Variants of Gradient Descent

What we will be covering in this module?

- Variants of Gradient Descent

What we will be covering in this module?

- Variants of Gradient Descent
- Challenges / Problems with Gradient Descent

What we will be covering in this module?

- Variants of Gradient Descent
- Challenges / Problems with Gradient Descent
- Different types of optimizers

What we will be covering in this module?

- Variants of Gradient Descent
- Challenges / Problems with Gradient Descent
- Different types of optimizers
- Implementing optimizers from scratch

Gradient Descent Recap

$$\theta_i = \theta_i - \alpha * dJ / d\theta_i$$

Gradient Descent Recap

Parameter

$$\theta_i = \theta_i - \alpha * dJ / d\theta_i$$

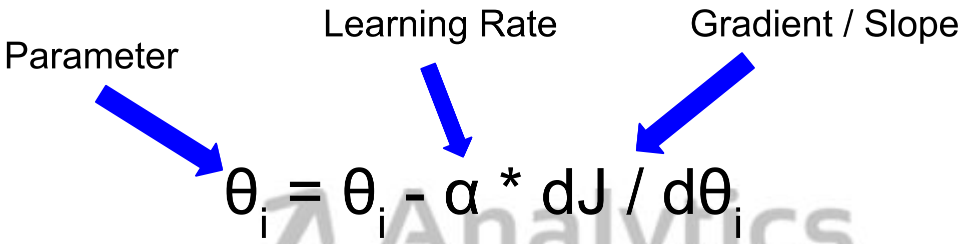
Gradient Descent Recap

Parameter

Gradient / Slope

$$\theta_i = \theta_i - \alpha * dJ / d\theta_i$$

Gradient Descent Recap



Parameter

Learning Rate

Gradient / Slope

$$\theta_i = \theta_i - \alpha * dJ / d\theta_i$$

Variants of Gradient Descent

Parameter

Learning Rate

Gradient / Slope

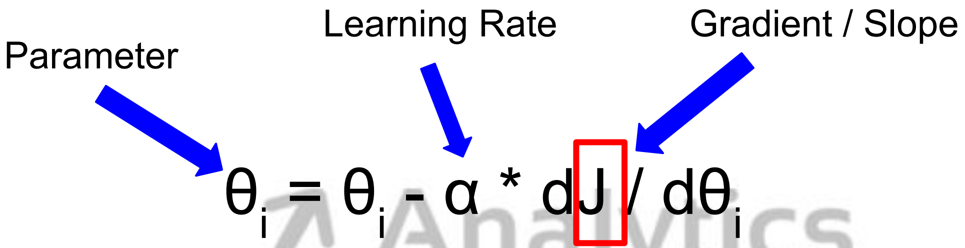
$$\theta_i = \theta_i - \alpha * d\boxed{J} / d\theta_i$$

Variants of Gradient Descent

$$\overset{\text{Parameter}}{\theta_i} = \theta_i - \overset{\text{Learning Rate}}{\alpha} * \overset{\text{Gradient / Slope}}{d\boxed{J} / d\theta_i}$$

Entire Training Set (m)

Variants of Gradient Descent



Parameter

Learning Rate

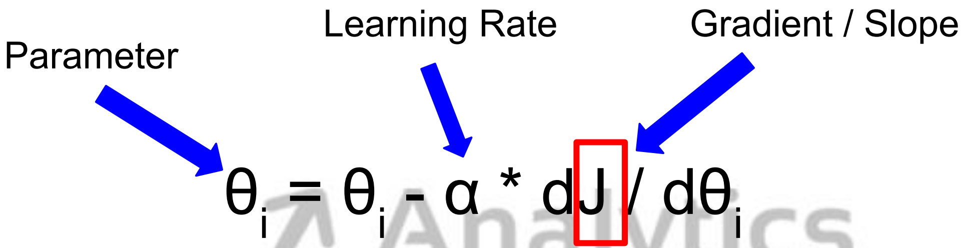
Gradient / Slope

$$\theta_i = \theta_i - \alpha * d\boxed{J} / d\theta_i$$

Entire Training Set (m)

Batch Gradient Descent

Variants of Gradient Descent



Parameter

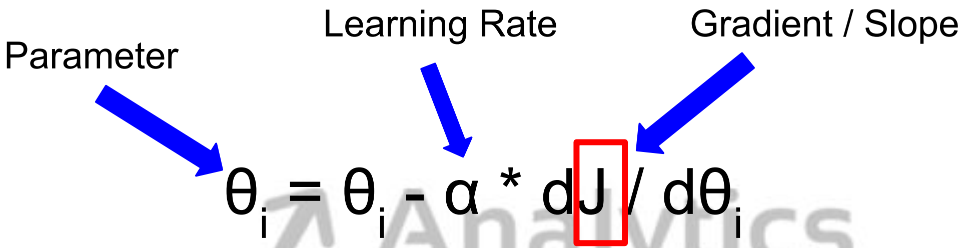
Learning Rate

Gradient / Slope

$$\theta_i = \theta_i - \alpha * d\boxed{J} / d\theta_i$$

Entire Training Set (m)	Batch Gradient Descent
Single Observation (1)	Stochastic Gradient Descent (SGD)

Variants of Gradient Descent



Parameter

Learning Rate

Gradient / Slope

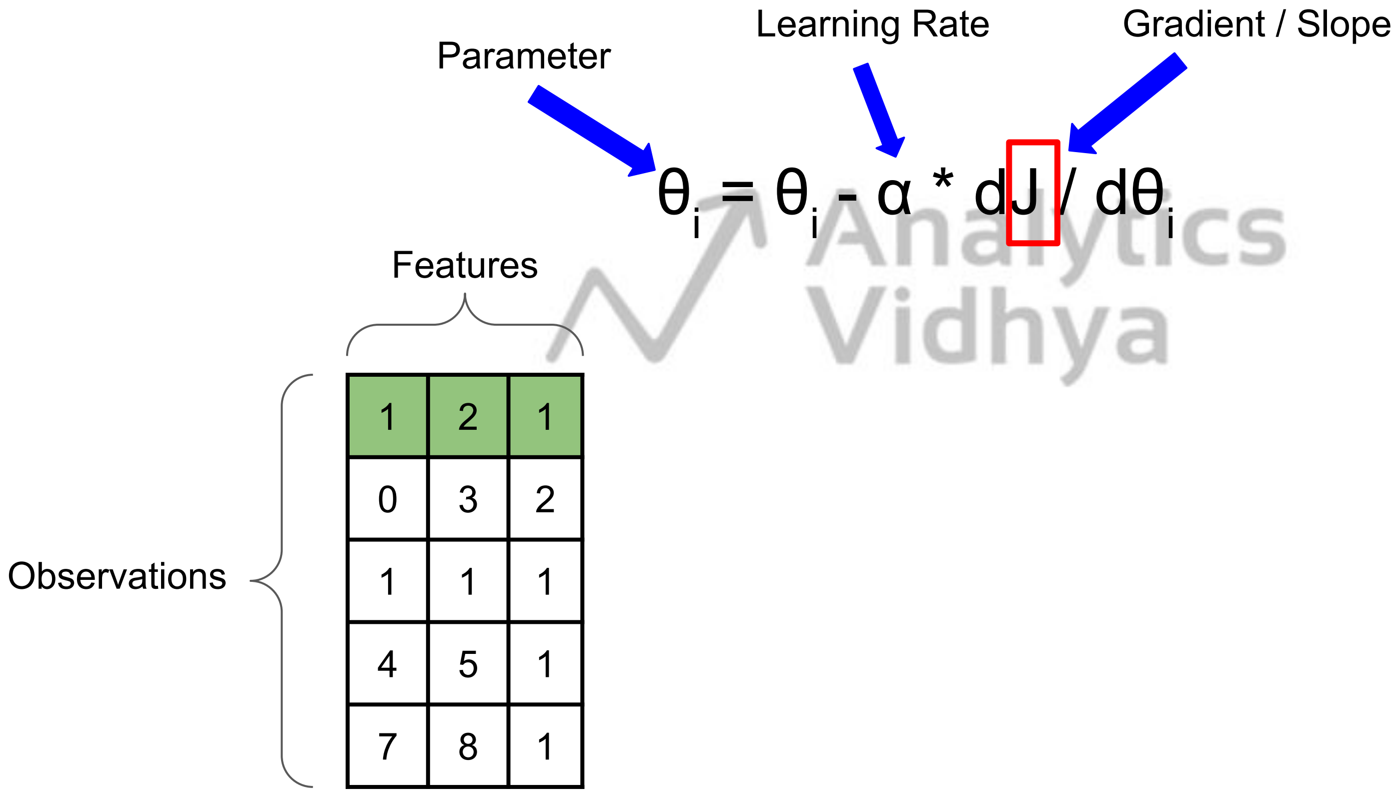
$$\theta_i = \theta_i - \alpha * d\boxed{J} / d\theta_i$$

Features

Observations

1	2	1
0	3	2
1	1	1
4	5	1
7	8	1

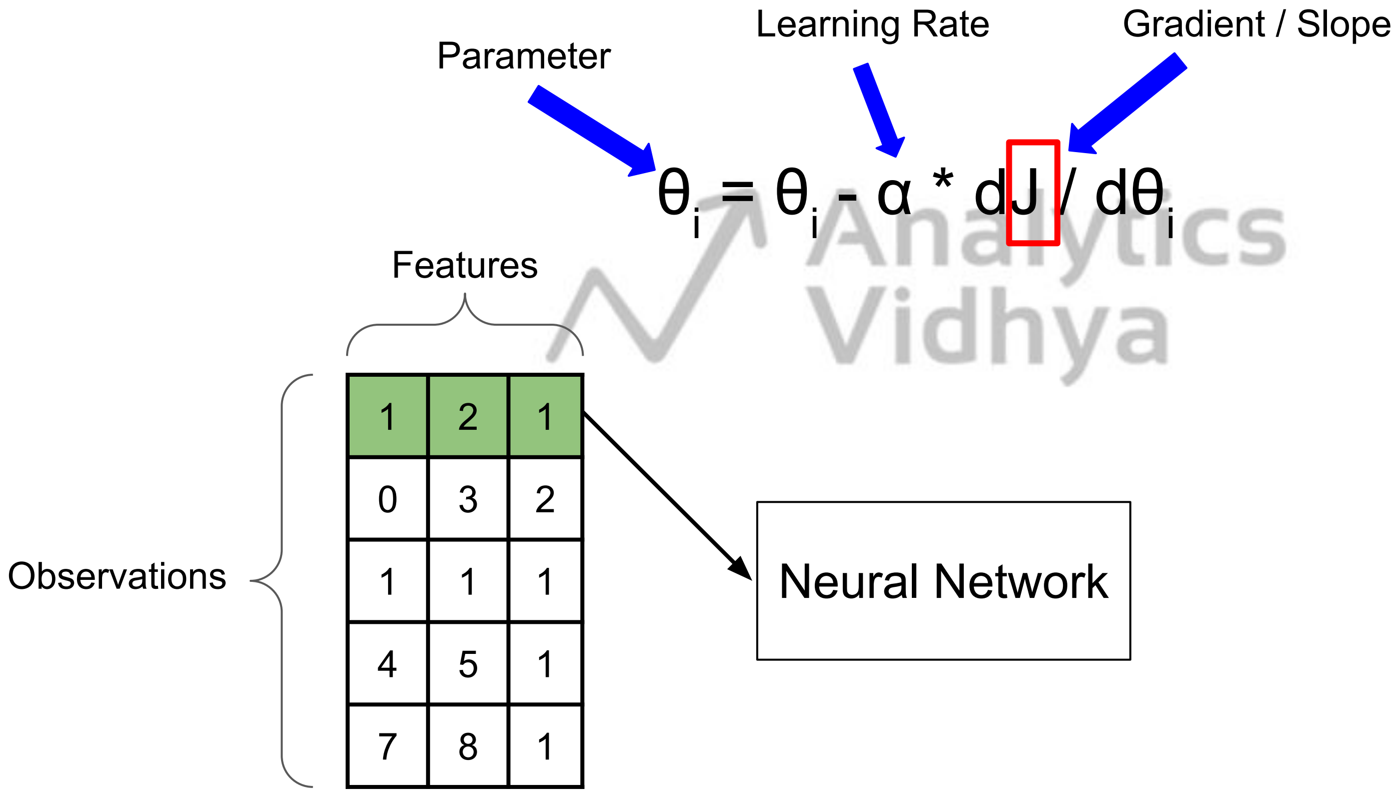
Stochastic Gradient Descent (SGD)



$$\theta_i = \theta_i - \alpha * d\boxed{J} / d\theta_i$$

1	2	1
0	3	2
1	1	1
4	5	1
7	8	1

Stochastic Gradient Descent (SGD)



Parameter

Learning Rate

Gradient / Slope

$$\theta_i = \theta_i - \alpha * d\boxed{J} / d\theta_i$$

Features

Observations

1	2	1
0	3	2
1	1	1
4	5	1
7	8	1

Neural Network

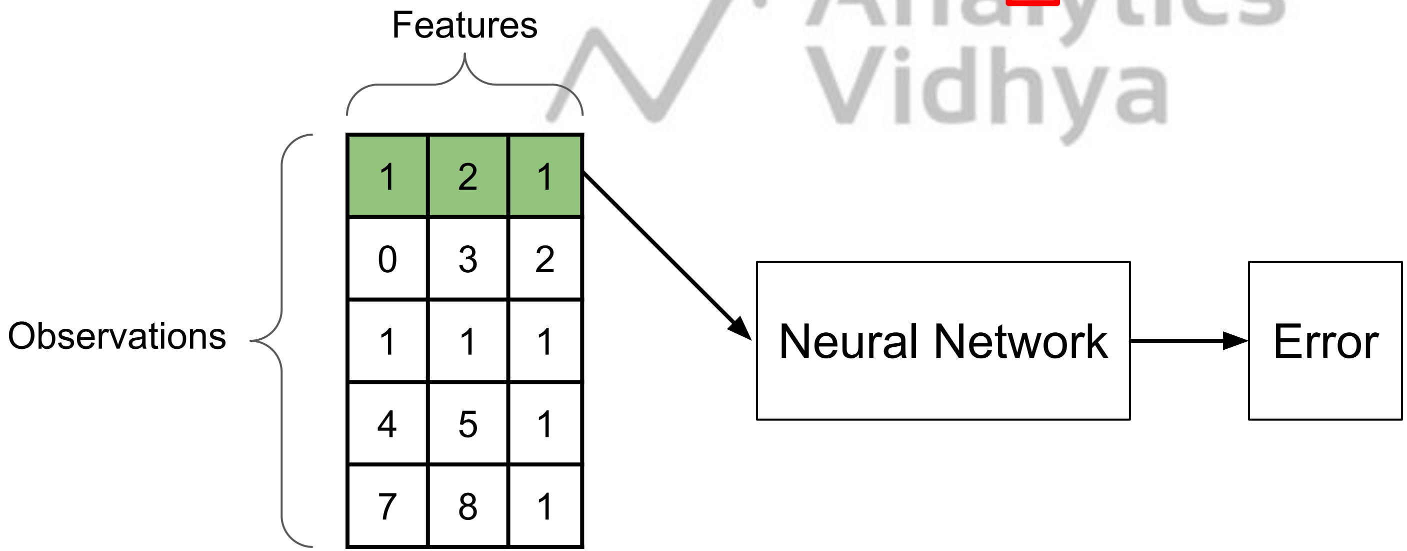
Stochastic Gradient Descent (SGD)

Parameter

Learning Rate

Gradient / Slope

$$\theta_i = \theta_i - \alpha * d\boxed{J} / d\theta_i$$

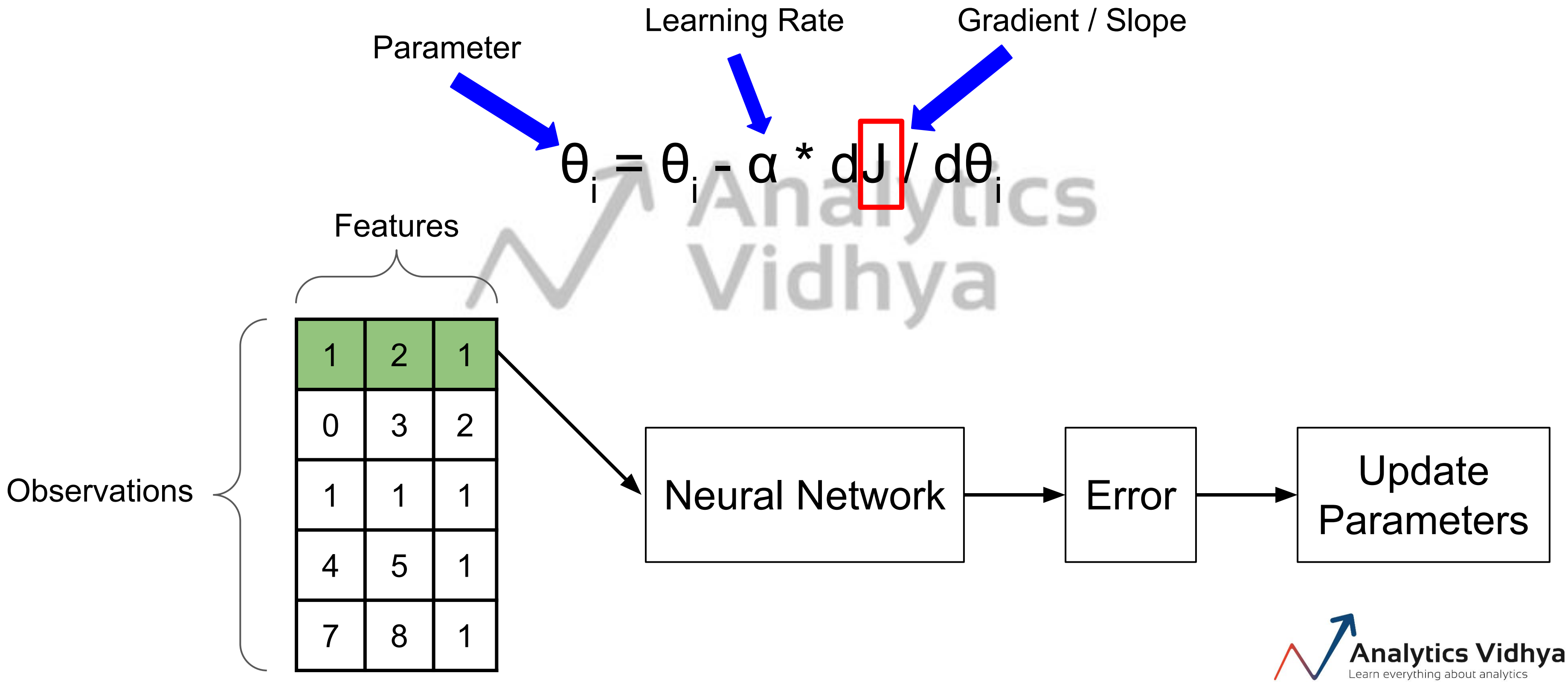


Features

Observations

1	2	1
0	3	2
1	1	1
4	5	1
7	8	1

Stochastic Gradient Descent (SGD)



$$\theta_i = \theta_i - \alpha * d\boxed{J} / d\theta_i$$

1	2	1
0	3	2
1	1	1
4	5	1
7	8	1

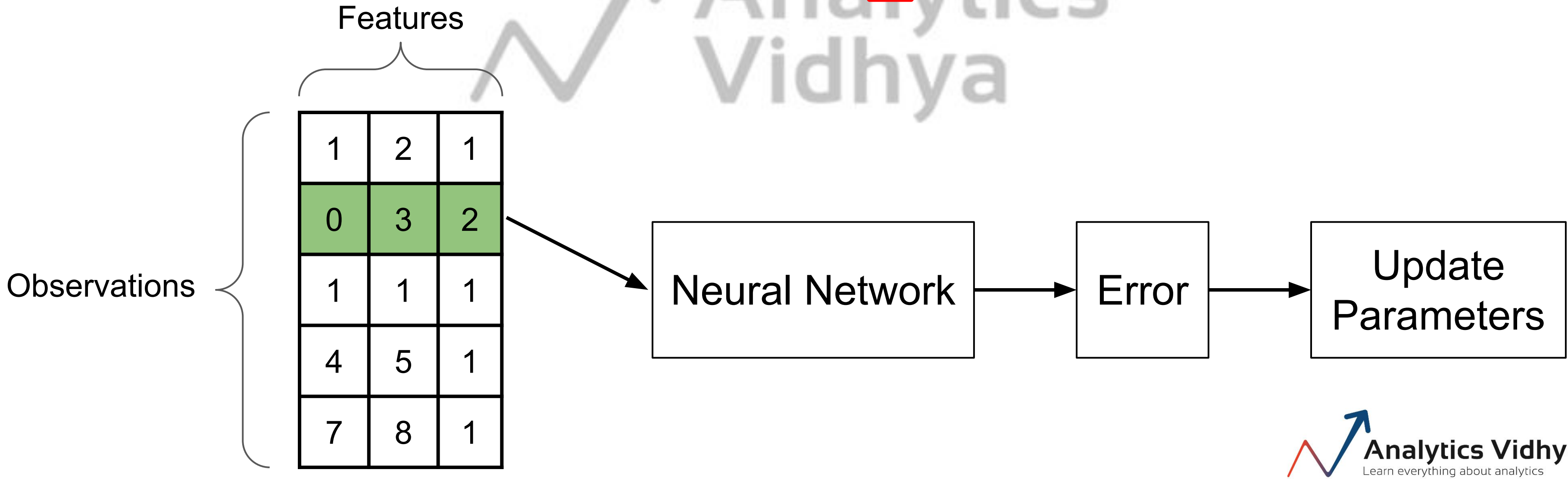
Stochastic Gradient Descent (SGD)

Parameter

Learning Rate

Gradient / Slope

$$\theta_i = \theta_i - \alpha * d\boxed{J} / d\theta_i$$



Features

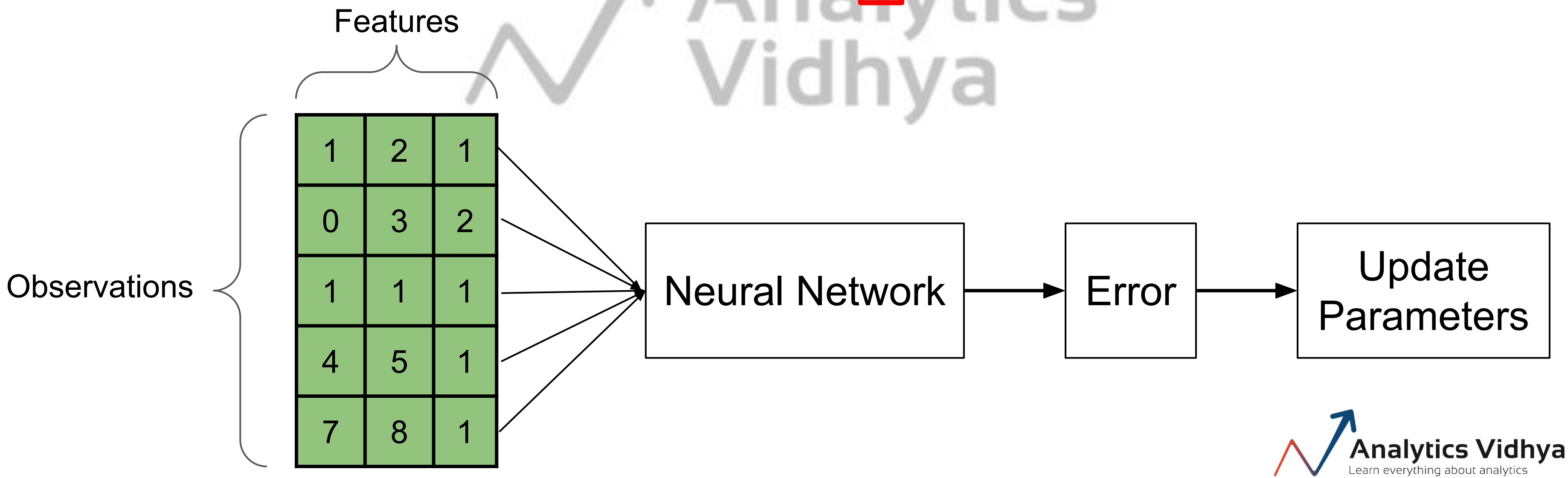
Observations

1	2	1
0	3	2
1	1	1
4	5	1
7	8	1

Neural Network → Error → Update Parameters

Stochastic Gradient Descent (SGD)

$$\overset{\text{Parameter}}{\theta_i} = \theta_i - \overset{\text{Learning Rate}}{\alpha} * d\overset{\text{Gradient / Slope}}{\boxed{J}} / d\theta_i$$



1	2	1
0	3	2
1	1	1
4	5	1
7	8	1

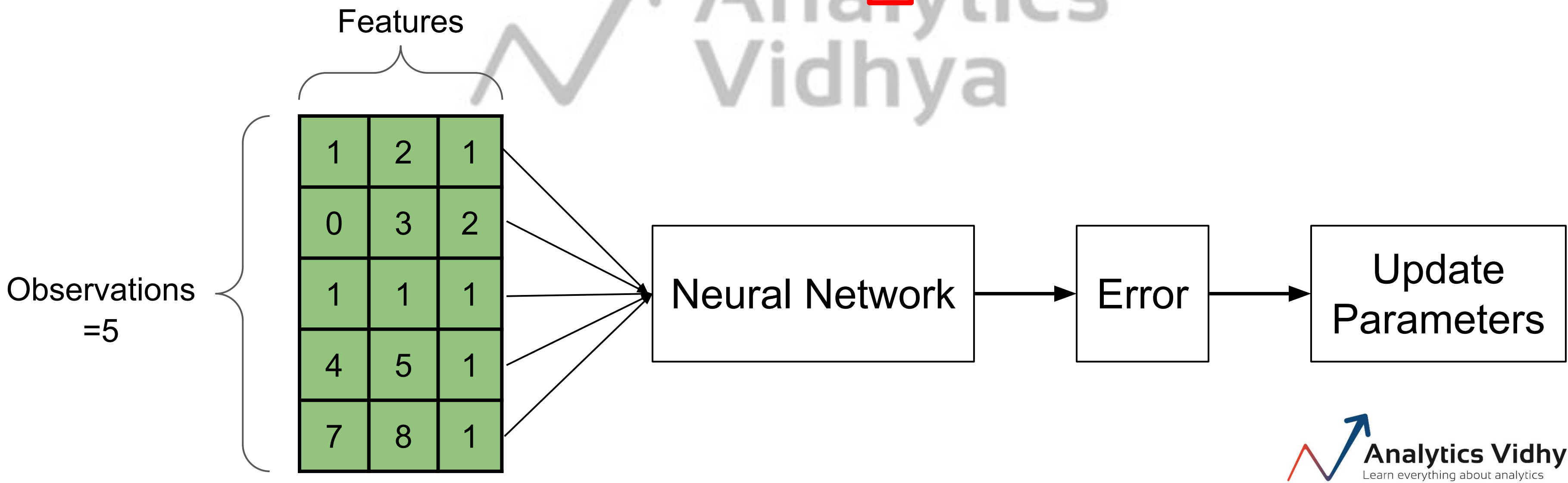
Stochastic Gradient Descent (SGD)

Parameter

Learning Rate

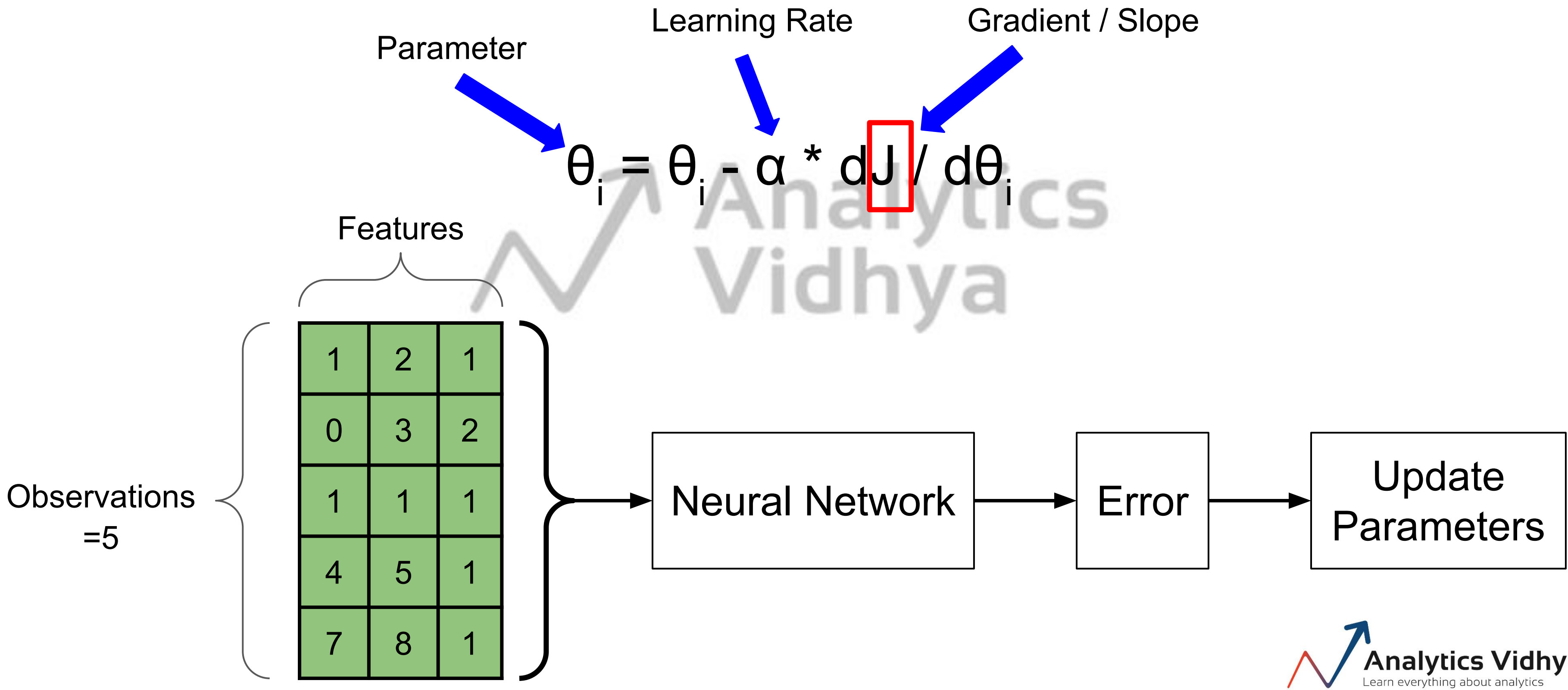
Gradient / Slope

$$\theta_i = \theta_i - \alpha * d\boxed{J} / d\theta_i$$



1	2	1
0	3	2
1	1	1
4	5	1
7	8	1

Batch Gradient Descent



$$\theta_i = \theta_i - \alpha * d\boxed{J} / d\theta_i$$

1	2	1
0	3	2
1	1	1
4	5	1
7	8	1

Variants of Gradient Descent

Parameter

Learning Rate

Gradient / Slope

$$\theta_i = \theta_i - \alpha * d\boxed{J} / d\theta_i$$

Entire Training Set (m)	Batch Gradient Descent
Single Observation (1)	Stochastic Gradient Descent (SGD)

Variants of Gradient Descent

Parameter

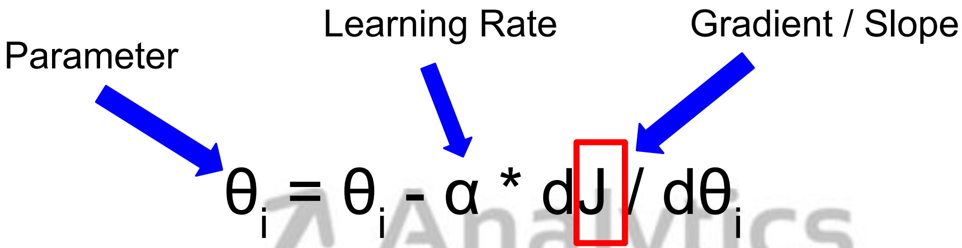
Learning Rate

Gradient / Slope

$$\theta_i = \theta_i - \alpha * d\boxed{J} / d\theta_i$$

Entire Training Set (m)	Batch Gradient Descent
Single Observation (1)	Stochastic Gradient Descent (SGD)
1 < x < m	Mini-Batch Gradient Descent

Mini-Batch Gradient Descent



Parameter

Learning Rate

Gradient / Slope

$$\theta_i = \theta_i - \alpha * d\boxed{J} / d\theta_i$$

Features

Observations

1	2	1
0	3	2
1	1	1
4	5	1
7	8	1

Mini-Batch Gradient Descent

Batch Size = 2

Parameter

Learning Rate

Gradient / Slope

$$\theta_i = \theta_i - \alpha * d\boxed{J} / d\theta_i$$

Features

Observations

1	2	1
0	3	2
1	1	1
4	5	1
7	8	1

Mini-Batch Gradient Descent

Batch Size = 2

Parameter

Learning Rate

Gradient / Slope

$$\theta_i = \theta_i - \alpha * d\boxed{J} / d\theta_i$$

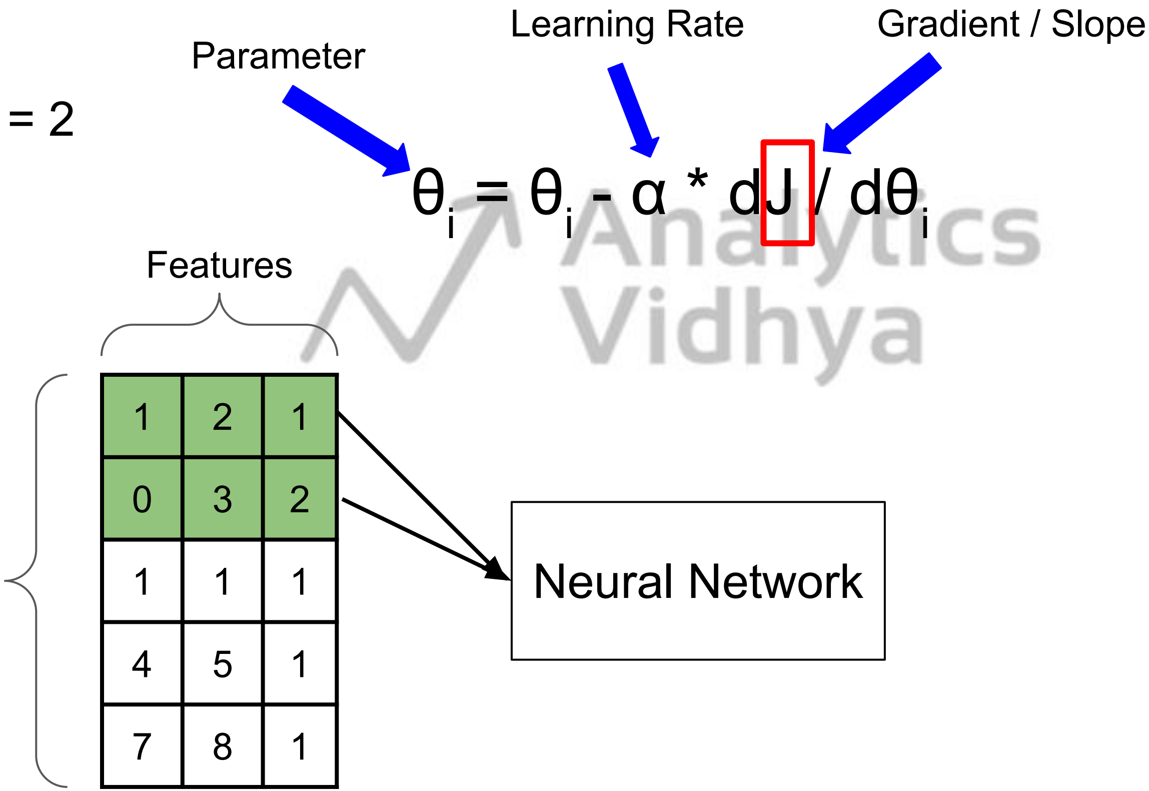
Features

Observations

1	2	1
0	3	2
1	1	1
4	5	1
7	8	1

Mini-Batch Gradient Descent

Batch Size = 2



Parameter

Learning Rate

Gradient / Slope

$$\theta_i = \theta_i - \alpha * d\boxed{J} / d\theta_i$$

Features

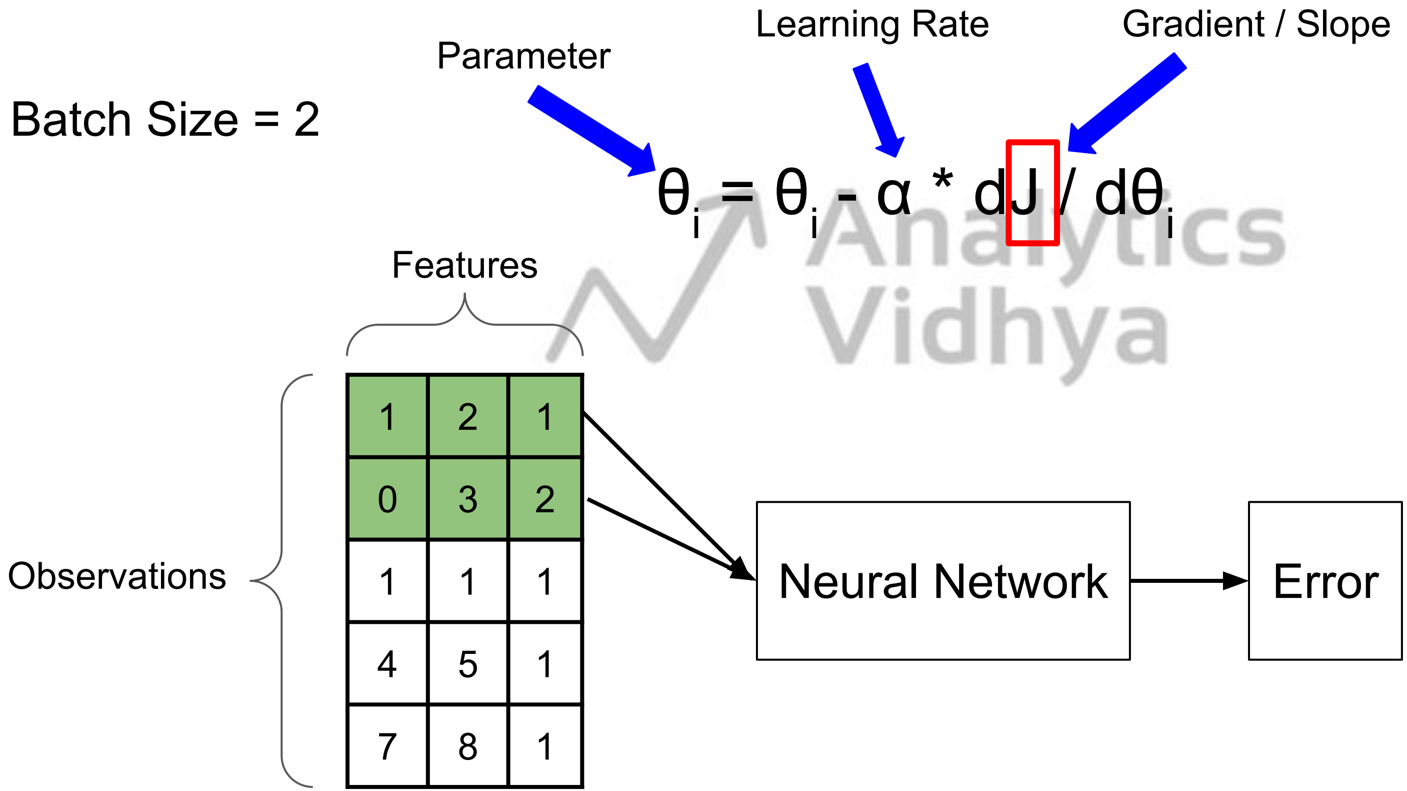
Observations

1	2	1
0	3	2
1	1	1
4	5	1
7	8	1

Neural Network

Mini-Batch Gradient Descent

Batch Size = 2



$$\theta_i = \theta_i - \alpha * d\boxed{J} / d\theta_i$$

1	2	1
0	3	2
1	1	1
4	5	1
7	8	1

Mini-Batch Gradient Descent

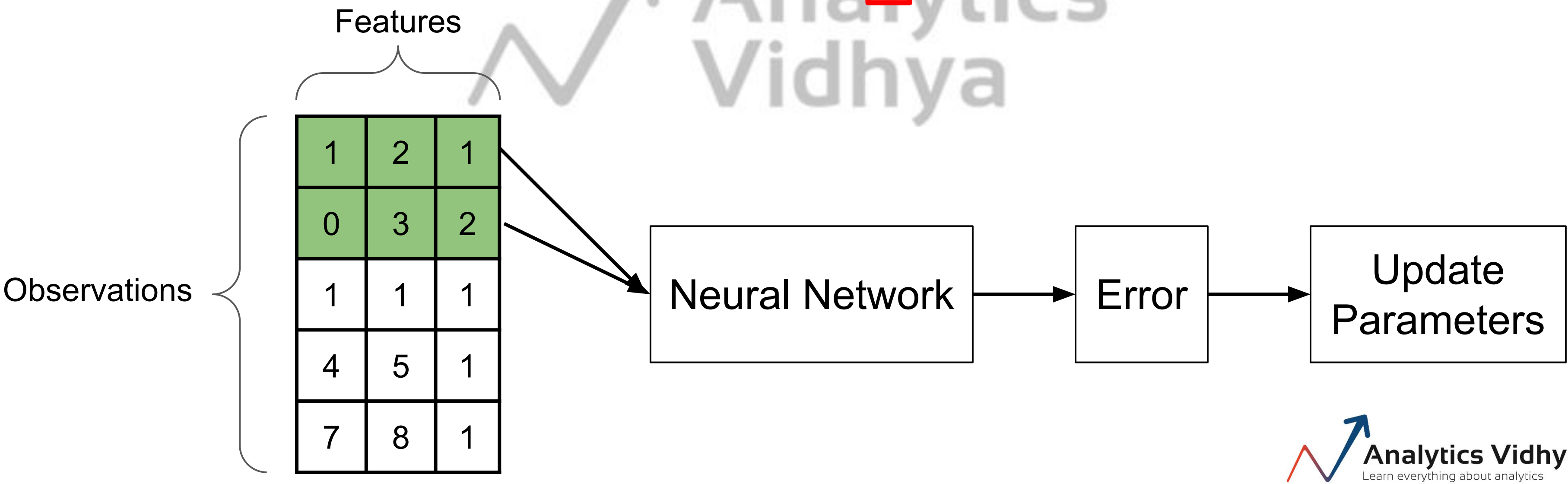
Batch Size = 2

Parameter

Learning Rate

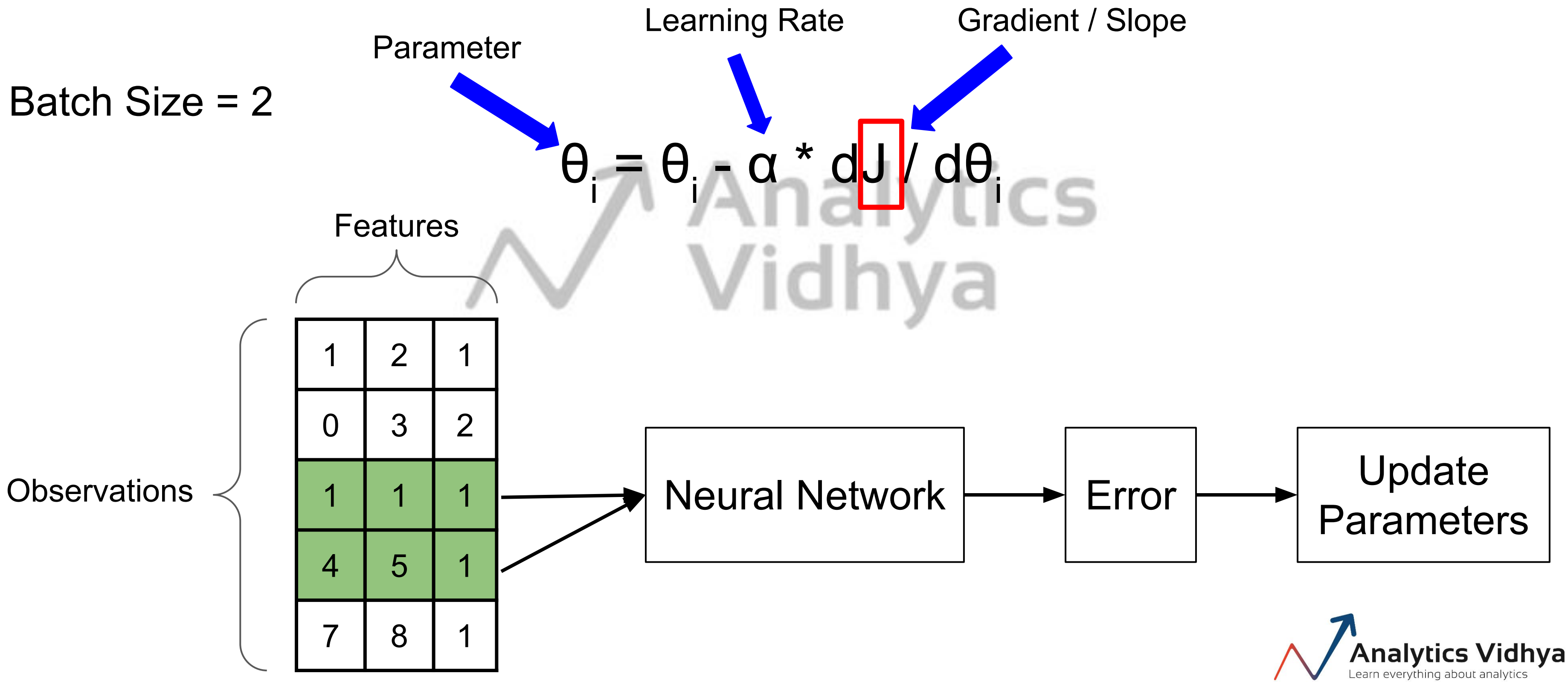
Gradient / Slope

$$\theta_i = \theta_i - \alpha * d\boxed{J} / d\theta_i$$



1	2	1
0	3	2
1	1	1
4	5	1
7	8	1

Mini-Batch Gradient Descent



Batch Size = 2

$$\theta_i = \theta_i - \alpha * d\boxed{J} / d\theta_i$$

1	2	1
0	3	2
1	1	1
4	5	1
7	8	1

Mini-Batch Gradient Descent

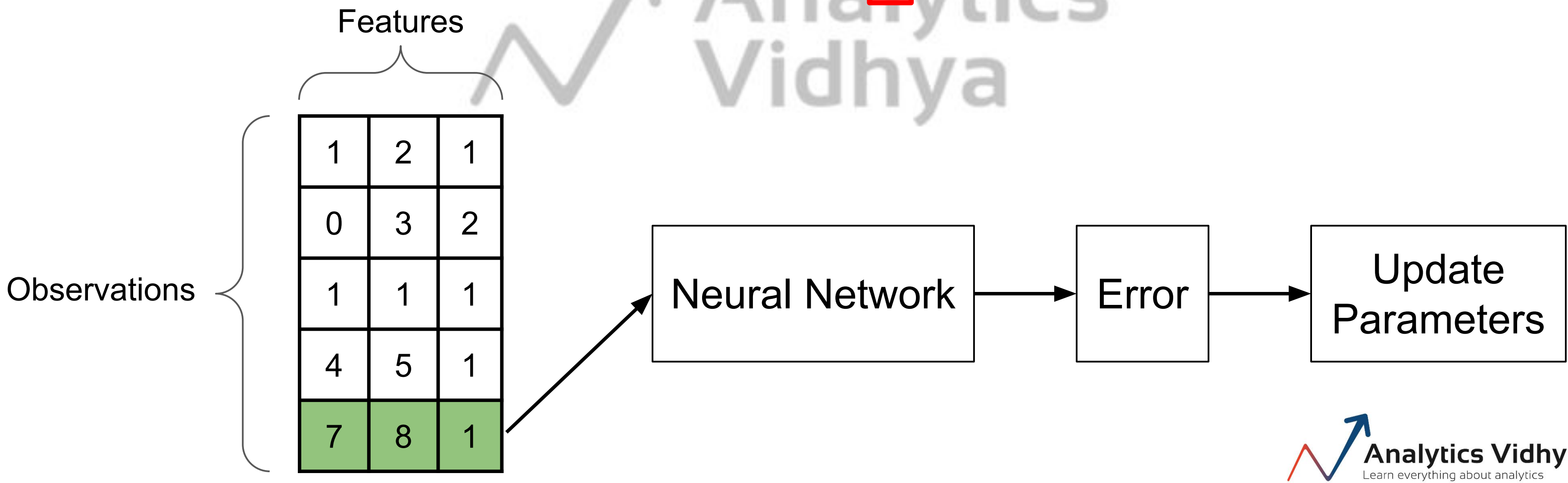
Batch Size = 2

Parameter

Learning Rate

Gradient / Slope

$$\theta_i = \theta_i - \alpha * d\boxed{J} / d\theta_i$$



1	2	1
0	3	2
1	1	1
4	5	1
7	8	1

Variants of Gradient Descent

Parameter

Learning Rate

Gradient / Slope

$$\theta_i = \theta_i - \alpha * d\boxed{J} / d\theta_i$$

Entire Training Set (m)	Batch Gradient Descent
Single Observation (1)	Stochastic Gradient Descent (SGD)
1 < x < m	Mini-Batch Gradient Descent

Comparison: Variants of Gradient Descent

Batch Gradient Descent	Stochastic Gradient Descent (SGD)	Mini-Batch Gradient Descent

Comparison: Number of observations used for updation

Batch Gradient Descent	Stochastic Gradient Descent (SGD)	Mini-Batch Gradient Descent
● Entire dataset for updation		

Comparison: Number of observations used for updation

Batch Gradient Descent

- Entire dataset for updation

Stochastic Gradient Descent (SGD)

- Single observation for updation

Mini-Batch Gradient Descent

Comparison: Number of observations used for updation

Batch Gradient Descent

- Entire dataset for updation

Stochastic Gradient Descent (SGD)

- Single observation for updation

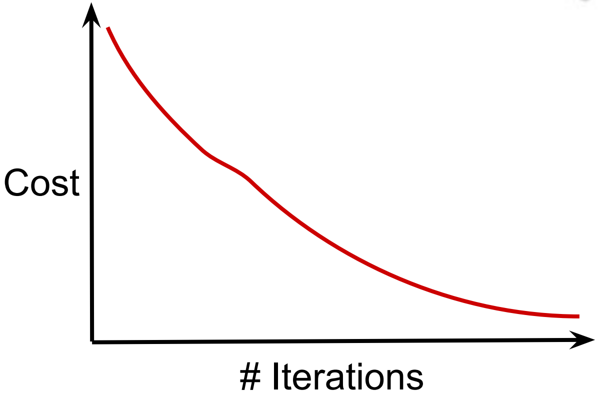
Mini-Batch Gradient Descent

- Subset of data for updation

Comparison: Cost Function

Batch Gradient Descent

- Entire dataset for updation
- Cost function reduces smoothly



Stochastic Gradient Descent (SGD)

- Single observation for updation

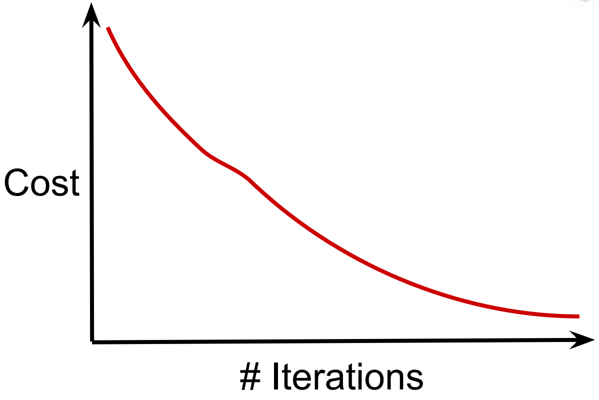
Mini-Batch Gradient Descent

- Subset of data for updation

Comparison: Cost Function

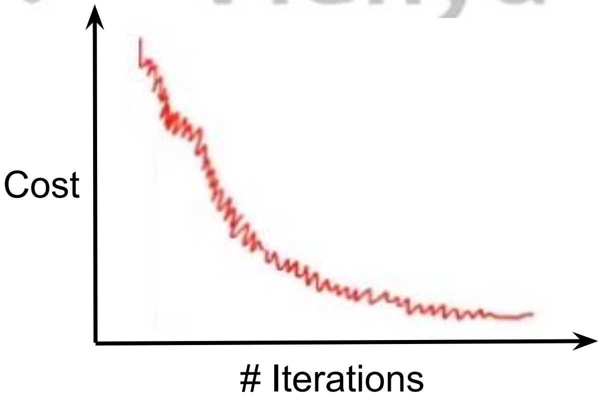
Batch Gradient Descent

- Entire dataset for updation
- Cost function reduces smoothly



Stochastic Gradient Descent (SGD)

- Single observation for updation
- Lot of variations in cost function



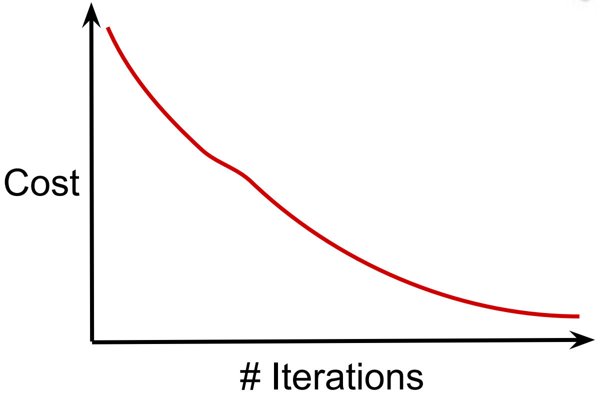
Mini-Batch Gradient Descent

- Subset of data for updation

Comparison: Cost Function

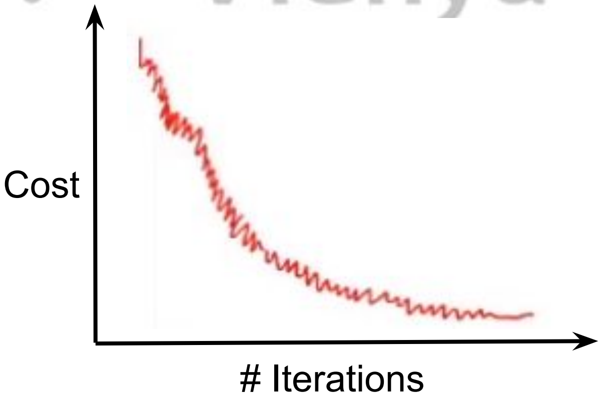
Batch Gradient Descent

- Entire dataset for updation
- Cost function reduces smoothly



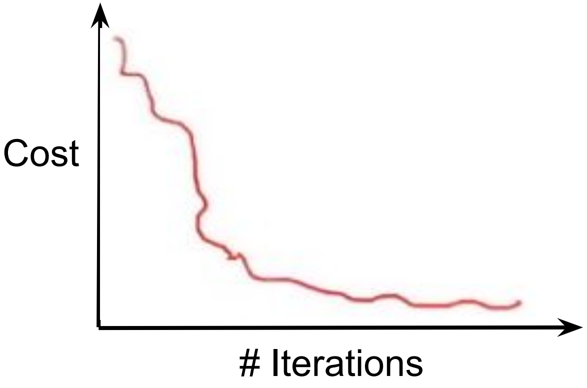
Stochastic Gradient Descent (SGD)

- Single observation for updation
- Lot of variations in cost function



Mini-Batch Gradient Descent

- Subset of data for updation
- Smoother cost function as compared to SGD



Comparison: Computation cost and time

Batch Gradient Descent

- Entire dataset for updation
- Cost function reduces smoothly
- Computation cost is very high

Stochastic Gradient Descent (SGD)

- Single observation for updation
- Lot of variations in cost function

Mini-Batch Gradient Descent

- Subset of data for updation
- Smoother cost function as compared to SGD

Comparison: Computation cost and time

Batch Gradient Descent	Stochastic Gradient Descent (SGD)	Mini-Batch Gradient Descent
● Entire dataset for updation	● Single observation for updation	● Subset of data for updation
● Cost function reduces smoothly	● Lot of variations in cost function	● Smoother cost function as compared to SGD
● Computation cost is very high	● Computation time is more	

Comparison: Computation cost and time

Batch Gradient Descent	Stochastic Gradient Descent (SGD)	Mini-Batch Gradient Descent
● Entire dataset for updation	● Single observation for updation	● Subset of data for updation
● Cost function reduces smoothly	● Lot of variations in cost function	● Smoother cost function as compared to SGD
● Computation cost is very high	● Computation time is more	● Computation time is lesser than SGD
		● Computation cost is lesser than Batch Gradient Descent

Thank You