Linear Regression

$$Y = XW + b$$

Derivation of the Normal Equation (Finding the Direct Solution)

Note: This is computationally expensive for large datasets because it involves inverting a matrix. In practice, iterative methods like gradient descent are often preferred for large-scale problems. We want to minimize the mean squared error loss function:

$$J = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$$

Where $\hat{Y}_i = X_i W + b$ To find the optimal parameters W and b, we take the derivative of the loss function with respect to W and b, set them to zero, and solve for the parameters.

First we need to add the bias term into the X matrix by adding a column of ones:

$$X_{new} = [\mathbf{1}|X]$$

Now we can rewrite the prediction as:

$$\hat{Y} = X_{new}W$$

Where $\theta = [b; W]$ is the parameter vector including both the bias and weights. The loss function can be rewritten in matrix form:

$$J = \frac{1}{N}(Y - X_{new}W)^T(Y - X_{new}W)$$

Taking the derivative of J with respect to W step by step:

$$\frac{\partial J}{\partial W} = \frac{\partial}{\partial W}\left(\frac{1}{N}(Y - X_{new}W)^T(Y - X_{new}W)\right)$$

The chain rule is defined as: If $z = f(y)$ and $y = g(x)$, then:

$$\frac{dz}{dx} = \frac{dz}{dy} \cdot \frac{dy}{dx}$$

The product rule is defined as: If $z = u(x)v(x)$, then:

$$\frac{dz}{dx} = u'(x)v(x) + u(x)v'(x)$$

Using these rules step by step, first we start with the outer function: Let $E = Y - X_{new}W$, then:

$$J = \frac{1}{N}E^T E$$

Taking the derivative with respect to W:

$$\frac{\partial J}{\partial W} = \frac{1}{N}\left(\frac{\partial E^T}{\partial W}E + E^T\frac{\partial E}{\partial W}\right)$$

Next, we need to find $\frac{\partial E}{\partial W}$: Since $E = Y - X_{new}W$, we have:

$$\frac{\partial E}{\partial W} = -X_{new}$$

Substituting this back into the derivative of J:

$$\frac{\partial J}{\partial W} = \frac{1}{N}\left(-X_{new}^T E + E^T(-X_{new})\right)$$

Since $E^T(-X_{new})$ is the transpose of $-X_{new}^T E$, we can combine the two terms:

$$\frac{\partial J}{\partial W} = \frac{1}{N}\left(-2X_{new}^T E\right)$$

Substituting back $E = Y - X_{new}W$:

$$\frac{\partial J}{\partial W} = \frac{1}{N}\left(-2X_{new}^T (Y - X_{new}W)\right)$$

Setting the derivative to zero to find the optimal weights:

$$-2X_{new}^T (Y - X_{new}W) = 0$$

Solving for W:

$$X_{new}^T Y = X_{new}^T X_{new} W$$
$$W = (X_{new}^T X_{new})^{-1} X_{new}^T Y$$

Regularization in Linear Regression

To prevent overfitting, we can add a regularization term to the loss function. Two common types of regularization are Lasso (L1) and Ridge (L2).

- **Ridge Regression (L2 Regularization):**

  $$J_{ridge} = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 + \lambda \sum_{j=1}^{p} W_j^2$$

  The closed-form solution for Ridge Regression is:

  $$W_{ridge} = (X_{new}^T X_{new} + \lambda I)^{-1} X_{new}^T Y$$

  Where I is the identity matrix and λ is the regularization parameter.

  In short, we arrived at this direct solution because when we took the derivative of the loss function with respect to W, the regularization term added $2\lambda W$ to the derivative. Setting the derivative to zero and solving for W led to the modified normal equation above.

- **Lasso Regression (L1 Regularization):**

  $$J_{lasso} = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 + \lambda \sum_{j=1}^{p} |W_j|$$

  Lasso does not have a closed-form solution like Ridge Regression. Instead, it is typically solved using optimization algorithms such as coordinate descent or least angle regression (LARS).