Assignment 1

Report

INF8245AE
Machine Learning

Name	Student ID
Ali Abbas	2078188

Polytechnique Montréal
21 September 2025

Table of Contents

1. Linear and Weighted Ridge Regression

1.1 Weighted Ridge Regression Solution

$$\mathcal{L}(w) = ||Xw - y||_2^2 + w^T \Lambda w$$

$$||Xw - y||_2^2 = (Xw - y)^T (Xw - y)$$

$$(Xw - y)^T (Xw - y) = (w^T X^T - y^T)(Xw - y)$$

$$(w^T X^T - y^T)(Xw - y) = w^T X^T Xw - w^T X^T y - y^T Xw + y^T y$$

$$w^T X^T y = (Xw)^T y \quad and \quad y^T Xw = (y^T X)w$$

$$y \in \mathbb{R}^n, X \in \mathbb{R}^{n \times d} \rightarrow w \in \mathbb{R}^d, \; y^T Xw \in \mathbb{R}^{1 \times 1} \; (Scalar)$$

$$(y^T Xw)^T = y^T Xw = w^T X^T y$$

$$\mathcal{L}(w) = w^T X^T Xw - y^T Xw - y^T Xw + y^T y + w^T \Lambda w$$

$$\mathcal{L}(w) = w^T (X^T X + \Lambda)w - 2y^T Xw + y^T y$$

Minimization:

$$\nabla_w \mathcal{L}(w) = 2(X^T X + \Lambda)w - 2y^T X \, , \; given \; that \; \frac{\partial (x^T Ax)}{\partial x} = 2Ax$$

$$\nabla_w \mathcal{L}(w) = 0$$

$$2(X^T X + \Lambda)w - 2y^T X = 0$$

$$(X^T X + \Lambda)w = X^T y,$$

$$tranposed \; y^T X \; to \; match \; dimension \; of \; \mathbb{R}^{d \times 1} \; on \; boths \; sided \; of \; the \; equation$$

$$w^* = (X^T X + \Lambda)^{-1} X^T y$$

1.2 Data Exploration and Results (q1_2)

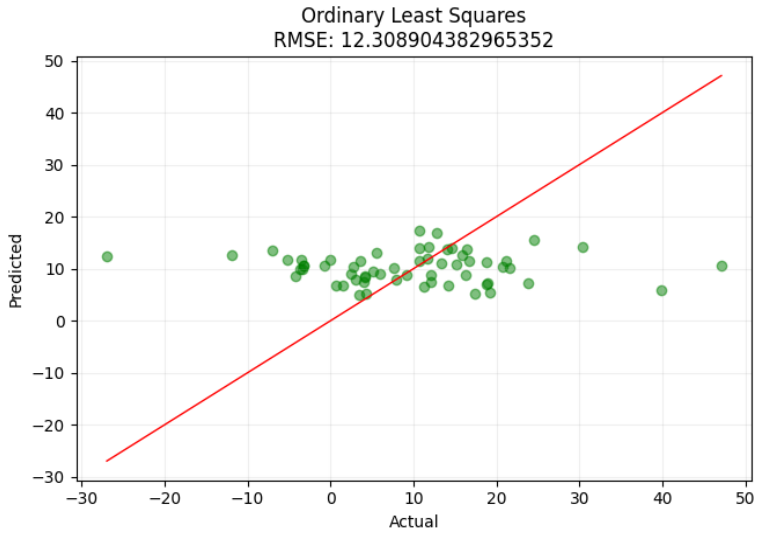

Figure 1 Actual vs Predicted Plot: OLS

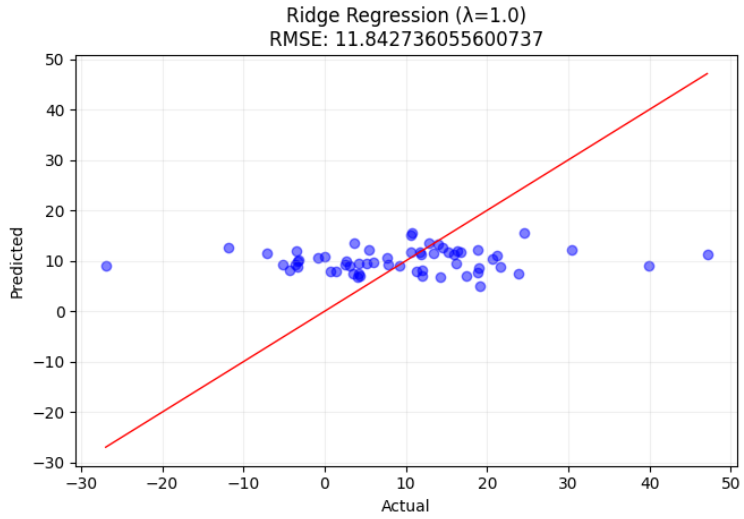

Figure 2 Actual vs Predicted Plot: Ridge Regression

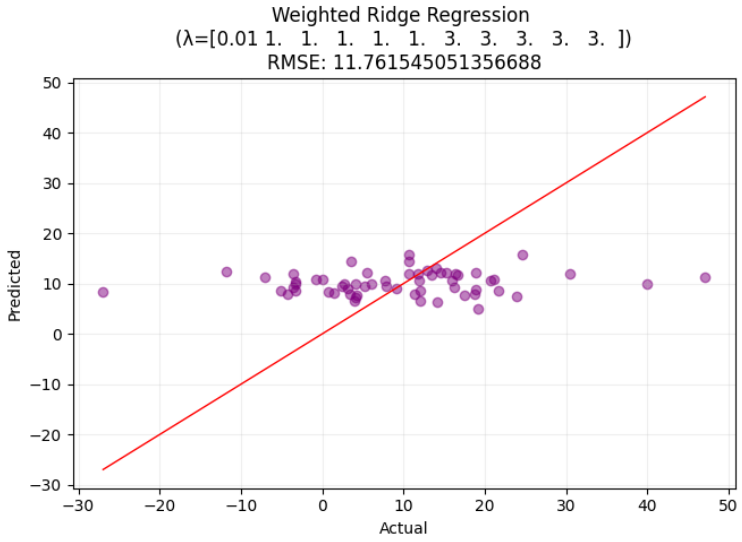

Figure 3 Actual vs Predicted Plot: Weighted Ridge Regression

The figures above indicate that the predicted data does not strongly match the test data, without further data exploration it is only possible to make assumptions as to why this is the case.

The horizontal scatter plots across all three models could be an indication of a nonlinear relationship in the data. For further analysis, Figure 4 illustrates the independent correlation of each feature to the y target (X_train vs y_train), the results show weak correlation between features and the target and a noisy scatter plot for each with no discernable pattern. This is a possible explanation as to why the linear models (OLS, Ridge Regression and Weighted Ridge Regression) predictions did not have a strong relationship to the actual data as can be seen in Figure 1, Figure 2 and Figure 3. The RMSEs for OLS, Ridge and Weighted ridge regression are 12.3089, 11.8427 and 11.7615 respectively. The models show signs of low variance and high bias (Underfitting), analyzing the RMSEs of each model show gains (lower RMSE) with regularization but not enough to avoid underfitting. Finally, Figure 5 shows a correlation heatmap between features, the results indicate high correlation (multicollinearity and redundancy of features is a strong possibility). It is important to note that the conclusions mentioned are not proof of the three models' predictive powers or the nature of relationship between the target and features (possible nonlinearity), but they simply are possible explanations for the results.

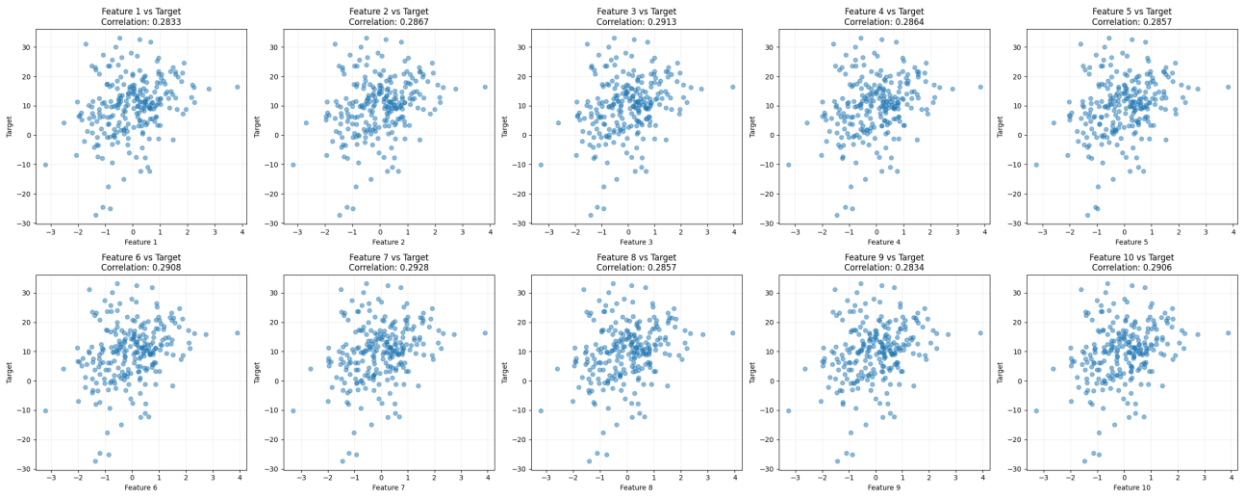

Figure 4 Correlation and scatter plots of features

Note: The analyses done for Figure 4 and Figure 5 are not explicitly required for this assignment but were included for better data exploration and a better understanding of the dataset. The methods used are drawn from materials in courses MTH3302 and INF8111.

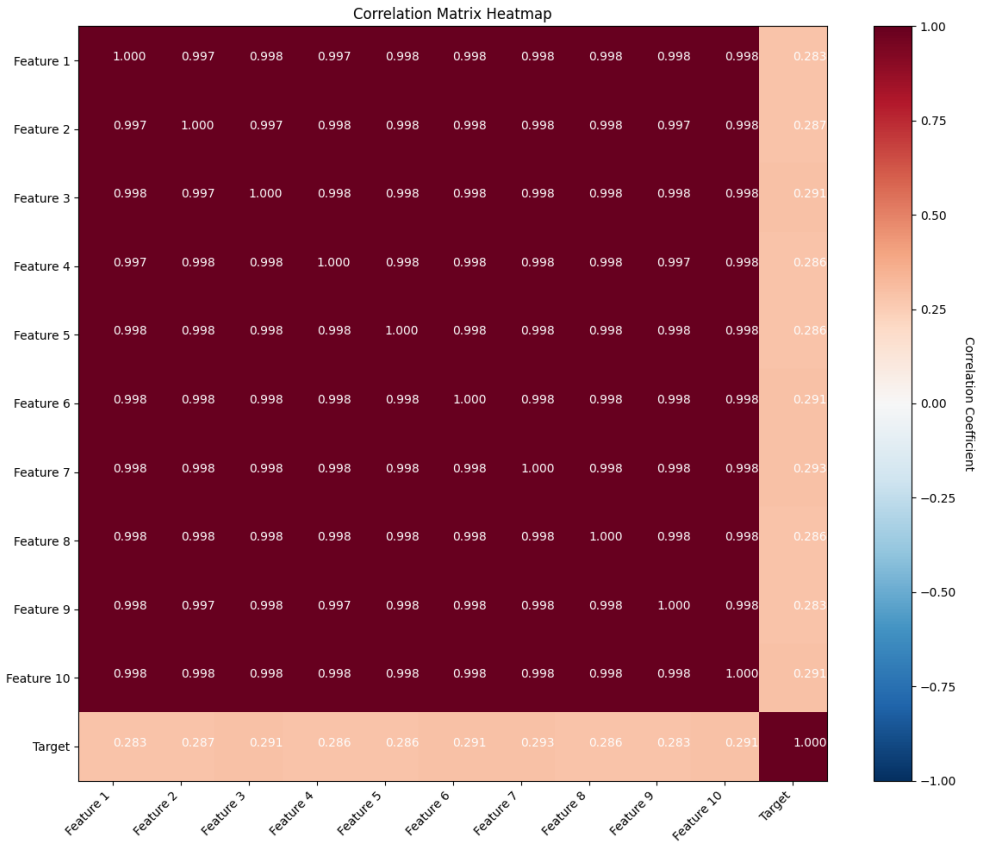

Figure 5 Correlation Heatmap

2 Cross-Validated Model Selection

2.1 Cross Validation Results (q2_2)

λ	MAE (best λ = 1)	Max Error (best λ= 100)	RMSE (best λ= 10)
0.01	8.091965	37.78716	9.742656
0.1	8.017052	37.2661	9.668058
1	7.873522	36.06789	9.521207
10	7.875033	35.29168	9.502451
100	8.470609	34.10074	10.04619

Table 1 Cross Validation Results

Table 1 shows the 5-fold cross-validation results for a ridge regression model. The results indicate that minimally regularized models have worse performance than more highly regularized models ($\lambda \geq 1$) except for MAE value at $\lambda = 100$. On average $\lambda = 1$ has the best MAE score, $\lambda = 100$ has the best Max Error score and finally $\lambda = 10$ has the best mean RMSE score. Of course, these results may vary slightly due to the random shuffle done in the k-fold algorithm.

3 Gradient Descent for Ridge Regression with Learning Rate Schedules

3.1 Results (q3_2)

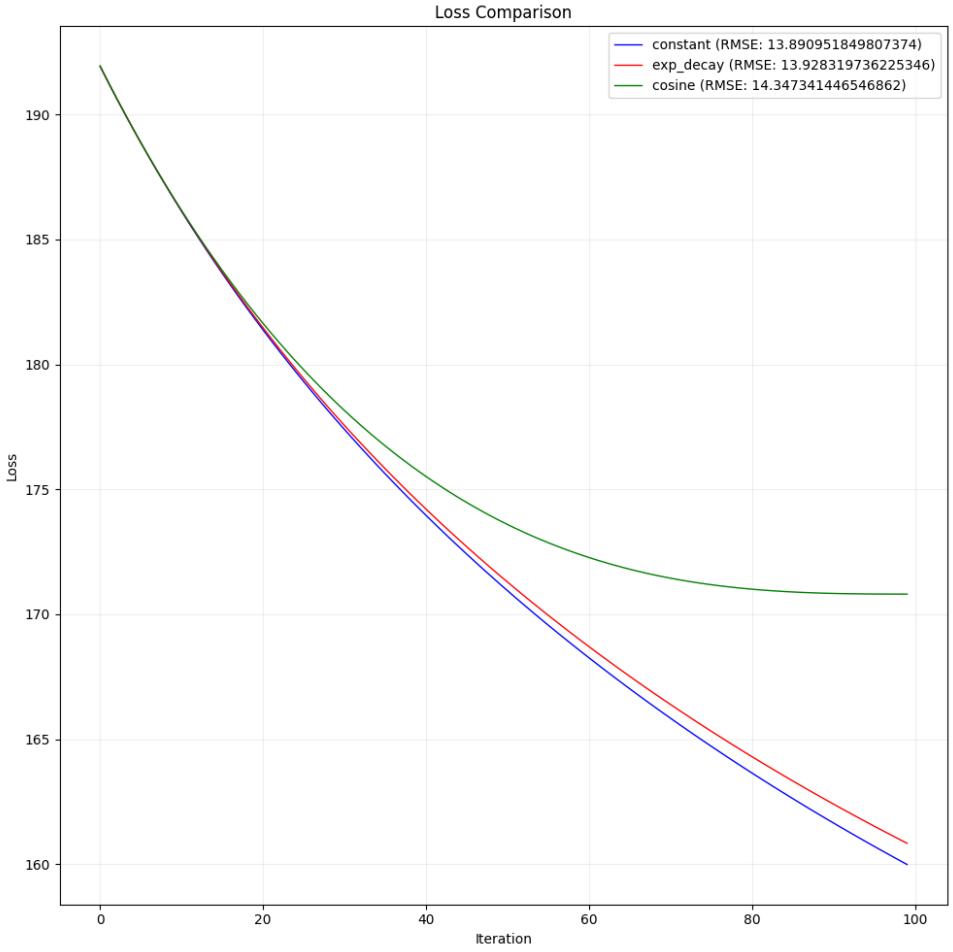

Figure 6 Loss vs Iteration

The constant schedule converges to the lowest loss faster than exponential decay or cosine schedules, the cosine reaches a plateau the fastest but is not the most optimal solution, the constant schedule is followed closely by the exponential decay schedule. In summary the constant schedule has the best performance at $\eta_0 = 0.001$, T = 100 and $\lambda = 1$.

Note: *Given that the gradient of Ridge regression loss is:*

$$\nabla_w \mathcal{L}(w) = \frac{-2}{n} X^T (y - Xw) + 2\lambda w$$

Then the loss function must be:

$$\mathcal{L}(w) = \frac{1}{n} ||y - Xw||^2 + \lambda ||w||^2$$