Model Selection with Regularization

Youri Halmaert

June 12, 2025

1 Introduction

This report documents a machine learning experiment focused on model selection using different regularization techniques. The study uses the Digits Dataset from scikit-learn to train a classifier that discriminates between small digits (0-4) and large digits (5-9). The primary objectives are:

- Understanding the effects of feature scaling
- Implementing and comparing L1 and L2 regularization
- Performing cross-validation for hyperparameter tuning
- Analyzing the sparsity properties of different regularization methods

2 Dataset Preparation

The Digits Dataset contains 1,797 samples with 64 features representing 8x8 pixel images of handwritten digits. The target variable was transformed into a binary classification problem:

$$y_{\text{new}} = \begin{cases} 0 & \text{if digit} \leq 4 \\ 1 & \text{if digit} > 4 \end{cases} \tag{1}$$

The data was split into training (80%) and testing (20%) sets. Features were standardized using StandardScaler, which transforms data to have zero mean and unit variance:

$$z = \frac{x - \mu}{\sigma} \tag{2}$$

3 Regularization Techniques

3.1 L2 Regularization (Ridge)

The L2 penalty adds the squared sum of coefficients to the loss function:

$$\text{Loss} + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{3}$$

Key observations:

- Best λ (inverse of C) found through cross-validation: 0.001
- Test accuracy: 91.39%
- Maintains all features but shrinks coefficients

3.2 L1 Regularization (Lasso)

The L1 penalty adds the absolute sum of coefficients:

$$\text{Loss} + \lambda \sum_{j=1}^{p} |\beta_j| \tag{4}$$

Key observations:

- Best λ found: 0.01
- Test accuracy: 91.39%
- Creates sparse solutions by setting some coefficients to zero

Table 1: Regularization Comparison

Metric	L1	L2
Best C value	100	1000
Sparsity at C=1	6.25%	4.69%
Sparsity at C=0.1	32.81%	4.69%
Sparsity at C=0.01	85.94%	4.69%
Test Accuracy	91.39%	91.39%

4 Cross-Validation

5-fold cross-validation was used to select the optimal regularization parameter. The average accuracy across folds was computed for each candidate λ value.

5 Key Findings

- L1 regularization produces sparser models than L2
- Both methods achieved similar test accuracy (91.4%)
- Feature scaling is crucial for regularization to work effectively
- The optimal λ differs between L1 and L2 regularization

6 Conclusion

This experiment demonstrated the importance of proper model selection and hyperparameter tuning. While both regularization techniques achieved similar predictive performance, they differ significantly in their effect on model coefficients. L1 regularization's ability to produce sparse models makes it particularly useful for feature selection, while L2 regularization tends to maintain all features with small coefficients.

The results suggest that for this particular classification task, the choice between L1 and L2 regularization may not significantly impact model performance, but could be important depending on whether feature selection is desired.