Linear Regression Algorithms

Kashyap Sureshchandra Patel

September 7, 2025

Algorithm 1: Linear Regression with Gradient Descent and Regularization

---

Input: Learning rate η, number of epochs T, batch size b, regularization flag r, regularization method $m \in \{\text{L1, L2, ElasticNet}\}$, regularization strength λ, mixing parameter α

Output: Optimal parameter vector θ

Initialize: $\theta \leftarrow \mathbf{0}$

Add bias column of ones to X: $X_b = [\mathbf{1}, X]$

$n \leftarrow$ number of samples, $d \leftarrow$ number of features (including bias)

for *epoch* $= 1$ *to* T **do**

- Shuffle dataset (X_b, y) randomly
- **for** *each mini-batch* (X_{batch}, y_{batch}) *of size* b **do**
  - $y_{\text{pred}} \leftarrow X_{\text{batch}} \cdot \theta$
  - Gradient: $\nabla J(\theta) \leftarrow \frac{2}{b} X_{\text{batch}}^{\top}(y_{\text{pred}} - y_{\text{batch}})$
  - **if** $r = \mathit{True}$ **then**
    - $\theta_{\text{reg}} \leftarrow \theta$, set $\theta_{\text{reg},0} \leftarrow 0$
    - **if** $m = \mathit{L1}$ **then**
      - $\nabla J(\theta) \leftarrow \nabla J(\theta) + \lambda \cdot \text{sign}(\theta_{\text{reg}})$
    - **else if** $m = \mathit{L2}$ **then**
      - $\nabla J(\theta) \leftarrow \nabla J(\theta) + 2\lambda \cdot \theta_{\text{reg}}$
    - **else if** $m = \mathit{ElasticNet}$ **then**
      - $\nabla J(\theta) \leftarrow \nabla J(\theta) + \lambda\Big(\alpha \cdot \text{sign}(\theta_{\text{reg}}) + 2(1-\alpha) \cdot \theta_{\text{reg}}\Big)$
  - Update step: $\theta \leftarrow \theta - \eta \cdot \nabla J(\theta)$

Algorithm 2: Linear Regression with Normal Equation (OLS and Ridge)

Input: Training data (X, y), regularization method $m \in \{\text{None, L2}\}$, regularization strength λ
Output: Optimal parameter vector θ

Preprocessing:
Add bias column of ones to X: $X_b = [\mathbf{1}, X]$
$n \leftarrow$ number of samples, $d \leftarrow$ number of features (including bias)

if $m = \textit{None (OLS)}$ **then**
  $\theta \leftarrow (X_b^\top X_b)^\dagger X_b^\top y$
else if $m = \textit{L2 (Ridge)}$ **then**
  $I_d \leftarrow d \times d$ identity matrix
  Set $I_{d,0,0} \leftarrow 0$ (do not regularize bias)
  $\theta \leftarrow \left(X_b^\top X_b + n\lambda I_d\right)^\dagger X_b^\top y$

return θ