

Handwritten Digit Recognition Using Machine Learning Techniques

Project Overview

This graduate-level project focuses on implementing and comparing different machine learning algorithms for handwritten digit recognition using the MNIST dataset. Students will explore various classification techniques including neural networks, linear regression, Naive Bayes, and logistic regression using scikit-learn.

Dataset

DIDA Dataset (10k version)

- Source: <https://didadataset.github.io/DIDA/>
- Dataset size: 10,000 samples
- Input: 28x28 images
- Output: 10 classes (digits 0-9)
- Format: Each image is represented as a 784-dimensional vector (28×28 pixels)

Project Objectives

1. Implement and compare three different machine learning approaches:
 - Multi-layer Perceptron Neural Network
 - Linear Regression (with One-vs-All strategy)
 - Logistic Regression
 - Naive Bayes Classifier
2. Analyze and evaluate model performance using various metrics:
 - Accuracy
 - Precision
 - Recall
 - F1-score
 - Confusion Matrix

Implementation Guidelines

1. **Data Preprocessing**
 - Normalize pixel values (scale to [0,1])
 - Split data into training (80%) and testing (20%) sets
2. **Neural Network Implementation**
 - Architecture: Multi-layer Perceptron
 - Hidden layers: Minimum 2 layers
 - Hyperparameters to tune:
 - Number of neurons per layer

- Learning rate
- Number of epochs
- Batch size

3. Linear Regression

- Implement One-vs-All strategy
- Feature engineering considerations

4. Logistic Regression

- Multi-class classification setup
- Convergence criteria

Analysis Requirements

Performance Analysis

1. Cross-Validation

- Implement k-fold cross-validation ($k=5$)
- Report mean and standard deviation of performance metrics

2. Comparative Analysis

- Compare computational efficiency
- Analyze training time vs. accuracy trade-offs

Results Discussion

Students should provide a comprehensive discussion addressing:

1. Model Performance

- Quantitative comparison of all three approaches
- Analysis of each model's strengths and weaknesses
- Discussion of failure cases and potential improvements

2. Hyperparameter Sensitivity

- Impact of different hyperparameter choices
- Justification for final parameter selections

3. Real-world Applicability

- Scalability considerations
- Potential applications and limitations

Deliverables

1. Code Repository

- Well-documented implementation of all three models
- Data preprocessing scripts
- Evaluation scripts

2. Technical Report

- Experimental results
- Comparative analysis

3. Presentation

- 5-minute presentation of findings
- Visual demonstrations of results