Linear Regression-2

Agenda

1. Model interpretability
2. feature importance
3. Mathematics
   - → gradient Descent

Model Interpretability

Model

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

$$\hat{y} = \dots (-10000).\text{age} + (-10)\,\text{odo.} + (200)\,\text{engine} + \dots$$

Case I : $w_j \rightarrow -ve$

$x_j \uparrow \rightarrow \hat{y} \downarrow$

odo $\uparrow \longrightarrow$ Price $\downarrow$

age + 1 $\longrightarrow \hat{y} : \hat{y} - 10000$

Case II $w_j \rightarrow +ve$

Case III $w_j \rightarrow 0$

No impact on $\hat{y}$ due to f_j

feature importances

"weights"

magnitude of $w_j \uparrow$ = importance$(f_j) \uparrow$

$$\hat{y} = (0.8) \cdot x_2 + \cdots + (0.42) x_5$$

x_2 has higher impact

$$\hat{y} = (-1.9) x_2 + \cdots + (-4.4) \cdot x_5$$

x_5 has higher impact

$$\hat{y} = (2.3) x_2 + \cdots + (-5.1) \cdot x_5$$

x_5 has higher impact

Feature importance in linear regression is determined by :

85 users have participated

✓	A	The magnitude of the regression coefficients.	92%
	B	The number of observations in the dataset.	0%
	C	The correlation between the independent variables.	2%
	D	The average squared difference between the predicted and actual values.	6%

End Quiz Now

Feature Scaling

age

odometer

$$\hat{y} = \ldots -10000\, \underset{[1-15]}{age} + -10 \cdot \underset{[10k-150k]}{odo}$$

age is Better odo ✗

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

86 users have participated

A	Feature scaling does not affect model interpretability	0%
B	Feature scaling improves model interpretability	35%
C	Feature scaling can help compare the magnitudes of different coefficients	65%

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 ?

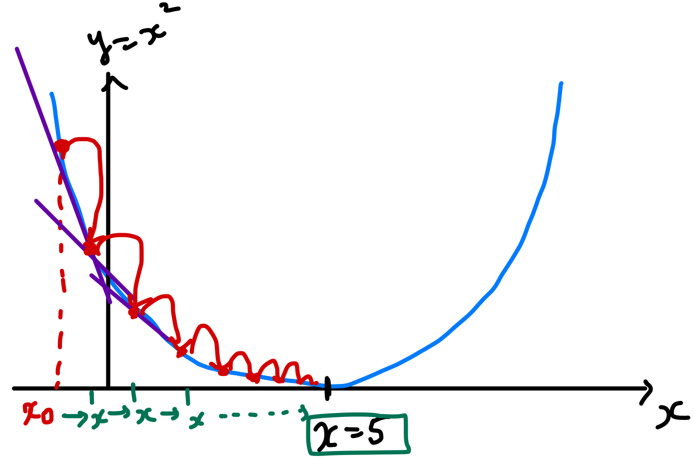
37 users have participated

A	x1	0%
B	x2	5%
C	x3	3%
D	x4	92%

End Quiz Now

$$\hat{y} = 5.2x_1 - 3.8x_2 + 2.1x_3 + 0.01x_4 - 1.5$$

Gradient Descent



$L = y = f(x) = (x-5)^2$

G.D

1. Pick randomly x_0

2. $\left.\frac{\partial L}{\partial x}\right|_{x_0}$ $-ve$ move towards neg deratives

3. $x = x - \eta \cdot \frac{\partial L}{\partial x}$

↳ learning rate : 0.1

In gradient descent, what does the gradient represent ?

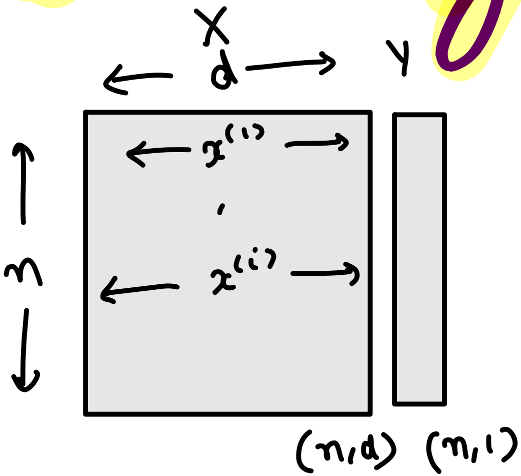
90 users have participated

A	The direction of steepest increase of the cost function	34%
B	The direction of steepest decrease of the cost function	62%
C	The number of training examples in the dataset	1%
D	The number of layers in the neural network	2%

(-gradient)

End Quiz Now

Linear Regression



$$w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}_{(d,1)}$$

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

$$\boxed{\hat{Y} = \underbrace{X}_{(n,d)} \cdot \underbrace{W}_{(d,1)} + w_0}$$

$$(n,d)\ (d,1) \rightarrow (n,1)$$

$$\hat{Y} = \begin{bmatrix} \hat{y}^{(1)} \\ \hat{y}^{(2)} \\ \vdots \\ \hat{y}^{(n)} \end{bmatrix}_{(n,1)}$$

Loss

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

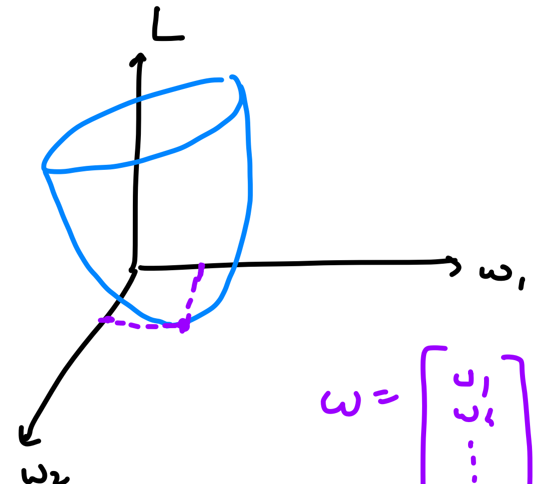
Gradient Descent

$$\text{Loss}(w) = \frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} - \hat{y}^{(i)}\right)^2$$

minimise MSE

optimization problem

$$\underset{w}{\text{argmin}} \ \frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} - \hat{y}^{(i)}\right)^2$$



$$w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}$$

1. randomly init 'w'
2. $\frac{\partial L}{\partial w} = \nabla_w L$
3. repeat n-iter times $\left\{ \ w_j' = w_j - \eta \cdot \frac{\partial L}{\partial w_j} \ \right\}$

Gradients

$$w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}$$

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

$$\to = w_1 x_1^{(i)} + w_2 x_2^{(i)} + \dots + w_d x_d^{(i)}$$

$$\frac{\partial L}{\partial w} = \begin{bmatrix} \frac{\partial L}{\partial w_1} \\ \frac{\partial L}{\partial w_2} \\ \vdots \\ \frac{\partial L}{\partial w_j} \\ \vdots \\ \frac{\partial L}{\partial w_d} \end{bmatrix}$$

away

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

$$\frac{\partial L}{\partial w_j}$$

for 1 dataPoint

$$\frac{\partial L}{\partial w_j} = \frac{\partial}{\partial w_j}(y - \hat{y})^2 \qquad \frac{d}{dx} f(g(x)) = \frac{d f(x)}{d g(x)} \cdot \frac{d g(x)}{dx}$$

$$= \frac{\partial (y-\hat{y})^2}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial w_j}$$

$$= -2(y-\hat{y}) \cdot \frac{\partial}{\partial w_j} \quad w_1x_1 + w_2x_2 + \dots w_jx_j + \dots w_dx_d$$

$$= -2(y-\hat{y}) \cdot x_j \Rightarrow 2(\hat{y}-y) \cdot x_j$$

derivative for all Points

$$\frac{\partial L}{\partial w_j} = \frac{2}{n} \sum_{i=1}^{n} (\hat{y}^{(i)} - y^{(i)}) \cdot x_j^{(i)}$$

G.D

1. Randomly init 'w'

2. $\dfrac{\partial L}{\partial w_j}$

3. Repeat {

$$w_j = w_j - \eta \cdot \left[\frac{1}{n} \sum_{i=1}^{n} \left(\hat{y}^{(i)} - y^{(i)} \right) \cdot x_j^{(i)} \right]$$

}

$$\frac{\partial L}{\partial w} \rightarrow \left[\quad \right]_{(d,1)}$$

$$\frac{\partial L}{\partial w} = \frac{2}{n} \sum_{i=1}^{n} \underbrace{\left(\hat{y}^{(i)} - y^{(i)} \right)}_{A} \cdot \underbrace{x_j^{(i)}}_{B}$$

$$\begin{bmatrix} \\ \\ \\ \end{bmatrix}_{(d,1)} \qquad \underbrace{\left(\hat{Y} - Y \right)}_{(n,1)} \qquad X \rightarrow (n,d)$$

$$X^T \cdot \left(\hat{Y} - Y \right)$$

$$(d,n) \cdot (n,1)$$

$$(d,1)$$

What is the objective of Gradient Descent in linear regression?

64 users have participated

A	Minimize the absolute error	14%
B ✓	Minimize the squared error	73%
C	Maximize the R-squared score	5%
D	Maximize the accuracy	8%

End Quiz Now

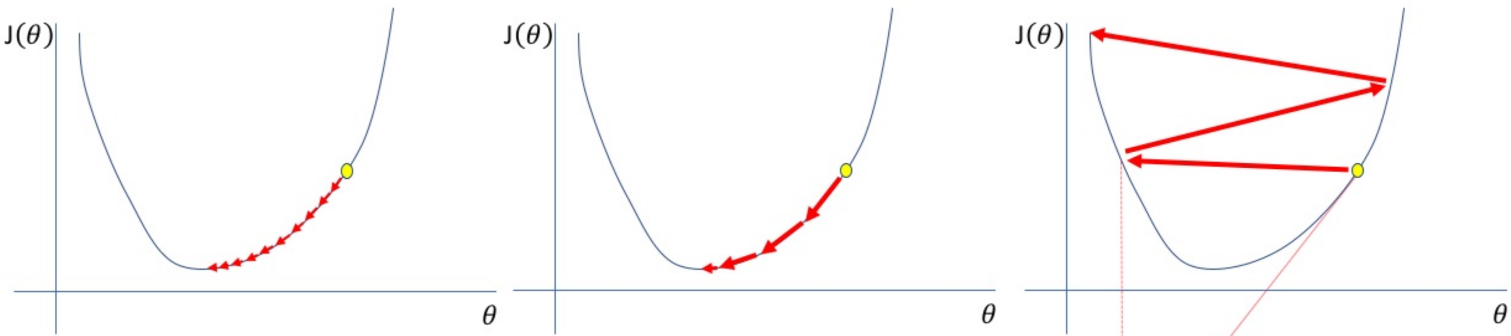
What happens if the learning rate in gradient descent for linear regression is set too large?

21 users have participated

A	The algorithm will converge faster to the optimal solution.	10%
B	The model will overfit the training data, leading to poor generalization.	10%
C ✓	The algorithm may fail to converge, and the coefficients may oscillate or diverge.	81%
D	The cost function will be overestimated, resulting in an inflated R2 score.	0%

End Quiz Now

Too low **Just right** **Too high**



A small learning rate requires many updates before reaching the minimum point

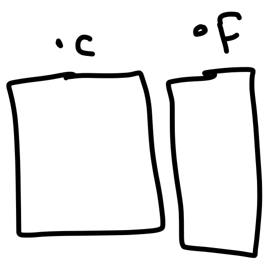
The optimal learning rate swiftly reaches the minimum point

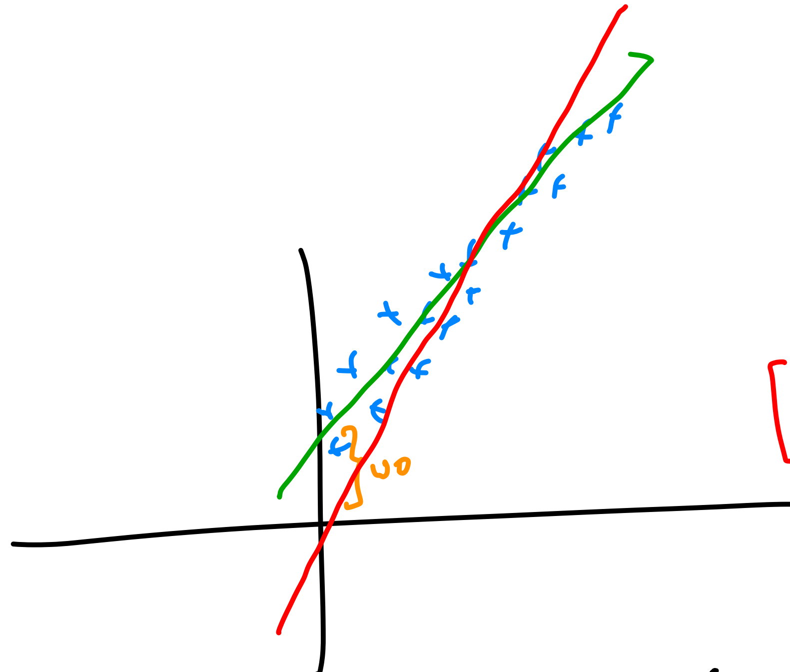
Too large of a learning rate causes drastic updates which lead to divergent behaviors

Vanilla / Batch G.D

Mini-Batch G.D

Stochastic G.D





[w0 is imp]