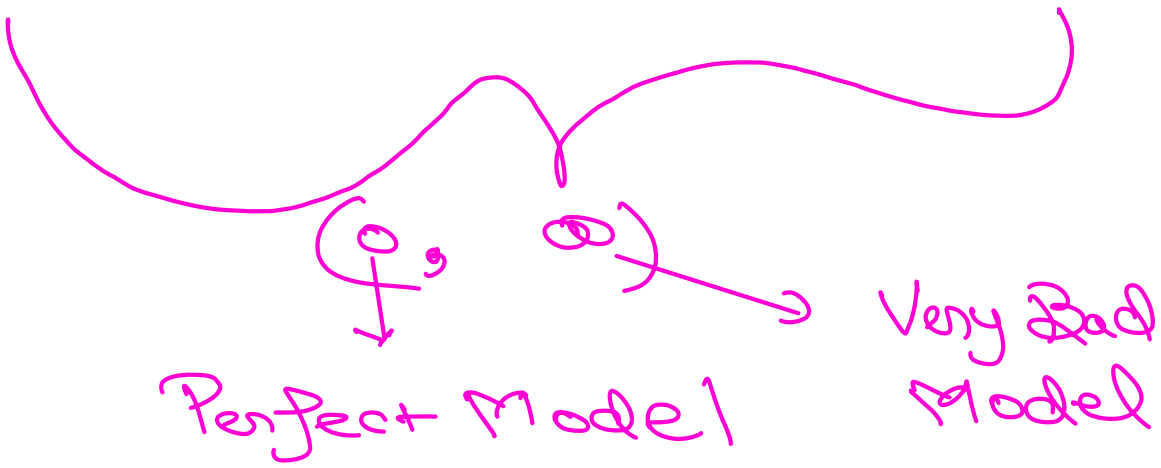
that helps find Best Model

Metric that you report end

③ Regression → MAE and MSE ≠ Eval Metric

RMSE → $\sqrt{MSE}$

MSE or MAE or RMSE



M_1, M_2 } Compare and decide which model

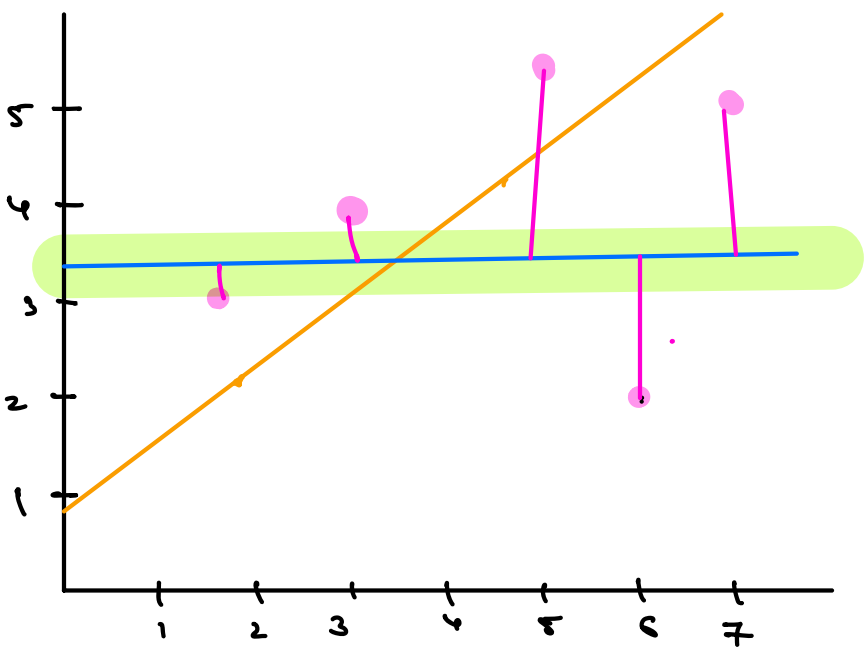
M_1 → R2_Score

(r squared Score)

(Coefficient of Determinism)

$$1 - \frac{SS_{Res}}{SS_{Total}}$$

$$1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$



Case-1:

$$1 - \frac{SS_{res}}{SS_{Total}}$$

$e(\hat{Y} - Y) \Rightarrow SS_{res} \Rightarrow 0$

$$\Rightarrow 1 - \frac{0}{SS_{Total}} \Rightarrow 1$$

Good Model will have score close to 1

Case-2:

$$SS_{res} = SS_{Total}$$

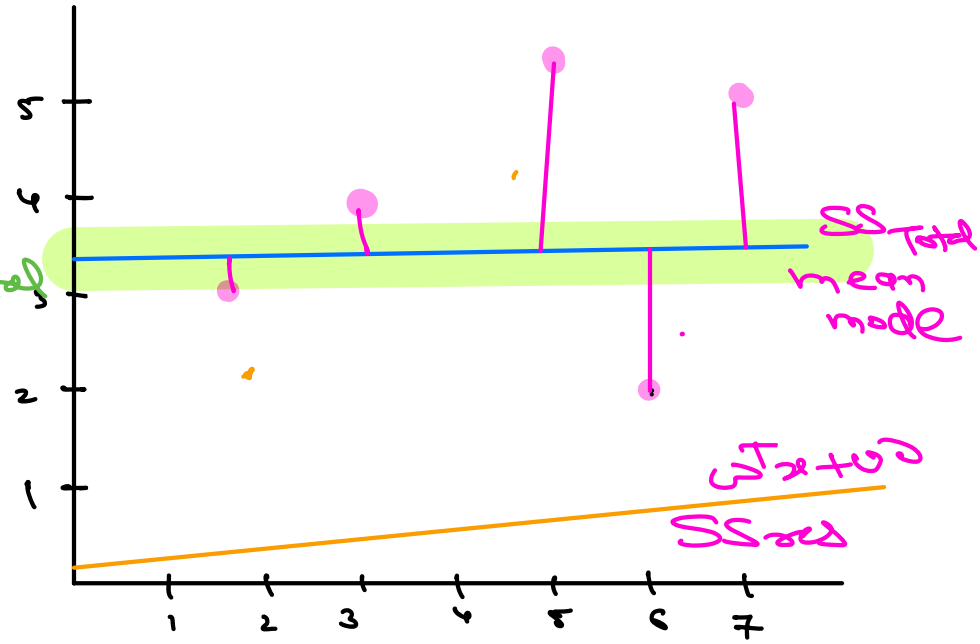
$$\Rightarrow 1 - \frac{SS_{res}}{SS_{Total}} \Rightarrow 1$$

$$\Rightarrow 0$$

Bad Model will have score Close to Zero

Case-3

$SS_{res} > SS_{Total}$



$$1 - \frac{SS_{Res}}{SS_{Total}}$$

$\downarrow$

$(-\infty, 0)$

R^2 Score $(-\infty, 1)$

$\downarrow$

Practical Value $\rightarrow (0, 1)$

Model Interpretability and Weights

$$\hat{y} = w \cdot x + w_0$$

$$\hat{y} = w_1 x_1 + w_2 x_2 + \cdots + w_0$$

Case-study

$$\hat{y} = w_0 + w_1 x_1 + (-1000) * \text{Age} + 1) + (-10) * \text{odometer} +$$

$-10k$ (Age)

-10 (odo)

Odo +1 ⇒ -10 Dollar

Age +1 ⇒ -10000 Dollar

Case ⇒ -ve w_+

$x \uparrow$ $\hat{y} \downarrow$

$x \downarrow$ $\hat{y} \uparrow$

-ve wt ⇒ x ∝ $\frac{1}{\hat{y}}$

Case ⇒ +ve w_+

$x \uparrow$ $\hat{y} \uparrow$

$x \downarrow$ $\hat{y} \downarrow$

-ve wt ⇒ x ∝ $\hat{y}$

Case $w_1 = 0$

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

$w_2 = 0$

Feature has no impact on Target

Magnitude Cases

Age wt ↓ -10000

Odometer ↓ -10

engine_capacity ↓ 1000

$$\begin{cases} \text{Age} + 1 \downarrow \hat{y} - 10000 \\ \text{Odo} + 1 \downarrow \hat{y} - 10 \end{cases}$$

Ages (0, 20 years) +1 @1000

Odo @ (0, 100,000)

Solution: Scaling

Ages (0, 20 years) @ (-1 1) 1.20

Odo @ (0, 100,000) @ (-1 1) 0.5

Now we can compare Magnitude weight to fairly estimate importan

- Magnitude x Normalized feature weight can be used to determine feature importance

Feature importance in linear regression is determined by :

4 options

Active Duration(Most preferred: 30 seconds)

Appears for	45 Secs

A	The magnitude of the regression coefficients.
B	The number of observations in the dataset.
C	The correlation between the independent variables.
D	The average squared difference between the predicted and actual values.

Quiz time!

Time Left: 0s

When assessing model interpretability in Linear Regression, what is the impact of feature scaling?

51 users have participated

A	Feature scaling does not affect model interpretability	10%
B	Feature scaling improves model interpretability	27%
✓ C	Feature scaling can help compare the magnitudes of different coefficients	63%

End Quiz Now

Consider the following Linear Regression model equation: y = 5.2x1 - 3.8x2 + 2.1x3 + 0.01x4 - 1.5 if we were to drop one feature, which one would be the best choice ?

9 users have participated

A	x1	22%
B	x2	11%
C	x3	0%
✓ D	x4	67%

End Quiz Now

Optimization

Gradient Descent

① Pick $\vec{w}$ and w_0 Randomly

② Calculate $\frac{\partial L}{\partial \vec{w}}$ and $\frac{\partial L}{\partial w_0}$

③ Update Step

$$\vec{w}_{t+1} \rightarrow \vec{w}_t - \eta \frac{\partial L}{\partial \vec{w}_t}$$

$$w_{0_{t+1}} \rightarrow w_{0_t} - \eta \frac{\partial L}{\partial w_0}$$

Converging to minima

Reach —

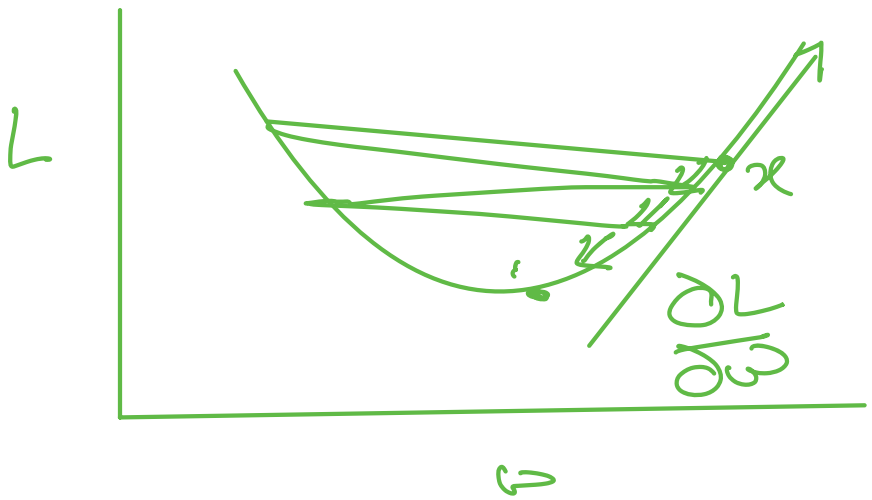
How do we find Global Minima?

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

In gradient descent, what does the gradient represent ?

43 users have participated

A	The direction of steepest increase of the cost function	35% ✓
B	The direction of steepest decrease of the cost function	58%
C	The number of training examples in the dataset	2%
D	The number of layers in the neural network	5%



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

Score ✓

Predict ✓

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

↓

$\frac{\partial L}{\partial w} = ?$

$\frac{\partial L}{\partial b} = ?$

① $\frac{\partial L}{\partial w}$

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

↓

$\frac{1}{n}$ $(y_1 -$ [illegible] $)^2$

$$\frac{\partial L}{\partial w_1} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}\right)^2$$

$$(y = mx + c)^2$$

$$\hat{y} = w_1x_1 + w_2x_2 + \cdots + w_dx_d + w$$

$$f(g(x)) = f'g(x) \times g'(x)$$

$$\frac{\partial L}{\partial w_1} = 2 \times (y_i - \hat{y}) \times \left(\frac{\partial y_i}{\partial w_1} - \frac{\partial \hat{y}}{\partial w_1}\right)$$

$$-$$

$$0$$

$$\frac{\partial (w_1x_1 + w_2x_2 + \cdots + w_d)}{\partial w_1}$$

$$0 \qquad 0$$

$$\frac{\partial \hat{y}}{\partial w_1} = (x_1)$$

$$\frac{\partial}{\partial w_2} = x_2$$

$$\frac{\partial}{\partial w_d} = x_d$$

$$\frac{\partial}{\partial w_0} = 1$$

$y = g(x) \times g'(x)$

$$\frac{\partial L}{\partial w_i} = 2 \times (y_i - \hat{y}) \times (-x_i)$$

$$= -2 (y_i - \hat{y}) \cdot x_i$$

$$\frac{\partial L}{\partial w_d} = \frac{1}{n} \sum_{i=1}^{n} -2 (y_i - \hat{y}) * x_d$$

$$\frac{\partial L}{\partial w_0} = \frac{1}{n} \sum_{i=1}^{n} -2 (y_i - \hat{y})$$