RIDGE REGRESSION – LASSO REGRESSION – ELASTIC NET

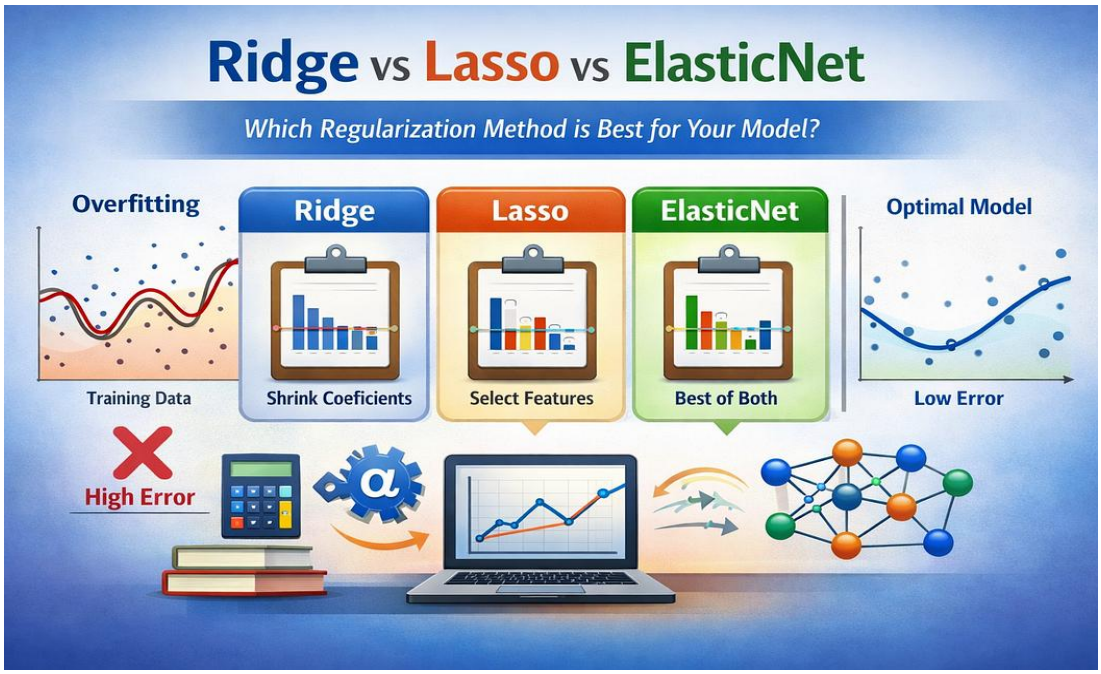
- Lasso, Ridge, and Elastic Net are regularization techniques used in linear regression to prevent overfitting and handle multicollinearity (high correlation among features) by adding a penalty to the size of the coefficients.

Core Differences

- Ridge Regression (L2 Penalty): Adds a penalty proportional to the square of the magnitude of the coefficients. It shrinks coefficients close to zero but never sets them exactly to zero, meaning it retains all features in the model.
- Lasso Regression (L1 Penalty): Adds a penalty proportional to the absolute value of the coefficients. It can drive coefficients down to absolute zero, performing automatic feature selection.
- Elastic Net (L1 + L2 Penalty): Combines both L1 and L2 penalties. It provides the feature selection of Lasso with the grouping and correlation handling of Ridge.

Introduction of Regularization

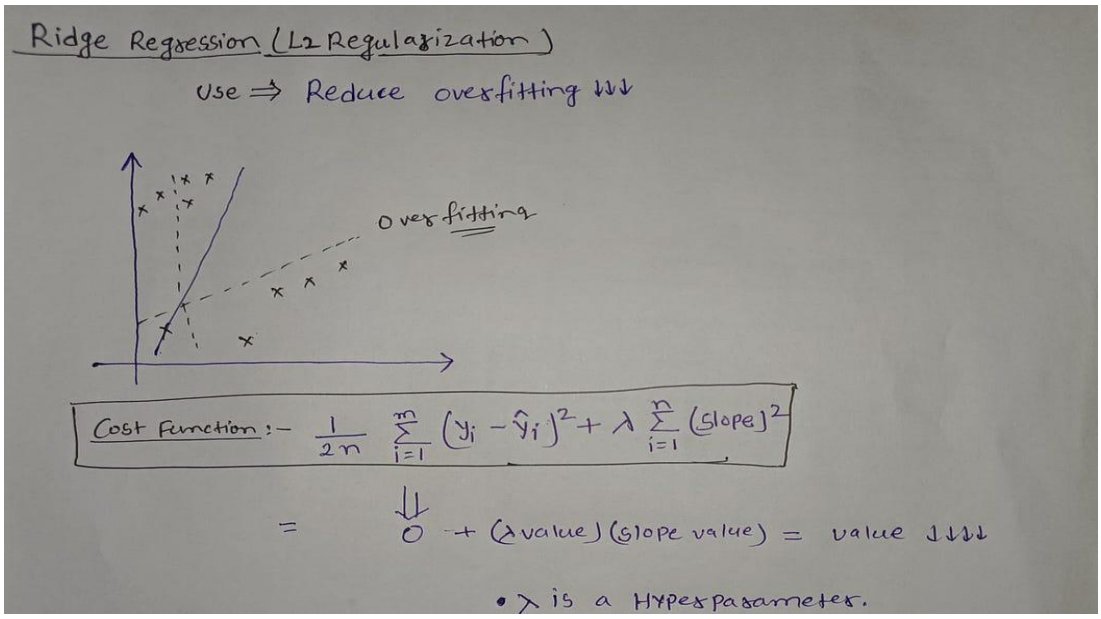
- In the area of linear models, ordinary least squares (OLS) regression is one of the chief ways to estimate the relationship between variables.
- But OLS also has serious limitations: it tends to overfit, especially with multicollinearity and when the number of predictors is greater than observations.
- Regularization methods such as Ridge or Lasso regression, however, have been developed to tackle such problems.



Ridge Regression (L2 Regularization)

- Ridge regression is a technique used to address overfitting by adding a penalty to the model's complexity.
- It introduces an L2 penalty (also called L2 regularization) which is the sum of the squares of the model's coefficients.
- This penalty term reduces the size of large coefficients but keeps all features in the model. This prevents overfitting with correlated features.

Ridge Regression Mathematical Intuition



S

- Formula for Ridge Regression:

- where:
- The first term calculates the prediction error.
- The second term penalizes large coefficients controlled by λ.

$$RidgeLoss = \sum_{i=1}^{m}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{n} \beta_j^2$$

- Example: Let's assume we are predicting house prices with features like size, location and number of rooms. The model might give coefficients like:
- β1= 5 (Size coefficient)
- β2= 3 (Number of rooms coefficient)
- λ= 0.1 (regularization strength).

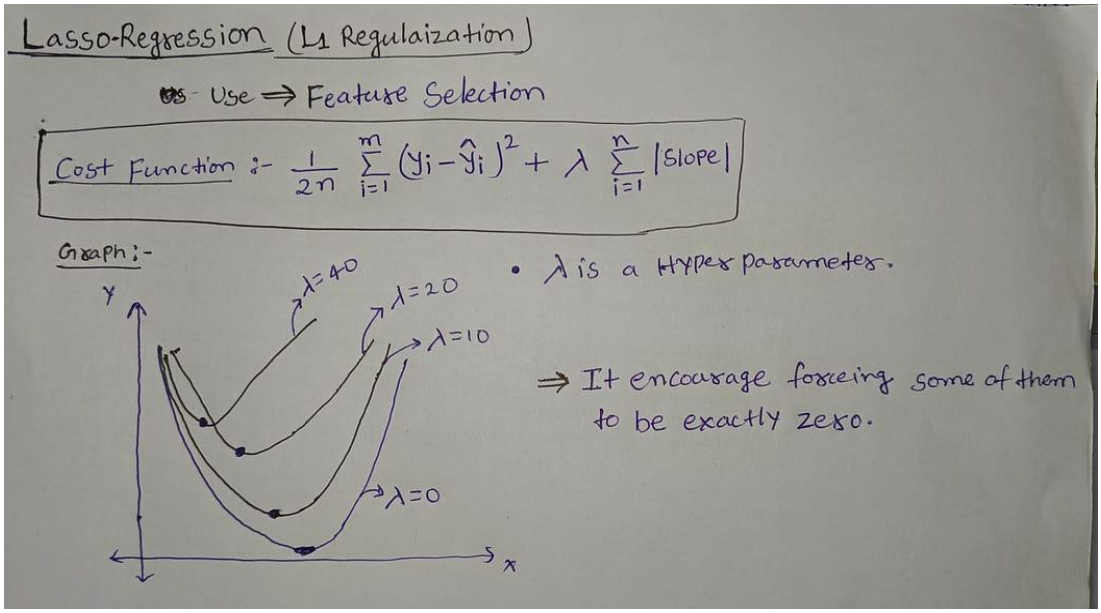
- The penalty term for Ridge would be calculated as:
- This penalty shrinks the coefficients to reduce overfitting but does not remove any features.

$$\lambda\left(\beta_1^2+\beta_2^2\right)=0.1\cdot\left(5^2+3^2\right)=0.1\cdot(25+9)=0.1\cdot 34=3.4$$

Lasso Regression (L1 Regularization)

- Lasso regression addresses overfitting by adding an L1 penalty i.e sum of absolute coefficients to the model's loss function.
- This encourages some coefficients to become exactly zero helps in effectively removing less important features. It also helps to simplify the model by selecting only the key features.

Lasso Regression Mathematical Intuition



- Formula for Lasso Regression:

- where:
- The first term calculates the prediction error.
- The second term encourages sparsity by shrinking some coefficients to zero.

$$LassoLoss = \sum_{i=1}^{m}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{n} |\beta_j|$$

- Example: Let's assume the same house price prediction example but now using Lasso. Assume:
- β1= 5 (Size coefficient)
- β2= 0 (Number of rooms coefficient is irrelevant and should be removed)
- λ= 0.1 (regularization strength).
- The penalty term for Lasso would be:
- λ· ∣β1∣=0.1· ∣5∣=0.1· 5=0.5
- Here Lasso forces β2= 0 removing the Number of Rooms feature entirely from the model.

Elastic Net Regression (L1 + L2 Regularization)

- Elastic Net regression combines both L1 (Lasso) and L2 (Ridge) penalties to perform feature selection, manage multicollinearity and balancing coefficient shrinkage.
- This works well when there are many correlated features helps in avoiding the problem where Lasso might randomly pick one and ignore others.

Formula for Elastic Net Regression:

- where:
- The first term calculates the prediction error.
- The second term applies the L1 penalty for feature selection.
- The third term applies the L2 penalty to handle multicollinearity.
- It provides a more stable and generalizable model compared to using Lasso or Ridge alone.

$$ElasticNetLoss = \sum_{i=1}^{m}(y_i - \hat{y}_i)^2 + \lambda_1 \sum_{j=1}^{n} |\beta_j| + \lambda_2 \sum_{j=1}^{n} \beta_j^2$$

- Example: Let's assume we are predicting house prices using Size and Number of Rooms. Assume:
- β1= 5 (Size coefficient)
- β2= 3 (Number of rooms coefficient)

- λ1= 0.1 (L1 regularization).
- λ2= 0.1 (L2 regularization).
- The penalty term for Elastic Net would be:

$$\lambda_1 \cdot (|\beta_1| + |\beta_2|) + \lambda_2 \cdot (\beta_1^2 + \beta_2^2) = 0.1 \cdot (|5| + |3|) + 0.1 \cdot (5^2 + 3^2) =$$
$$0.1 \cdot (5 + 3) + 0.1 \cdot (25 + 9) = 0.1 \cdot 8 + 0.1 \cdot 34 = 0.8 + 3.4 = 4.2$$

When to Use Which in Machine Learning

- Use Ridge Regression when: You believe that most or all of the features are useful and contribute to predicting the target, and you want to reduce model variance and manage multicollinearity.
- Use Lasso Regression when: You have a high-dimensional dataset where many features are irrelevant or redundant, and you need a sparse, interpretable model that drops noise variables entirely.
- Use Elastic Net when: You have highly correlated predictor variables or the number of features is much larger than the number of observations, causing Lasso to behave erratically or drop correlated groups arbitrarily

Problem on Ridge, Lasso, and Elastic Net Regression

- A regression model has the following coefficients obtained from ordinary linear regression:
- Apply the following regularization methods (assume the coefficients after regularization are given):
- Ridge Regression: Shrinks each coefficient by 20%.
- Lasso Regression: Sets coefficients less than or equal to 3 to zero and shrinks the others by 10%.
- Elastic Net Regression: Shrinks coefficients by 15%, and sets coefficients less than 2 to zero after shrinkage.
- Find the new coefficients for each method.

Feature	Coefficient
X_1	8
X_2	6
X_3	4
X_4	2

1. Ridge Regression

- Each coefficient is reduced by 20%.

- New coefficient=Old coefficient×(1−0.20)

Answer (Ridge):

- [6.4,4.8,3.2,1.6]

Feature	Original	Ridge
X_1	8	(8 * 0.8 = 6.4)
X_2	6	(6 * 0.8 = 4.8)
X_3	4	(4 * 0.8 = 3.2)
X_4	2	(2 *0.8 = 1.6)

2. Lasso Regression

- First, shrink coefficients by 10%.

Feature	After 10% Shrinkage
X_1	(8 * 0.9 = 7.2)
X_2	(6 *0.9 = 5.4)
X_3	(4 *0.9 = 3.6)
X_4	(2 *0.9 = 1.8)

- Now set coefficients ≤ 3 to zero.

Answer (Lasso):

- [7.2,5.4,3.6,0]
-

Feature	Lasso
X_1	7.2
X_2	5.4
X_3	3.6
X_4	0

3. Elastic Net Regression

- Shrink coefficients by 15%.

Feature	After 15% Shrinkage
X_1	(8 *0.85 = 6.8)
X_2	(6 *0.85 = 5.1)
X_3	(4 *0.85 = 3.4)
X_4	(2 *0.85 = 1.7)

- Now set coefficients < 2 to zero.

- Answer (Elastic Net):
- [6.8,5.1,3.4,0]
-

Feature	Elastic Net
X_1	6.8
X_2	5.1
X_3	3.4
X_4	0

- **Final Comparison**

- Key Observation
- Ridge: Shrinks all coefficients but none become zero.
- Lasso: Shrinks coefficients and eliminates less important features by setting them to zero.
- Elastic Net: Combines Ridge and Lasso, shrinking coefficients while also removing some features.
-

Feature	Original	Ridge	Lasso	Elastic Net
X_1	8	6.4	7.2	6.8
X_2	6	4.8	5.4	5.1
X_3	4	3.2	3.6	3.4
X_4	2	1.6	0	0