COMP 6915 - Machine Learning
Assignment 2 Report

by
© Soroush Baghernezhad
& Ramyar Zarza
& Mantra .

Instructor: Karteek Popuri
Department of Computer Science & Scientific Computing
Memorial University of Newfoundland

January 2025

St. John's Newfoundland

Question 1:

In part 1 we traind regression model with and without crossvalidation and with different n_fold, the n_folds were chosen from this set $n_folds = \{5, 10, 20, 50, 100, len(Train_D)\}$

```
Simple Linear Regression - Validation Approach Test RSE: 11.65, R^2: 0.54
Simple Linear Regression - Cross-Validation for 5 folds   RSE: 10.37, R^2: 0.63
Simple Linear Regression - Cross-Validation for 10 folds   RSE: 10.4, R^2: 0.63
Simple Linear Regression - Cross-Validation for 20 folds   RSE: 10.35, R^2: 0.63
Simple Linear Regression - Cross-Validation for 50 folds   RSE: 10.26, R^2: 0.63
Simple Linear Regression - Cross-Validation for 100 folds   RSE: 10.07, R^2: 0.63
Simple Linear Regression - Cross-Validation for 800 folds   RSE: 8.22, R^2: 0.63
```

With increasing the number of folds, we can see that RSE reduced because we had more training data in each training, however R^2 remains the same.
Without using cross-validation, we can see that our RSE is higher than ones with cross-validation and that is because we are only testing with part of data.

Question 2:

Ridge regression is a regularized version of ordinary least squares (OLS) that addresses multicollinearity and overfitting by adding an L_2 penalty to the loss function. The ridge regression objective minimizes the sum of squared residuals while penalizing large coefficients:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \alpha \sum_{j=1}^{p} \beta_j^2$$

where X is the design matrix, y is the response vector, β represents the regression coefficients, and $\alpha \geq 0$ is the regularization parameter controlling the strength of the penalty. When $\alpha = 0$, ridge regression reduces to OLS, while larger values of α shrink the coefficients toward zero, reducing model complexity and preventing overfitting. Unlike Lasso regression, ridge does not force coefficients to be exactly zero, making it more suitable when all predictors contribute to the response.
In this question we searched for differernt values of α to find the least RSE. In Figure 1, we can see the RSE error for different values of α and the optimal parameter of it with red vertical line.
For the ridge model, we have $RSE : 10.00$ and $R^2 : 0.59$ in 5-fold cross-validation, whereas in the vanilla regression model, we had $RSE : 10.37$ and $R^2 : 0.63$. This indicates that the ridge model achieves a lower residual standard error (RSE), suggesting better predictive accuracy and reduced variance compared to the vanilla regression model. However, the slight decrease in R^2 from 0.63 to 0.59 suggests that the ridge penalty introduces some bias by shrinking coefficients, which slightly reduces the proportion of variance explained. This trade-off between bias and variance is expected in ridge regression, as it aims to improve generalization by preventing overfitting.

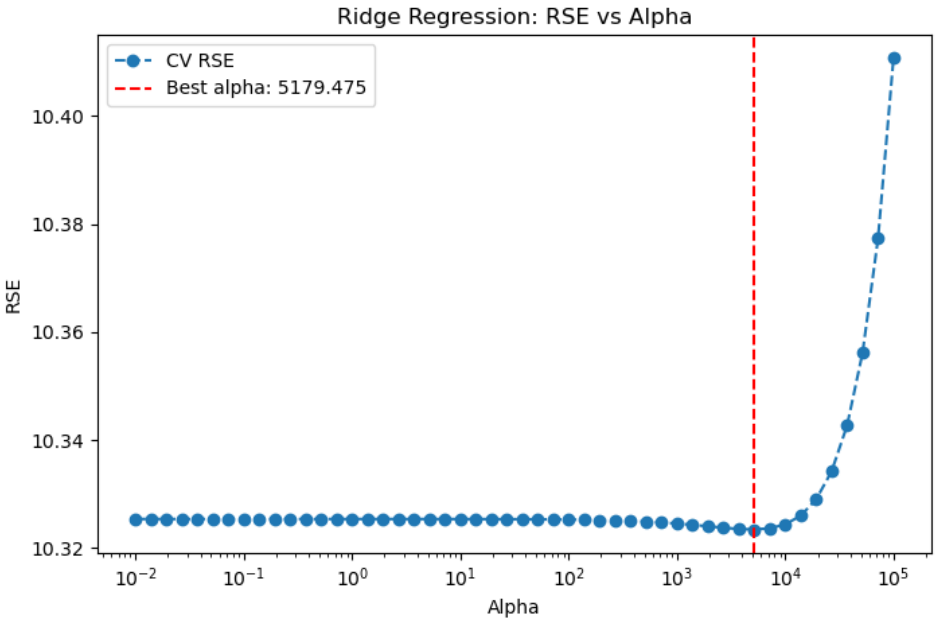
Ridge Regression: RSE vs Alpha
CV RSE
Best alpha: 5179.475
RSE
10.40
10.38
10.36
10.34
10.32
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
10^{4}
10^{5}
Alpha

Figure 1

Question 3:

Lasso (Least Absolute Shrinkage and Selection Operator) regression is a regularization technique that extends ordinary least squares (OLS) by adding an L_1 penalty to the loss function, promoting sparsity by shrinking some regression coefficients to exactly zero. The Lasso objective function is given by:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \alpha \sum_{j=1}^{p} |\beta_j|$$

where X is the design matrix, y is the response vector, β represents the regression coefficients, and $\alpha \geq 0$ is the regularization parameter that controls the degree of shrinkage. Unlike ridge regression, which applies an L_2 penalty and shrinks coefficients continuously, Lasso enforces sparsity by setting some coefficients exactly to zero, making it useful for feature selection. When $\alpha = 0$, Lasso reduces to OLS, whereas larger values of α increase sparsity, effectively selecting a subset of relevant predictors. For this question we compared RSE error for different values of α in the range of $\{10^{-2}\}$ to $\{10^{5}\}$ In figure 2 it can be seen that RSE error tend to increase with increasing the α value. We saw that lasso had $RSE : 10.00$ same as ridge in previous question and $R^2 : 0.59$ which is again same as ridge algorithm. However, it still has less residual sum of error compared to vanilla regression model.

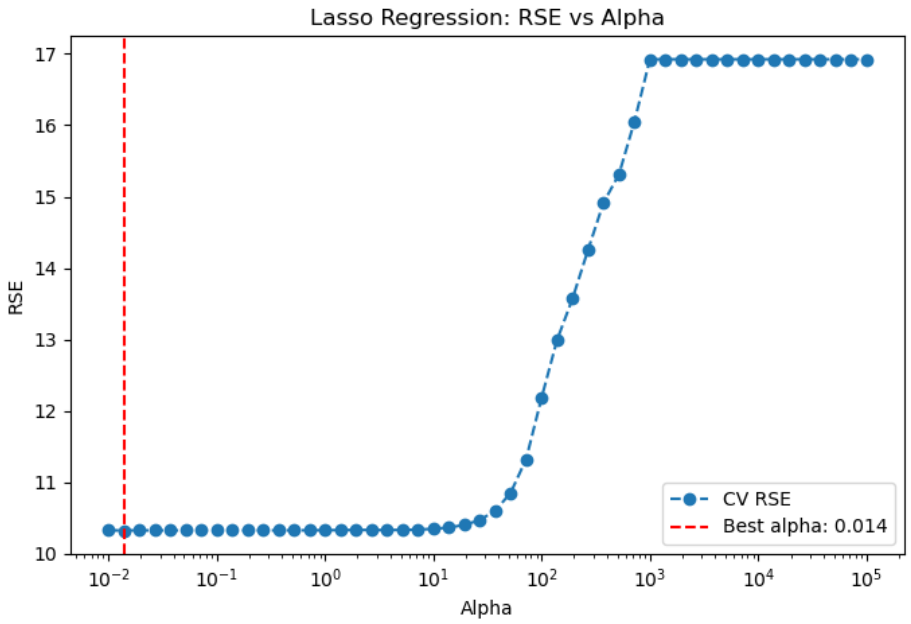

Figure 2

Question 4:

In Question 4, we utilized three methods: the Ridge and Lasso algorithms, which use L_2 and L_1 norms, respectively, and their combination, the Elastic Net algorithm. Elastic Net combines both penalties to balance regularization and feature selection, and its objective function is given by:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \alpha_1 \sum_{j=1}^{p} |\beta_j| + \alpha_2 \sum_{j=1}^{p} \beta_j^2$$

where X is the design matrix, y is the response vector, β represents the regression coefficients, and $\alpha_1, \alpha_2 \geq 0$ are the regularization parameters controlling the L_1 (Lasso) and L_2 (Ridge) penalties, respectively. This formulation allows Elastic Net to handle multicollinearity like Ridge regression while performing automatic feature selection like Lasso.

We tested our methods on 100 different α values over 20 folds. The RSE chart for each fold across different α values is depicted in Figure 3, where the vertical line represents the optimal α value, and the solid black line indicates the average RSE across all 20 folds. Both Lasso and Elastic Net exhibited similar performance.

Additionally, we incorporated Z-score normalization into our pipeline to ensure that all data followed a standard distribution with a mean of 0 and a variance of 1. We also tested the effect of adding PCA-based feature reduction, which reduced the number of features from 8 to 3. However, this did not result in a significant difference in our RSE metric.

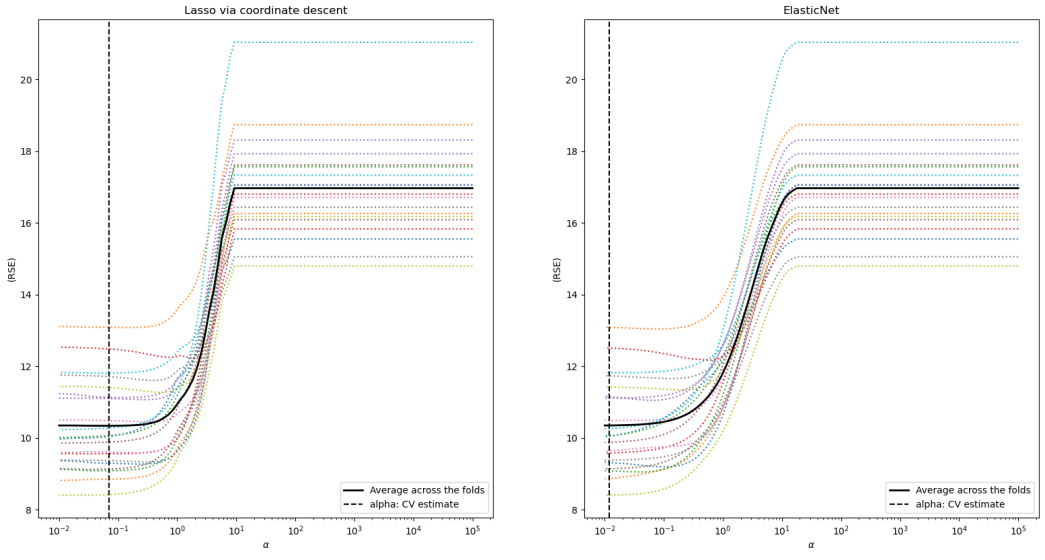

Figure 3