Machine Learning and Data Mining

Regression - Lasso, Ridge, Elastic Net

Claudio Sartori

DISI
Department of Computer Science and Engineering – University of Bologna, Italy
claudio.sartori@unibo.it

Lasso Regression

- Lasso (Least Absolute Shrinkage and Selection Operator) Regression:
  - A linear regression method that adds L_1-regularization to the cost function
  - Encourages sparse models by shrinking some coefficients to exactly zero
- Useful for feature selection and regularization in high-dimensional data

Objective Function

The Lasso Regression objective function minimizes:

$$\text{Loss} = \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

- Components
  - Residual sum of squares: Measures prediction error
  - L_1-norm regularization: Penalizes the sum of absolute values of coefficients
- Effects
  - Encourages sparsity in β_j
  - Controls the trade-off between error minimization and sparsity with λ

Computational Complexity

- Training Complexity
  - Depends on the number of features D, samples N, and iterations T
  - for coordinate descent:

$$O(TND)$$

  - for large datasets, this is linear in $N$and D
- Convergence
  - Faster convergence if many coefficients are sparse
  - Slower for high-dimensional dense datasets
- Prediction Complexity
  - Linear in $\bar{D}$ (number of nonzero coefficients):

Understanding β in Lasso Regression

- β represents the coefficients** or weights** of the linear regression model
- Structure of β:
  - $\beta = [\beta_0, \beta_1, \ldots, \beta_p]$
    - β_0: The intercept term of the model
    - β_j: The weight for the j-th feature, where $j = 1, \ldots, p$
- Predicted value $\hat{y}_i$ for a sample x_i:

$$\hat{y}_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

  - $\hat{y}_i$: Predicted output for the i-th sample
  - x_{ij}: Value of the j-th feature for the i-th sample

Role of β in Lasso Regression

- The optimization process adjusts β to:
  - Minimize residual error: $\frac{1}{2n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
  - Penalize large values of β_j using L_1-norm regularization: $\lambda \sum_{j=1}^{p} |\beta_j|$
- Effect of L_1-regularization:
  - Encourages sparsity in β
    - Many coefficients β_j are set to exactly zero
- β embodies the importance of each feature in the regression model, while ensuring simplicity and robustness

Lasso: Summary

- Advantages
  - Produces sparse models for feature selection
  - Scales linearly with the size of the dataset
- Limitations
  - Struggles with collinearity among features
  - Computationally expensive for very large p due to iterative updates
- Applications
  - High-dimensional datasets where feature selection is essential

Ridge Regression

- Ridge Regression is a type of linear regression
  - It adds a penalty term to the cost function to prevent overfitting
- Key Features:
  - Reduces model complexity
  - Improves generalization performance

The Ridge Regression Cost Function

- Ordinary Least Squares (OLS) cost function:

$$J(\mathbf{w}) = \sum_{i=1}^{n} (y_i - \mathbf{w}^T \mathbf{x}_i)^2$$

- Ridge Regression modifies OLS by adding a penalty:

$$J(\mathbf{w}) = \sum_{i=1}^{n} (y_i - \mathbf{w}^T \mathbf{x}_i)^2 + \lambda \|\mathbf{w}\|_2^2$$

  - λ: Regularization parameter controlling penalty strength
  - $\|\mathbf{w}\|_2^2$: L2 norm of the weight vector

Effects of Regularization

- High λ:
  - More penalty, leading to smaller weights
  - Reduces variance but increases bias
- Low λ:
  - Less penalty, resembling OLS regression
  - Retains variance but may overfit the data
- Choosing λ:
  - Cross-validation is commonly used to find the optimal λ

Ridge - Applications and Summary

- Applications:
  - Multicollinear data where features are highly correlated
  - Scenarios requiring reduced overfitting
- Summary:
  - Ridge Regression introduces a regularization term
  - Balances bias and variance for better generalization
  - Cross-validation helps in optimal parameter selection

Elastic Net Regression

- Elastic Net Regression is a linear regression method
  - Combines penalties from Ridge Regression and Lasso Regression
- Why Elastic Net?
  - Addresses limitations of Ridge and Lasso:
    - Ridge cannot perform feature selection
    - Lasso struggles when features are highly correlated
  - Offers a balance between these methods

The Elastic Net Cost Function

- Ordinary Least Squares (OLS) cost function:

$$J(\mathbf{w}) = \sum_{i=1}^{n} (y_i - \mathbf{w}^T \mathbf{x}_i)^2$$

- Elastic Net modifies OLS with two penalties:

$$J(\mathbf{w}) = \sum_{i=1}^{n} (y_i - \mathbf{w}^T \mathbf{x}_i)^2 + \lambda_1 \|\mathbf{w}\|_1 + \lambda_2 \|\mathbf{w}\|_2^2$$

  - λ_1: Controls the Lasso penalty (L1 norm)
  - λ_2: Controls the Ridge penalty (L2 norm)

Properties and Advantages

- Properties:
  - Encourages sparsity in coefficients (like Lasso)
  - Groups correlated features (like Ridge)
- Advantages:
  - Handles multicollinear data effectively
  - Can select relevant features while maintaining stability
  - Useful in high-dimensional data scenarios

Applications and Summary

- Applications:
  - Genomics (e.g., selecting gene expressions)
  - Financial modeling with highly correlated features
  - High-dimensional datasets with potential multicollinearity
- Summary:
  - Elastic Net combines Lasso and Ridge penalties
  - Effective in handling multicollinear data and sparse solutions
  - Requires hyperparameter tuning (λ_1, λ_2)

Comparison of regularized regression techniques

- Lasso, Ridge, and Elastic Net are regularization techniques used in regression
- They address overfitting and multicollinearity by introducing penalties in the cost function
- This presentation compares their real-world use cases, strengths, and limitations

Lasso Regression

- **Strengths:**
  - Performs feature selection, producing sparse models by setting some coefficients to zero
  - Useful for high-dimensional datasets with many irrelevant features
- **Limitations:**
  - Struggles with datasets where predictors are highly correlated
- **Use Cases:**
  - **Genomics:** Identifying relevant genes influencing a disease
  - **Text Processing:** Selecting keywords or n-grams in sentiment analysis
  - **Sparse Sensor Networks:** Identifying critical sensors in IoT or environmental monitoring

Ridge Regression

- **Strengths:**
  - Handles multicollinearity by shrinking coefficients
  - Retains all predictors, avoiding the elimination of variables
- **Limitations:**
  - Does not perform feature selection
- **Use Cases:**
  - **Finance:** Predicting stock prices using correlated economic indicators
  - **Marketing:** Modeling customer demand influenced by correlated factors
  - **Engineering:** Calibration of multivariate systems like chemical processes
  - **Medical Imaging:** Predicting outcomes from high-dimensional MRI or CT data

Elastic Net Regression

- **Strengths:**
  - Combines Lasso and Ridge penalties, balancing sparsity and multicollinearity handling
  - Selects groups of correlated features, unlike Lasso alone
- **Limitations:**
  - Requires careful tuning of two parameters (λ_1 and λ_2)
- **Use Cases:**
  - **Genomics:** Selecting groups of genes associated with traits
  - **Healthcare Analytics:** Modeling patient outcomes from clinical predictors
  - **Customer Segmentation:** Identifying clusters of customer behaviors in retail
  - **Climate Science:** Modeling climate variables with correlated predictors
  - **Social Media Analysis:** Predicting trends from sparse and correlated features

Comments

- Lasso, Ridge, and Elastic Net offer distinct strengths tailored to different data characteristics
- Choosing the right method depends on:
  - Presence of multicollinearity
  - Sparsity of the solution required
  - Dimensionality of the dataset
- Elastic Net is often a robust choice when both sparsity and correlation must be addressed

Comparison of Lasso, Ridge, and Elastic Net Regression

Feature	Lasso	Ridge	Elastic Net
Feature Selection	Yes	No	Yes groups correlated features
Handles Multicollinearity	Weak	Strong	Strong
Model Interpretability	High (sparse coefficients)	Moderate	Moderate (sparse, but groups features)
Dataset Characteristics	High-dimensional, sparse predictors	Correlated predictors	Sparse and correlated predictors

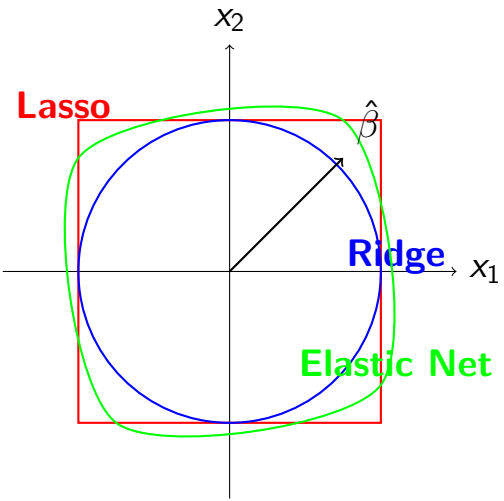
Explanation

- **Sparsity:**
  - Elastic Net can set some coefficients to zero, removing irrelevant predictors
  - This results in a simpler and more interpretable model, similar to Lasso
- **Groups Features:**
  - When predictors are highly correlated, Elastic Net:
    - Tends to select them together rather than choosing one arbitrarily
    - Shrinks their coefficients toward each other using the Ridge-like penalty
  - This behavior arises because Elastic Net combines:
    - L1 penalty (Lasso) for sparsity
    - L2 penalty (Ridge) for handling multicollinearity

Practical Example: Correlated Predictors

- Suppose two predictors, x_1 and x_2, are highly correlated:
  - **Lasso:**
    - May select only x_1 or x_2, ignoring the other entirely
  - **Ridge:**
    - Keeps both x_1 and x_2, but shrinks their coefficients
  - **Elastic Net:**
    - Selects both x_1 and x_2, but their coefficients may be reduced (shrunk) in different proportions
    - Balances between sparsity and correlation handling

Visualization of Sparse and Grouping Behavior



Description of the figure I

- Visual representation of the constraints applied by Lasso, Ridge, and Elastic Net regression
- The axes represent the coefficient of two predictors
- Shapes of Constraints
  - Lasso: A diamond-shaped constraint indicating L1 penalty, which promotes sparsity (coefficients set to zero)
  - Ridge: A circular constraint indicating L2 penalty, which shrinks coefficients uniformly but does not set them to zero
  - Elastic Net: A combination of Lasso and Ridge constraints, allowing both sparsity and handling of correlated groups

Description of the figure II

- Interpretation of Coefficient Paths
  - In Lasso, coefficients are pushed to the edges, setting some to zero
  - In Ridge, coefficients shrink but remain non-zero, resulting in a smoother path
  - Elastic Net provides a balance, with paths that follow the L1 and L2 constraints, enabling feature selection and correlation handling

Key Takeaways

- Elastic Net combines the best of Lasso and Ridge:
  - **Sparsity:** Sets some coefficients to zero for simpler models
  - **Handles Correlated Predictors:** Selects groups of features rather than one arbitrarily
- Ideal for:
  - High-dimensional datasets with multicollinearity
  - Applications requiring both feature selection and robust performance