Mathematical Description of Linear Regression

Linear regression is a statistical method used to model the relationship between a dependent variable y and one or more independent variables $\mathbf{x}$. It assumes a linear relationship between the variables and seeks to find the best-fitting line that minimizes prediction errors. Below is the mathematical formulation.

1. Model Representation

For a dataset with n features, the dependent variable $y \in \mathbb{R}$ is modeled as a linear combination of the feature vector $\mathbf{x} \in \mathbb{R}^n$:

$$y = \mathbf{w}^T \mathbf{x} + b + \epsilon$$

- $\mathbf{w} \in \mathbb{R}^n$: Weight vector, representing the coefficients of the features.
- $b \in \mathbb{R}$: Bias term (intercept).
- $\mathbf{w}^T \mathbf{x} + b$: Predicted value, denoted as $\hat{y}$.
- ϵ: Random error term, assumed to be normally distributed with mean 0 and variance σ^2.

For a dataset with m samples $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^m$, the prediction for the i-th sample is:

$$\hat{y}^{(i)} = \mathbf{w}^T \mathbf{x}^{(i)} + b$$

2. Objective Function

The goal is to find the parameters $\mathbf{w}$ and b that minimize the prediction error. The **Mean Squared Error (MSE)** is commonly used as the loss function:

$$J(\mathbf{w}, b) = \frac{1}{m} \sum_{i=1}^{m} (y^{(i)} - \hat{y}^{(i)})^2 = \frac{1}{m} \sum_{i=1}^{m} (y^{(i)} - (\mathbf{w}^T \mathbf{x}^{(i)} + b))^2$$

The objective is to minimize $J(\mathbf{w}, b)$ with respect to $\mathbf{w}$ and b.

3. Optimization

The loss function $J(\mathbf{w}, b)$ is convex, allowing optimization through **closed-form solutions** or **iterative methods** such as gradient descent.

3.1 Closed-Form Solution

Linear regression allows the optimal parameters to be derived analytically using the **normal equation**. Let $\mathbf{X} \in \mathbb{R}^{m \times (n+1)}$ be the design matrix, where each row is $[\mathbf{x}^{(i)}, 1]$ (appending a 1 for the bias term), and $\mathbf{y} \in \mathbb{R}^m$ the vector of target values. The parameters $\boldsymbol{\theta} = [\mathbf{w}, b]$ are computed as:

$$\boldsymbol{\theta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

This solution assumes that $\mathbf{X}^T\mathbf{X}$ is invertible, requiring that the features are not perfectly collinear and $m \geq n + 1$.

3.2 Gradient Descent

Alternatively, gradient descent iteratively updates the parameters to minimize the loss. The gradients of the loss function are:

$$\frac{\partial J}{\partial \mathbf{w}} = \frac{2}{m} \sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)}) \mathbf{x}^{(i)}$$

$$\frac{\partial J}{\partial b} = \frac{2}{m} \sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)})$$

The parameters are updated as:

$$\mathbf{w} \leftarrow \mathbf{w} - \alpha \frac{\partial J}{\partial \mathbf{w}}, \quad b \leftarrow b - \alpha \frac{\partial J}{\partial b}$$

where α is the learning rate.

4. Regularization

To prevent overfitting, especially when n is large or features are correlated, regularization can be applied. Common approaches include:

- **L2 Regularization (Ridge Regression)**: Adds a penalty on the magnitude of the weights to the loss function:

$$J(\mathbf{w}, b) = \frac{1}{m} \sum_{i=1}^{m} (y^{(i)} - (\mathbf{w}^T \mathbf{x}^{(i)} + b))^2 + \lambda \|\mathbf{w}\|^2$$

  - λ: Regularization parameter.

  The normal equation for Ridge Regression is:

$$\boldsymbol{\theta} = (\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$$

  where $\mathbf{I}$ is the identity matrix (excluding the bias term from regularization).

- **L1 Regularization (Lasso Regression)**: Adds a penalty on the absolute values of the weights:

$$J(\mathbf{w}, b) = \frac{1}{m} \sum_{i=1}^{m} (y^{(i)} - (\mathbf{w}^T \mathbf{x}^{(i)} + b))^2 + \lambda \|\mathbf{w}\|_1$$

  Lasso encourages sparsity, setting some weights to zero, but typically requires iterative optimization.

5. Assumptions

Linear regression assumes:

- Linearity: The relationship between $\mathbf{x}$ and y is linear.
- Independence: Observations are independent.
- Homoscedasticity: Constant variance of errors.
- Normality: Errors are normally distributed (for inference purposes).
- No multicollinearity: Features are not perfectly collinear.

6. Evaluation

The model's performance is commonly evaluated using metrics such as:

- **Mean Squared Error (MSE)**: As defined above.
- **Root Mean Squared Error (RMSE)**: $\sqrt{\text{MSE}}$.
- R^2: Proportion of variance in y explained by the model:

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(y^{(i)} - \hat{y}^{(i)})^2}{\sum_{i=1}^{m}(y^{(i)} - \bar{y})^2}$$

  where $\bar{y}$ is the mean of the target values.

7. Summary

Linear regression models the relationship between features and a continuous outcome using a linear function, optimized by minimizing the mean squared error. It can be solved analytically or via gradient descent, with regularization to improve generalization. The model is simple, interpretable, and widely used for predictive tasks.