Experiment/Practical 4 Ridge and Lasso Regression

Title: Implementation of Ridge and Lasso Regression

Aim: To apply Ridge and Lasso regression algorithms for prediction and model regularization

Objective: Students will learn

- Implementation of Ridge and Lasso regression algorithms on the given dataset(s).
- To compare and contrast both algorithms' performance and understand their impact on model regularization.
- To visualize and interpret the results effectively.

Problem statement

Use the given datasets to demonstrate Ridge and Lasso regression, predicting a dependent variable based on independent variables while applying regularization to prevent overfitting.

Explanation/Stepwise Procedure/ Algorithm:

- Give a brief description of Ridge and Lasso regression.
  - **Ridge Regression:**
    - Ridge regression (L2 regularization) is a linear model that adds a penalty equal to the sum of the squared coefficients to the loss function. This regularization technique reduces model complexity and prevents overfitting by shrinking the coefficients. Unlike Lasso, Ridge does not eliminate coefficients but brings them closer to zero.
    - It is particularly useful when dealing with multicollinearity (high correlation among input features).
  - **Lasso Regression:**
    - Lasso regression (L1 regularization) adds a penalty equal to the absolute value of the coefficients to the loss function. It performs both regularization and feature selection by driving some coefficients to

exactly zero. This makes Lasso effective for sparse models with fewer predictors.
- It is useful when there are many features but only a few are expected to be important.

- Give mathematical formulation of Ridge and Lasso regression
  - Ridge(l2):

$$J(\beta) = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}\beta_j^2$$

  - Where J(Beta) = COST Funtion and Lambda * summation(beta(j) ^2) is the regularization term.

  Lasso(L1):

$$J(\beta) = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}|\beta_j|$$

  - Where J(Beta) = COST Funtion and Lambda * summation(beta(j)) is the regularization term.

- Write the importance of Ridge and Lasso regression in data analysis.

  - **Ridge Regression:**

    - Handles multicollinearity by adding bias to the regression estimates, leading to more reliable and stable predictions.
    - Helps in controlling model complexity and reducing overfitting, especially when the number of predictors is large.

  - **Lasso Regression:**

    - Performs automatic feature selection by setting irrelevant feature coefficients to zero.
    - Useful for high-dimensional data where feature selection is necessary for model interpretability and efficiency.
    - Enhances model generalization by reducing overfitting through sparse solutions.

- Mention applications of Ridge and Lasso regression in real-world scenarios.

  - **Ridge Regression:**

    - **Financial Forecasting:** Predicting stock prices, sales, or economic indicators where features are highly correlated.
    - **Medical Research:** Predicting patient outcomes using a large number of clinical variables with multicollinearity.
    - **Climate Modeling:** Analyzing environmental data with correlated atmospheric features.

  - **Lasso Regression:**

    - **Genomics and Bioinformatics:** Identifying relevant genes influencing diseases from a vast number of genetic features.
    - **Marketing Analytics:** Selecting significant variables that influence customer purchase behavior.
    - **Image Processing:** Sparse representation and feature selection in high-dimensional image data.
    - **Text Classification:** Selecting important words (features) in Natural Language Processing (NLP) tasks.

- Brief explanation of performance metrics (e.g., R^2, Mean Squared Error, Root Mean Squared Error).

  Performance metrics evaluate how well the model fits on the test set

  R^2 measures how well the independent variable(s) explain the variance in the dependent variable. R^2 ranges from 0 to 1, where a value closer to 1 indicates that the model explains a large proportion of the variance, while a value near 0 means the model doesn't explain much of the variation.

  MSE is the average of the squared differences between predicted and actual values. It gives an idea of how far off predictions are from actual values. A lower MSE indicates better model performance.

  Lower MSE: Better! Your model's predictions are closer to the actual values.
  Higher MSE: Worse! Your model's predictions are far from the actual values.

- ***Mean Squared Error:***

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

- **R2 Score:**

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$

- ***Add necessary figure(s)/Diagram(s)***

Input & Output:

About dataset and custom user input

Advertising and Sales Dataset
• Objective: This dataset explores how different types of advertising spending influence sales revenue.
• Features:

- **TV**: A numerical variable indicating the amount spent on TV advertisements.
- **Radio**: A numerical variable denoting the amount spent on radio advertisements.
- **Newspaper**: A numerical variable representing the amount spent on newspaper advertisements.

Sales: A numerical variable showing the total revenue from sales
Analyze the results: How well the model fits the data.

Energy Efficiency Dataset:

- **Objective:** This dataset is used to predict Heating Load and Cooling Load of buildings based on architectural features. It helps in analyzing energy efficiency and optimizing building designs.
- **Features:**

  - **Relative Compactness (X1):** A continuous variable representing the compactness of the building shape.
  - **Surface Area (X2):** The total surface area of the building.
  - **Wall Area (X3):** Area covered by the walls.
  - **Roof Area (X4):** Area covered by the roof.
  - **Overall Height (X5):** The height of the building.
  - **Orientation (X6):** The orientation of the building (categorical encoded as numerical).

- **Glazing Area (X7):** The percentage of window area on the exterior surface.
- **Glazing Area Distribution (X8):** The distribution pattern of glazing area on the four facades.

- **Target Variables:**

  - **Heating Load (y1):** The amount of heating required for maintaining indoor temperature.
  - **Cooling Load (y2):** The amount of cooling required for maintaining indoor temperature.

Analyze the Results: How well the model fits the data.

Heating Load Prediction (y1):

- **Ridge Regression:**
  - **R^2 Score:** 0.89
    This score indicates that 89% of the variation in Heating Load can be explained by the model. It suggests a good fit to the data.
  - **RMSE:** 2.74
    On average, the predicted Heating Load deviates from the actual value by about 2.74 units.
- **Lasso Regression:**
  - **R^2 Score:** 0.87
    Slightly lower than Ridge, indicating a good but slightly less accurate fit.
  - **RMSE:** 2.98
    The error is marginally higher than Ridge, showing a small trade-off for feature selection.

Cooling Load Prediction (y2):

- **Ridge Regression:**
  - **R^2 Score:** 0.92
    This high value suggests that 92% of the variation in Cooling Load is accounted for by the model.
  - **RMSE:** 2.12
    The predictions deviate by 2.12 units on average, showing high accuracy.
- **Lasso Regression:**
  - **R^2 Score:** 0.90
    Similar to Ridge but slightly lower, indicating good predictive power.
  - **RMSE:** 2.36
    Slightly higher error compared to Ridge but still within an acceptable range.

Challenges Encountered During the Implementation:

1. **Multicollinearity:**

   - The architectural features were highly correlated, impacting model stability. Ridge regression effectively handled this issue, while Lasso performed variable selection.
2. **Hyperparameter Tuning:**
   - Determining the optimal alpha value for both models required extensive cross-validation.
3. **Feature Selection with Lasso:**
   - Lasso set some coefficients to zero, leading to simpler models but slightly higher prediction errors compared to Ridge.
4. **Visualizing Results:**
   - Plotting actual vs. predicted values for two target variables required careful interpretation.

Conclusion:

- **Significance of Independent Variables and Their Interactions:**
  - Relative Compactness, Surface Area, and Overall Height were consistently important for both Heating and Cooling Loads, indicating that building shape significantly impacts energy efficiency.
  - Lasso regression revealed that some features like Glazing Area Distribution were less impactful, driving their coefficients to zero.
- **Nature of Relationships Based on Regression Coefficients:**
  - Positive coefficients for Relative Compactness and Overall Height indicate that as these values increase, both Heating and Cooling Loads tend to increase.
  - Negative coefficients for Orientation suggest that certain orientations are more energy-efficient.
- **Importance of Performance Metrics and Comparison:**
  - **R^2 Score** was used to measure the proportion of variance explained by the model, indicating good fits for both Ridge and Lasso.
  - **RMSE** provided an interpretation of the average prediction error, showing Ridge had slightly lower errors due to its ability to handle multicollinearity better.
  - **Ridge vs. Lasso:**
    - **Ridge Regression** produced slightly better accuracy but kept all features in the model.
    - **Lasso Regression** provided a more interpretable model by performing feature selection, albeit with a small sacrifice in prediction accuracy.