Machine Learning 101

Rajdeep Chatterjee, Ph.D.
Amygdala AI, Bhubaneswar, India *

January 2025

Linear Models with Regularization

1 L1 Norm (Manhattan Norm or Taxicab Norm)

The **L1 norm** of a vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is defined as the sum of the absolute values of its components:

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|$$

1.1 Geometrical Interpretation of L1 Norm:

In a 2D space, the L1 norm can be visualized as the distance traveled along grid lines, like a taxi moving along streets. The unit ball for the L1 norm in 2D is a diamond shape.

Thus, the unit ball for the L1 norm in 2D is described by the equation:

$$|x_1| + |x_2| = 1$$

This equation represents a diamond-shaped region centered at the origin.

L2 Norm (Euclidean Norm)

The **L2 norm** of a vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is the square root of the sum of the squares of its components:

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

*Amygdala AI, is an international volunteer-run research group that advocates for *AI for a better tomorrow* `http://amygdalaai.org/`.

1.2 Geometrical Interpretation of L2 Norm:

The L2 norm measures the straight-line (Euclidean) distance between the origin and the point **x**. In 2D, the unit ball for the L2 norm is a circle. The equation of the unit circle is:

$$x_1^2 + x_2^2 = 1$$

This represents a circle centered at the origin.

2 Lasso and Ridge Regression: Relationship to L1 and L2 Norms

In machine learning, particularly in linear regression, **Lasso Regression** and **Ridge Regression** are techniques that add regularization terms to prevent overfitting by penalizing large coefficients. These regularization terms use the **L1 norm** and **L2 norm**, respectively.

2.1 Lasso Regression (L1 Regularization)

Lasso stands for **Least Absolute Shrinkage and Selection Operator**. Lasso regression minimizes the residual sum of squares (RSS) with an additional penalty proportional to the **L1 norm** of the coefficients. This can be formulated as:

$$\hat{\beta}_{\text{Lasso}} = \underset{\beta}{\operatorname{argmin}} \left(\sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right)$$

Where:

- y_i is the target variable.
- $\mathbf{x}_i$ are the feature vectors.
- β_j are the coefficients of the model.
- λ is the regularization parameter controlling the strength of the penalty.

The λ term controls the regularization strength. When $\lambda = 0$, Lasso regression reduces to ordinary least squares regression. As λ increases, the coefficients are penalized, and many of them become exactly zero, leading to sparse solutions (feature selection).

2.2 Ridge Regression (L2 Regularization)

Ridge regression minimizes the residual sum of squares (RSS) with a penalty proportional to the **L2 norm** of the coefficients. This can be formulated as:

$$\hat{\beta}_{\text{Ridge}} = \underset{\beta}{\text{argmin}} \left(\sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right)$$

Where:

- y_i is the target variable.
- $\mathbf{x}_i$ are the feature vectors.
- β_j are the coefficients of the model.
- λ is the regularization parameter controlling the strength of the penalty.

The λ term controls the regularization strength. When $\lambda = 0$, Ridge regression reduces to ordinary least squares regression. As λ increases, the coefficients are penalized, but unlike Lasso, Ridge tends to shrink the coefficients towards zero, but they do not become exactly zero.

2.3 Geometrical Interpretation of Regularization

Both Lasso and Ridge regression introduce constraints on the optimization problem. These constraints can be interpreted geometrically:

- **Lasso (L1 Regularization):** The constraint region is a diamond in 2D. The sharp corners of the diamond lead to sparse solutions where many coefficients are exactly zero.
- **Ridge (L2 Regularization):** The constraint region is a circle (or sphere in higher dimensions). Ridge regression shrinks the coefficients smoothly towards zero without forcing them to be exactly zero.

3 Summary of the Differences

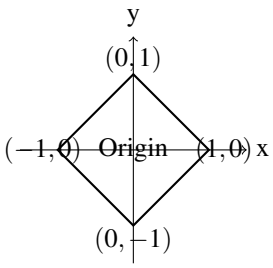
- **Lasso Regression:** Uses the L1 norm penalty and leads to sparse solutions, making it useful for feature selection.
- **Ridge Regression:** Uses the L2 norm penalty and tends to shrink coefficients towards zero, but doesn't force them to be exactly zero.

4 Summary of Geometrical Shapes for L1 and L2 Norms

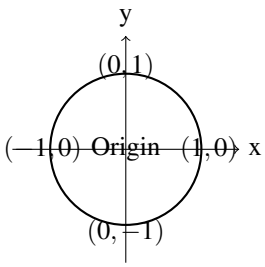
- The **L1 norm unit ball** in 2D: Diamond shape.
- The **L2 norm unit ball** in 2D: Circle shape.

In higher dimensions, the L1 norm forms a polyhedron, while the L2 norm forms a sphere.

4.1 L1 Norm Unit Ball (Diamond Shape) with Axes



4.2 L2 Norm Unit Ball (Circle Shape) with Axes



5 Mathematical Derivations and Analysis of Regression Models

5.1 Linear Regression

Linear regression models the relationship between features X and a continuous target variable y.

5.1.1 Model Definition

$$y = X\beta + \varepsilon$$

where:

- $X \in R^{m \times n}$ is the feature matrix,
- $\beta \in R^n$ is the coefficient vector,
- $\varepsilon \in R^m$ is the error term.

5.1.2 Objective Function

The goal is to minimize the Mean Squared Error (MSE):

$$J(\beta) = \frac{1}{m} \sum_{i=1}^{m} (y_i - X_i\beta)^2$$

5.1.3 Gradient Derivation

1. Expand the cost function:

$$J(\beta) = \frac{1}{m}(y - X\beta)^T(y - X\beta)$$

2. Compute the gradient:

$$\nabla_\beta J(\beta) = -\frac{2}{m}X^T(y - X\beta)$$

3. Set the gradient to zero and solve:

$$X^T X\beta = X^T y \quad \Longrightarrow \quad \beta = (X^T X)^{-1} X^T y$$

5.1.4 Merits and Demerits

Merits:

- Simple and interpretable.
- Works well with linearly separable data.

Demerits:

- Sensitive to outliers.
- Poor performance on non-linear relationships.

5.2 Logistic Regression

Logistic regression models the probability of binary outcomes using the sigmoid function.

5.2.1 Model Definition

The probability of the positive class is:

$$P(y = 1|x) = \sigma(z) = \frac{1}{1 + e^{-z}}, \quad z = X\beta$$

5.2.2 Likelihood Function

The likelihood for m examples is:

$$L(\beta) = \prod_{i=1}^{m} \sigma(z_i)^{y_i}(1 - \sigma(z_i))^{1-y_i}$$

5.2.3 Log-Likelihood

$$\ell(\beta) = \sum_{i=1}^{m} [y_i \log(\sigma(z_i)) + (1 - y_i)\log(1 - \sigma(z_i))]$$

5.2.4 Gradient and Update Rule

1. Compute the gradient:

$$\nabla_\beta \ell(\beta) = X^T(y - \sigma(X\beta))$$

2. Update using gradient ascent:

$$\beta := \beta + \alpha X^T(y - \sigma(X\beta))$$

5.2.5 Merits and Demerits

Merits:

- Probabilistic interpretation.
- Handles binary classification problems.

Demerits:

- Assumes linear decision boundary.
- Not suitable for multi-class problems without extensions.

5.3 Ridge Regression (L2 Regularization)

Ridge regression adds an L2 penalty to the cost function.

5.3.1 Objective Function

$$J(\beta) = \frac{1}{m}\sum_{i=1}^{m}(y_i - X_i\beta)^2 + \lambda\|\beta\|_2^2$$

5.3.2 Solution

The closed-form solution is:

$$\beta = (X^TX + \lambda I)^{-1}X^Ty$$

5.3.3 Merits and Demerits

Merits:

- Reduces overfitting.
- Keeps all features in the model.

Demerits:

- Coefficients are not sparse.
- Sensitive to the choice of λ.

5.4 Lasso Regression (L1 Regularization)

Lasso regression adds an L1 penalty to the cost function.

5.4.1 Objective Function

$$J(\beta) = \frac{1}{m}\sum_{i=1}^{m}(y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{n}|\beta_j|$$

5.4.2 Optimization

Lasso does not have a closed-form solution and requires iterative methods like coordinate descent.

5.4.3 Merits and Demerits

Merits:

- Performs feature selection.
- Effective for sparse data.

Demerits:

- Can remove important features.
- Computationally intensive.

5.5 Elastic Net

Elastic Net combines L1 and L2 regularization for more flexibility.

5.5.1 Objective Function

$$J(\beta) = \frac{1}{m}\sum_{i=1}^{m}(y_i - X_i\beta)^2 + \lambda_1 \sum_{j=1}^{n}|\beta_j| + \lambda_2 \sum_{j=1}^{n}\beta_j^2$$

5.5.2 Optimization

Elastic Net is typically solved using gradient descent or coordinate descent.

5.5.3 Merits and Demerits

Merits:

- Combines strengths of L1 and L2 regularization.
- Handles multicollinearity well.

Demerits:

- Requires tuning of two hyperparameters.
- Computationally expensive for large datasets.