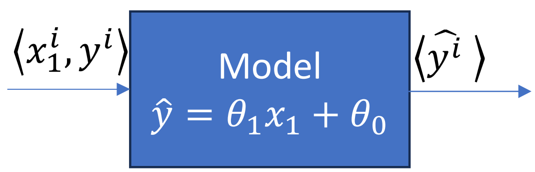
Regression

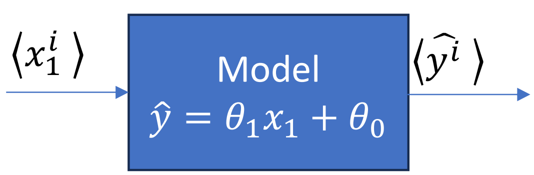
Regression

- The goal of regression is to learn a mapping function that can predict the target label / output variable (dependent variable) based on one or more input features / variables (independent variables).
- The relationship between the input variables and the output variable is typically represented by a mathematical equation.
- Regression algorithms aim to minimize the difference between the actual observed values and the values predicted by the model.
- This is often done by optimizing a loss /cost function, such as mean squared error (MSE) during the training process.

Regression

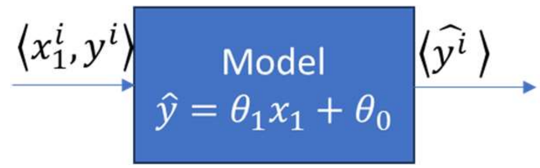
- Consider the following dataset.
- $\hat{y} = h_\theta(x) = \theta_1 x_1 + \theta_0$
- $\hat{y} = h_\theta(x) = \theta_1 x_1 + \theta_0 x_0$



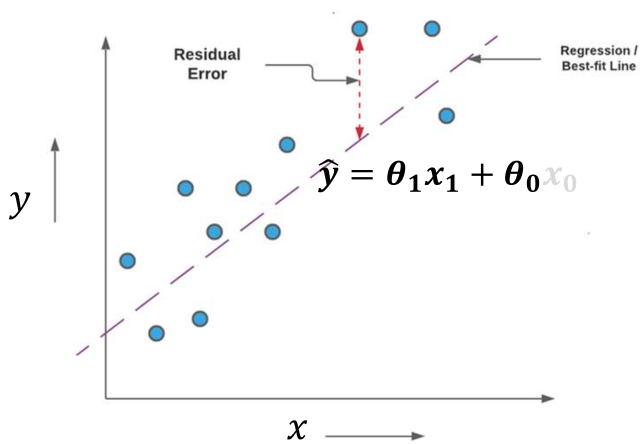


x_0	No. of hours of study (x_1)	CGPA (y)
1	1	2
1	2	3.5
1	3	5
1	4	6.5
1	5	8
1	6	9.5

Regression

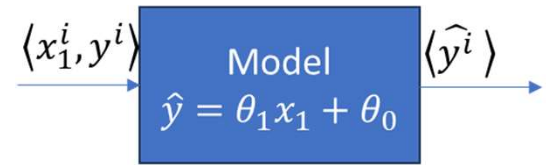


- $\hat{y} = \theta_1 x_1 + \theta_0 x_0$
- Assume θ_0, θ_1
- $\theta_0 = 0.1, \theta_1 = 0.2$
- Compute $\hat{y} = h_\theta(x)$



No. of hours of study (x_1)	CGPA (y)	$h(\theta) = \hat{y}$	$(\hat{y} - y)^2$
1	2	0.3	2.89
2	3.5	0.5	9
3	5	0.7	18.5
4	6.5	0.9	31.4
5	8	1.1	47.6
6	9.5	1.3	67.3
Total			176.7

Regression

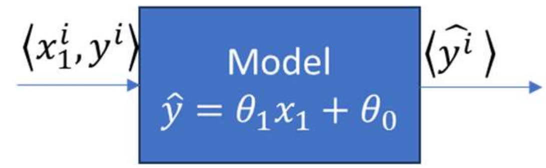


No. of hours of study (x_1)	CGPA (y)	$\hat{y}$	$(\hat{y} - y)^2$
1	2	0.3	2.89
2	3.5	0.5	9
3	5	0.7	18.5
4	6.5	0.9	31.4
5	8	1.1	47.6
6	9.5	1.3	67.3
Total			

- *Minimize*
- $J(\theta) = \frac{1}{2n}\sum_{i=1}^{n}\left(h_\theta(x) - y^i\right)^2$
- $J(\theta) = \frac{1}{2n}\sum_{i=1}^{n}\left(\widehat{y^i} - y^i\right)^2$
- $J(\theta) = \frac{1}{2n}\sum_{i=1}^{n}\left((\theta_1 x_1^i + \theta_0 x_0^i) - y^i\right)^2$
- $J(\theta) = \frac{1}{2n}\sum_{i=1}^{n}(\sum_{j=0}^{d}\theta_j\, x_j^i) - y^i)^2$

$\bar{x} = [x_0, x_1, \dots x_d]$ $\quad \bar{\theta} = [\theta_0, \theta_1, \dots \theta_d]$

Regression

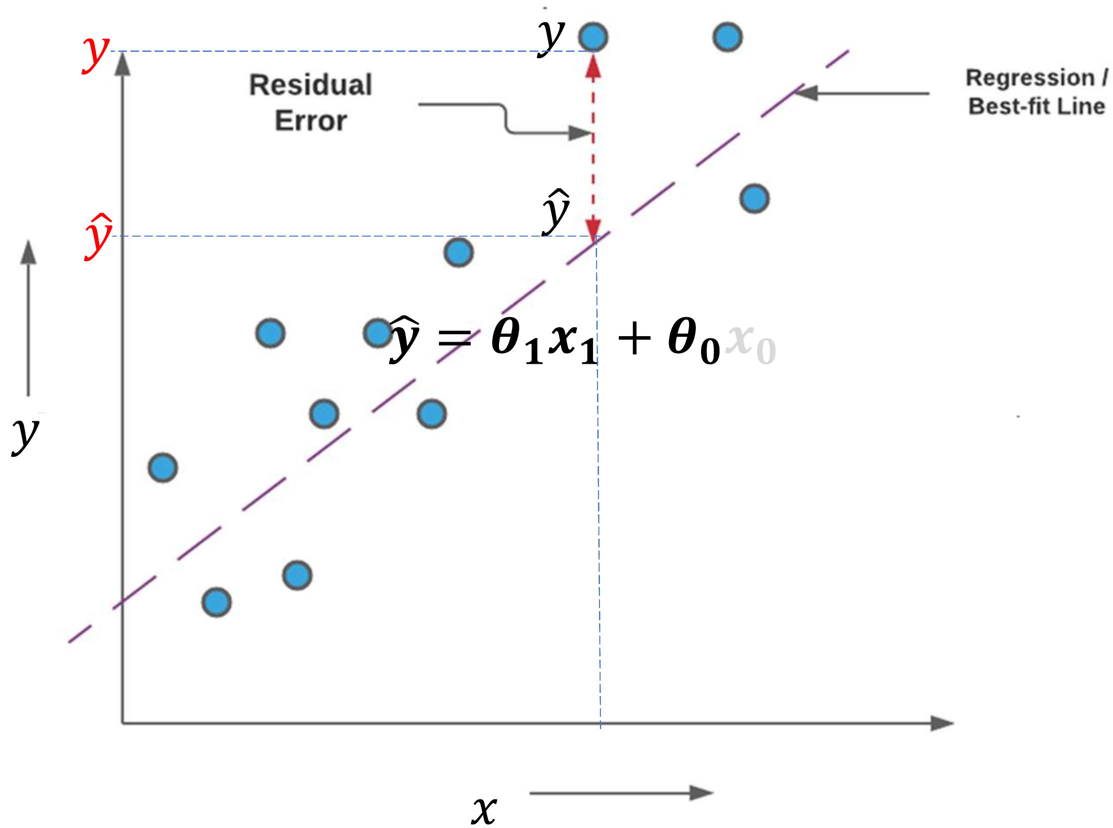


- $Minimize$
- $J(\bar{\theta}) = \frac{1}{2n} \sum_{i=1}^{n} \left(h_\theta(\bar{x}) - y^i\right)^2$
- $J(\bar{\theta}) = \frac{1}{2n} \sum_{i=1}^{n} \left(\widehat{y^i} - y^i\right)^2$
- $J(\bar{\theta}) = \frac{1}{2n} \sum_{i=1}^{n} \left((\theta_1 x_1^i + \theta_0 x_0^i) - y^i\right)^2$
- $J(\bar{\theta}) = \frac{1}{2n} \sum_{i=1}^{n} \left(\sum_{j=0}^{d} \theta_j \, x_j^i - y^i\right)^2$

$\bar{x} = [x_0, x_1, \dots x_d]$ $\quad \bar{\theta} = [\theta_0, \theta_1, \dots \theta_d]$

No. of hours of study (x_1)	CGPA (y)	$\hat{y}$	$(\hat{y} - y)^2$
1	2	0.3	2.89
2	3.5	0.5	9
3	5	0.7	18.5
4	6.5	0.9	31.4
5	8	1.1	47.6
6	9.5	1.3	67.3
Total			

Regression

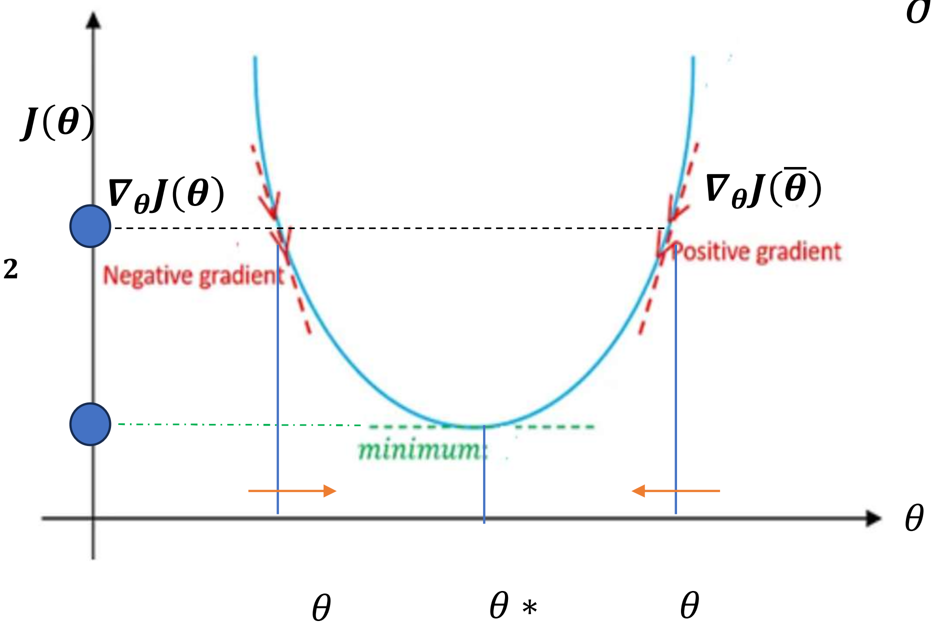


y
y
Residual Error
Regression / Best-fit Line
$\hat{y}$
$\hat{y}$
$\hat{y} = \theta_1 x_1 + \theta_0 x_0$
y
x

Gradient Descent

$$\nabla_\theta J(\theta) = \frac{\partial}{\partial\theta} J(\theta)$$

$$\boldsymbol{J(\theta)} = \frac{1}{2n}\sum_{i=1}^{n}\left((\theta_1 x_1^i + \theta_0) - y^i\right)^2$$



- If the slope/gradient at the current value of $\theta > 0$, this means that we are to the right of optimal $\theta *$.

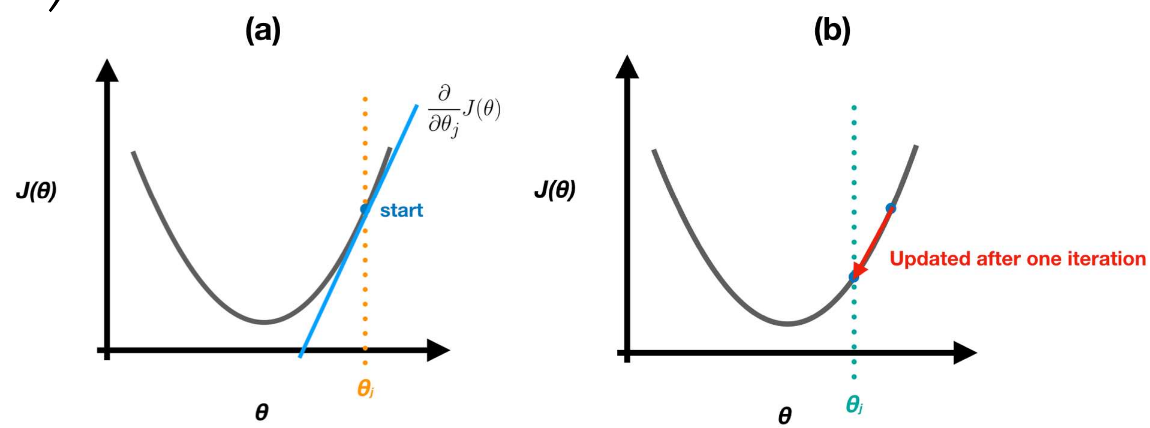
Gradient Descent

$$J(\theta) = \frac{1}{2n}\sum_{i=1}^{n}\left(\sum_{j=0}^{d}\theta_j\, x_j^i - y^i\right)^2$$

Repeat until converge {

$$\boxed{\theta_j} := \boxed{\theta_j} - \alpha \boxed{\frac{\partial}{\partial \theta_j} J(\theta)}$$

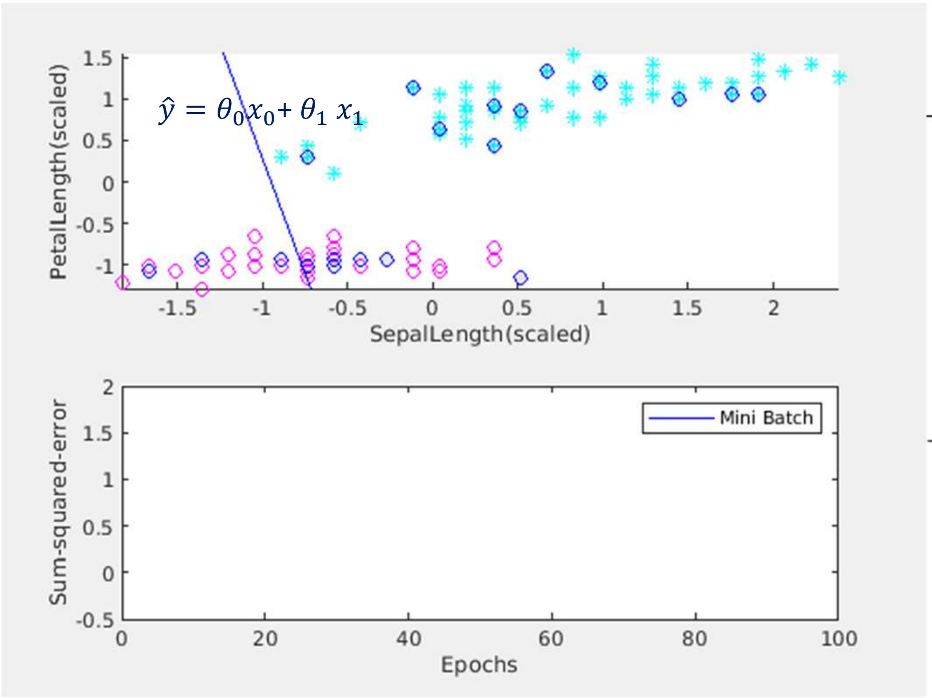
} where j represents the feature index number.

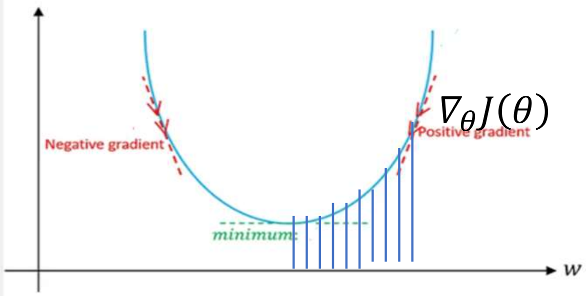


Gradient Descent

$$J(\boldsymbol{\theta}) = \frac{1}{2n}\sum_{i=1}^{n}\left((\boldsymbol{\theta}_1 x_1^i + \boldsymbol{\theta}_0 x_0^i) - y^i\right)^2$$

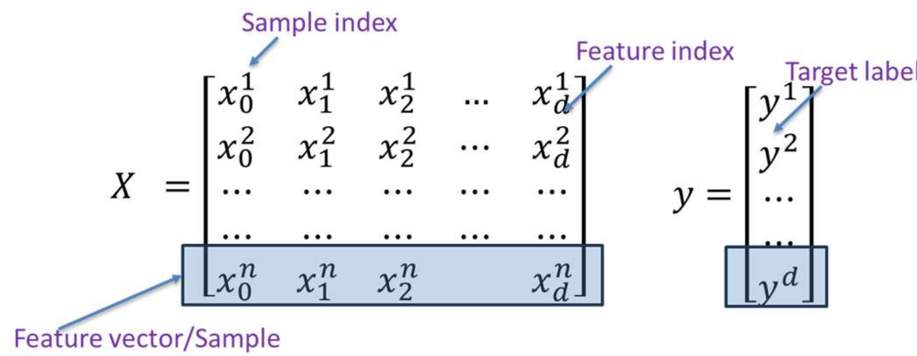
$$\nabla_\theta J(\theta) = \frac{\partial}{\partial\theta} J(\theta)$$





Gradient Descent

- $\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{2n} \sum_{i=1}^{n} \left(\sum_{j=1}^{d} \theta_j x_j^i - y^i \right)^2$
- $\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{2n} \sum_{i=1}^{n} 2 * \left(\sum_{j=1}^{d} \theta_j x_j^i - y^i \right) * \frac{\partial}{\partial \theta_j} \left(\theta_0 x_0^i + \cdots + \theta_j x_j^i + \cdots \theta_d x_d^i - y^i \right)$
- $\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{n} \sum_{i=1}^{n} \left(\widehat{y}^i - y^i \right) * x_j^i$
- $\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{n} \sum_{i=1}^{n} e^i * x_j^i$
- $n = 1,$ *Stochstic gradient descent*
- $\frac{\partial J(\theta)}{\partial \theta_j} = e * x_j$



Gradient Descent

$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{n}\sum_{i=1}^{n} e^i * x_j^i$$

$$\frac{\partial J(\theta)}{\partial \theta_0} = \frac{1}{n}\sum_{i=1}^{n} e^i * x_0^i$$

$$\frac{\partial J(\theta)}{\partial \theta_1} = \frac{1}{n}\sum_{i=1}^{n} e^i * x_1^i$$

x_0	No. of hours of study (x_1)	CGPA (y)	$\hat{y}$	$e = \hat{y} - y$	$(y - \hat{y})x_0$	$(y - \hat{y})x_0$
1	1	2	0.3	-1.7	-1.7	-1.7
1	2	3.5	0.5	-3	-3	-6
1	3	5	0.7	-4.3	-4.3	-12.9
1	4	6.5	0.9	-5.6	-5.6	-22.4
1	5	8	1.1	-6.9	-6.9	-34.5
1	6	9.5	1.3	-8.2	-8.2	-49.2
	Total				-29.7	-126.7

Gradient Descent

- $\left[\frac{\partial J(\theta)}{\partial \theta_0}, \quad \dots, \quad \dots, \quad \frac{\partial J(\theta)}{\partial \theta_d}\right] = e * [x_0, .., x_d]$
- $[\theta_0^{new}, \dots, \dots, \quad \theta_d^{new}] = [\theta_0^{old}, \dots, \dots, \quad \theta_d^{old}] - \alpha \left[\frac{\partial J(\theta)}{\partial \theta_0}, \dots, \dots, \quad \frac{\partial J(\theta)}{\partial \theta_d}\right]$
- $[\theta_0^{ne} \quad , \dots, \dots, \quad \theta_d^{new}]_{1\times d} = [\theta_0^{old}, \dots, \dots, \quad \theta_d^{old}]_{1\times d} - \alpha * e * [x_0, .., x_d]_{1\times d}$

Sample index

Feature index

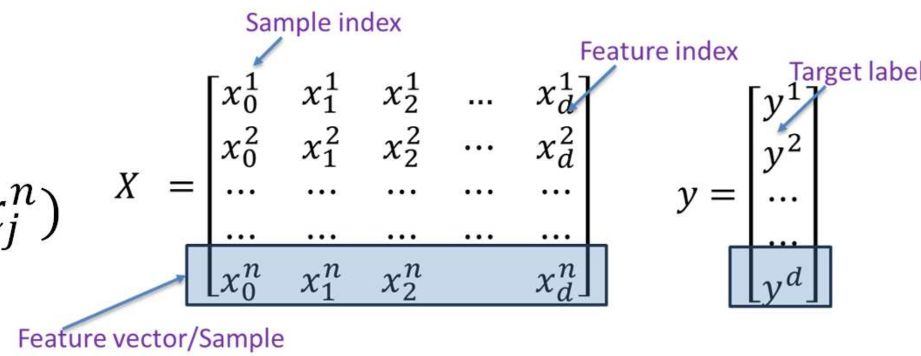
Target label

$$X = \begin{bmatrix} x_0^1 & x_1^1 & x_2^1 & \dots & x_d^1 \\ x_0^2 & x_1^2 & x_2^2 & \dots & x_d^2 \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ x_0^n & x_1^n & x_2^n & & x_d^n \end{bmatrix} \qquad y = \begin{bmatrix} y^1 \\ y^2 \\ \dots \\ \dots \\ y^d \end{bmatrix}$$

Feature vector/Sample

Gradient Descent

- $\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{2n} \sum_{i=1}^{n} \left(\sum_{j=1}^{d} \theta_j x_j^i - y^i \right)^2$
- $\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{2n} \sum_{i=1}^{n} 2 * \left(\sum_{j=1}^{d} \theta_j x_j^i - y^i \right) * \frac{\partial}{\partial \theta_j} \left(\theta_0 x_0^i + \cdots + \theta_j x_j^i + \cdots \theta_d x_d^i - y^i \right)$
- $\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{n} \sum_{i=1}^{n} (\widehat{y}^i - y^i) * x_j^i$
- $\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{n} \sum_{i=1}^{n} e^i * x_j^i$
- $\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{n} (e^1 x_j^1 + \cdots + e^p x_j^p + \cdots e^n x_j^n)$



Gradient Descent

- $\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{n}\sum_{i=1}^{n} e^i * x_j^i$
- $\left[\frac{\partial J(\theta)}{\partial \theta_0}, \quad \dots, \quad \dots, \quad \frac{\partial J(\theta)}{\partial \theta_d}\right] = \frac{1}{n}\sum_{i=1}^{n} e^i \left[x_0^i, \quad \dots, \quad \dots, \quad x_d^i\right]$
- $[\theta_0^{new}, \dots, \dots, \quad \theta_d^{new}] = [\theta_0^{old}, \dots, \dots, \quad \theta_d^{old}] - \alpha \left[\frac{\partial J(\theta)}{\partial \theta_0}, \dots, \dots, \quad \frac{\partial J(\theta)}{\partial \theta_d}\right]$

Sample index

Feature index

Target label

$$X = \begin{bmatrix} x_0^1 & x_1^1 & x_2^1 & \dots & x_d^1 \\ x_0^2 & x_1^2 & x_2^2 & \dots & x_d^2 \\ \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \boxed{x_0^n} & \boxed{x_1^n} & \boxed{x_2^n} & & \boxed{x_d^n} \end{bmatrix} \qquad y = \begin{bmatrix} y^1 \\ y^2 \\ \dots \\ \dots \\ \boxed{y^d} \end{bmatrix}$$

Feature vector/Sample

Regression-Vectorized Computations

- $\hat{y} = \theta_0 x_0 + \theta_1 x_1$
- $\theta = [\theta_0, \theta_1]$
- $x = [x_0, x_1]$
- $\hat{y} = np.dot(\theta, x)$

Regression-Vectorized Computations

- $\begin{bmatrix} \widehat{y^1} \\ \cdots \\ \cdots \\ \widehat{y^n} \end{bmatrix} = \begin{bmatrix} x_0^1 & x_1^1 \\ \cdots & \cdots \\ \cdots & \cdots \\ x_0^n & x_1^n \end{bmatrix} \circ [\theta_0 \quad \theta_1] = \begin{bmatrix} \theta_0 x_0^1 & + & \theta_1 x_1^1 \\ \cdots & + & \cdots \\ \cdots & + & \cdots \\ \theta_0 x_0^n & + & \theta_1 x_1^n \end{bmatrix}$

- $\begin{bmatrix} e^1 \\ \cdots \\ \cdots \\ e^n \end{bmatrix} = \begin{bmatrix} \widehat{y^1} \\ \cdots \\ \cdots \\ \widehat{y^n} \end{bmatrix} - \begin{bmatrix} y^1 \\ \cdots \\ \cdots \\ y^n \end{bmatrix}$

Regression-Vectorized Computations

- $\begin{bmatrix} \frac{\partial J(\theta)}{\partial \theta_0} & \frac{\partial J(\theta)}{\partial \theta_1} \end{bmatrix} = np.sum\left(\begin{bmatrix} e^1 \\ \dots \\ \dots \\ e^n \end{bmatrix} * \begin{bmatrix} x_0^1 & x_1^1 \\ \dots & \dots \\ \dots & \dots \\ x_0^n & x_1^n \end{bmatrix}, axis = 0 \right) / n$
- $\begin{bmatrix} \frac{\partial J(\theta)}{\partial \theta_0} & \frac{\partial J(\theta)}{\partial \theta_1} \end{bmatrix} = np.sum\left(\begin{bmatrix} e^1 x_0^1 & e^1 x_1^1 \\ \dots & \dots \\ \dots & \dots \\ e^1 x_0^n & e^1 x_1^n \end{bmatrix}, axis = 0 \right) / n$
- $[\theta_0^{new} \quad \theta_1^{new}] = [\theta_0^{old} \quad \theta_1^{old}] - \alpha \begin{bmatrix} \frac{\partial J(\theta)}{\partial \theta_0} & \frac{\partial J(\theta)}{\partial \theta_1} \end{bmatrix}$

Regression-Metrics

- R^2
- In regression analysis, the coefficient of determination, denoted as R^2, is a statistical measure that represents the proportion of the variance in the dependent variable that is predictable from the independent variables.
- In simpler terms, it quantifies the goodness of fit of the regression model to the data.

Regression-Metrics

- R^2
- 0 indicates that the model does not explain any of the variability of the target valuesaround its mean.
- 1 indicates that the model explains all the variability of the target values around its mean.
- We calculate the value of R^2 using the test dataset.

Regression-Metrics

- R^2
- $SS_{tot} = \sum_{i=1}^{m}\left(y_{test}^{i} - \overline{y_{test}}\right)^2, m\ is\ the\ no.of\ test\ samples$
- $SS_{res} = \sum_{i=1}^{m}\left(y_{test}^{i} - \widehat{y_{pred}^{i}}\right)^2$
- $R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$