Normal Equation Solution

Closed-form solution for linear regression

Matrix Form

$$Y = X\beta + \varepsilon$$

X is the design matrix

↓

Normal Equation

$$X^TX\beta = X^TY$$

↓

General Solution

$$\hat{\beta} = (X^TX)^{-1}X^TY$$

When X^TX is invertible

Simple Regression

$$\hat{\beta_1} = \text{Cov}(X,Y) \,/\, \text{Var}(X)$$

$$\hat{\beta_0} = \bar{Y} - \hat{\beta_1}\,\bar{X}$$

Complexity

Computational cost

$$O(np^2 + p^3)$$

n: samples
p: features

Alternative

For large-scale:
Gradient Descent
(iterative method)

Advantage

Closed-form
Direct computation
No iterations needed

Gradient Descent Process

Parameter update with MSE Loss - Step by Step

Mean Squared Error (MSE) Loss

$$L(\beta) = (1/n)\Sigma(y_i - \hat{y}_i)^2 = (1/n)\Sigma(y_i - \beta_0 - \beta_1 x_i)^2$$

↓

1 Compute Partial Derivatives

$$\partial L/\partial \beta_0 = -(2/n)\Sigma(y_i - \beta_0 - \beta_1 x_i)$$

$$\partial L/\partial \beta_1 = -(2/n)\Sigma x_i(y_i - \beta_0 - \beta_1 x_i)$$

Calculate gradient for each parameter

2 Evaluate at Current Parameters

$$g_0 = \partial L/\partial \beta_0 | \beta^{(t)}, \quad g_1 = \partial L/\partial \beta_1 | \beta^{(t)}$$

Substitute current $\beta_0^{(t)}$ and $\beta_1^{(t)}$ values

3 Update Parameters

$$\beta_0^{(t+1)} = \beta_0^{(t)} - \alpha \cdot g_0$$

$$\beta_1^{(t+1)} = \beta_1^{(t)} - \alpha \cdot g_1$$

α is the learning rate (step size)

4 Check Convergence

$$|L^{(t+1)} - L^{(t)}| < \varepsilon \text{ or } ||\nabla L|| < \varepsilon$$

Repeat steps 1-3 until convergence criterion is met

Learning Rate α

Controls step size

Too large: diverge

Too small: slow

Matrix Form

$$\beta^{(t+1)} = \beta^{(t)} - \alpha \cdot \nabla L$$

$$\nabla L = -(2/n)X^T(Y - X\beta)$$

Convergence

Iterative approach

Works for large-scale

O(np) per iteration