Impact of Regression and Regularization Techniques on Predictive Modeling of Diabetes Progression

Iqra Jawaid (25K-7620)

Muhammad Abdullah Panhwar (25K-7638)

Abstract

This report implements and compares five foundational regression techniques—Ordinary Least Squares (OLS), SVD-based Regression, Gradient Descent (GD), Principal Component Analysis (PCA) with OLS, and Ridge Regression—to predict the progression of diabetes disease one year post-baseline. The study utilized the Diabetes dataset (N=442, D=10) after standardizing features and appending a bias term. Results demonstrated that models incorporating complexity control outperformed the baseline OLS model. Specifically, PCA combined with OLS on 8 principal components achieved the lowest Mean Squared Error (MSE: 2848.3308) on the test set. Ridge Regression also provided strong regularization, achieving an MSE of 2866.4258. Computationally, SVD proved the fastest closed-form solution (3.52 ms), while GD provided a viable, scalable iterative approach.

1. Introduction

1.1 Dastaset Description and Preprocessing: The project employs the standard Diabetes dataset, which consists of N=442 patient samples and D=10 normalized physiological and serum features. The target variable (y) is a quantitative measure of disease progression one year after the initial examination. Prior to model training, the data was split (80% train, 20% test). Feature matrices were Standard Scaled and augmented with a column of ones to incorporate the model's bias/intercept term (θ_0).

1.2 Problem Formulation: The objective is to solve the linear regression problem ($X\theta \approx y$) where the goal is to find the optimal weight vector θ that minimizes the prediction error on the training data. The primary performance metric for comparing the methods is the Mean Squared Error (MSE) on the held-out test set, defined as MSE = $\frac{1}{N}\sum_{i=1}^{N}(y_i - X_i\theta)^2$.

2. Methodology

2.1 Closed-Form Solutions: OLS and SVD

The Ordinary Least Squares (OLS) method determines the optimal θ analytically by solving the Normal Equation:

$$\theta_{OLS} = (X^T X)^{-1} \boldsymbol{X^T y}$$

The SVD-based Regression approach is theoretically equivalent but computationally more stable. It utilizes the Singular Value Decomposition ($X = USV^T$) to compute the pseudo-inverse.

2.2 Iterative Optimization: Gradient Descent (GD)

Gradient Descent iteratively adjusts the weight vector θ by moving in the direction opposite to the gradient of the cost function, $J(\theta) = ||X\theta - y||^2$. The update rule is:

$$\theta_{t+1} = \theta_t - \eta\nabla J(\theta_t)$$

The model was trained for 1000 epochs, testing three learning rates ($\eta \in \{0.1, 0.01, 0.001\}$) to find the optimal convergence speed and stability.

2.3 Regularization: Ridge Regression

Ridge Regression is a form of L2 regularization that minimizes a penalized cost function, adding a squared magnitude of the coefficient vector to the OLS objective. This shrinks coefficients towards zero, which helps mitigate multicollinearity and prevents overfitting.

$$\theta_{Ridge} = (X^T X + \lambda I)^{-1} \mathbf{X^T y}$$

The model was evaluated using regularization parameters $\lambda \in \{0.1, 1, 10, 100\}$.

2.4 Dimensionality Reduction: PCA

Principal Component Analysis (PCA) was used to transform the feature space onto a lower-dimensional subspace that captures maximum variance. PCA was implemented using SVD on the centered feature matrix. OLS was then performed on the data projected onto ***k*** principal components, varying ***k*** from 1 to 10 to determine the optimal dimensionality.

3. Results

3.1 Comparative Performance of Regression Methods

Table 1 summarizes the best performing configuration for each core regression method based on the test set MSE.

Method	Best Parameter	Training Runtime	Test MSE
OLS (Normal Eq.)	N/A	31.45 ms	2900.1936
SVD-based Reg.	N/A	**3.52 ms**	2900.1936
Gradient Descent	η=0.01	50.62 ms	2899.7029
Ridge Regression	λ=1	32.53 ms	2866.4258
PCA + OLS	k=8 Components	N/A	**2848.3308**

3.2 Optimization and Convergence

Gradient Descent successfully converged. Figure 1 illustrates the MSE reduction over epochs for the tested learning rates. The learning rate of η=0.01 provided the best balance of speed and stability.

Figure 1. Gradient Descent Convergence. The plot shows the decrease in MSE over 1000 epochs. η=0.01 (middle curve) provided the fastest convergence without oscillation.

3.3 Impact of Regularization and Dimensionality

Ridge Regression: Figure 2 illustrates the effect of L2 regularization. As λ increases, the magnitude of the coefficients is driven toward zero. The minimum test error was found at λ=1.

PCA: Regression on the PCA-transformed data achieved the lowest overall Test MSE (2848.3308) when using 8 out of 10 principal components. This result indicates that the noise or irrelevant information was successfully filtered out by discarding the two components with the lowest variance.

4. Discussion

4.1 Insights and Takeaways

Efficiency and Stability: OLS and SVD-based methods yield identical results but SVD is 10 times faster and more robust to ill-conditioned data, making it the preferred closed-form technique for moderate datasets.

Generalization: Both Ridge Regression (λ=1) and PCA (k=8) improved the model's generalization performance by ~ 1% and ~ 1.8% respectively, compared to the unregularized OLS (2900.1936). This highlights the benefit of managing model complexity (variance) over raw fit (bias).

Dimensionality vs. Regularization: The minor difference in performance (PCA: 2848.3308 vs. Ridge: 2866.4258) suggests that while regularization effectively shrinks coefficients, removing the two least important orthogonal directions via PCA was slightly more effective in isolating the signal for this specific dataset.

4.2 Limitations

The study employed a standard Batch Gradient Descent, which is less computationally efficient and slower to converge than modern variants like Mini-Batch or Stochastic Gradient Descent for large-scale problems. Furthermore, the hyperparameter search space for λ in Ridge Regression and η in GD was limited, potentially missing slightly better optimal parameters.

5. Conclusion

The comparative analysis demonstrated that approaches which actively manage model complexity—either through feature subspace transformation (PCA) or coefficient regularization (Ridge)—provide superior predictive accuracy on unseen data compared to standard OLS. The PCA-based regression model (using k=8 components) was the optimal predictor. Future work should focus on implementing Lasso (L1 regularization) for feature selection and exploring more computationally efficient iterative optimizers, such as Stochastic Gradient Descent (SGD), to enhance scalability.

References

Project Implementation Notebook: MFAI_Project.ipynb.