Report: MNIST Classification Using Least Squares

Roll No: 22CS30010

November 1, 2024

1 Objective

To classify the MNIST dataset using a least squares approach and evaluate its performance. This assignment restricts the solution method to the `np.linalg.lstsq` function for solving the least squares problem.

2 Assignment Tasks and Implementation

2.1 Data Loading

- **Method**: Used `fetch_openml` from `sklearn.datasets` to load the MNIST dataset.
- **Splitting**: The dataset was split into training and test sets using `train_test_split`.

2.2 Data Preprocessing

- **Normalization**: Pixel values were scaled to fall between 0 and 1 for improved training performance.
- **One-Hot Encoding**: Implemented a function to manually convert integer labels to one-hot encoded vectors.

2.3 Least Squares Model

- **Weight Matrix Computation**: Applied `np.linalg.lstsq` to calculate the weight matrix W.
- **Implementation**:

```
W, _, _, _ = np.linalg.lstsq(X_train, y_train_one_hot, rcond=None)
```

2.4 Prediction and Evaluation

- **Training and Test Prediction**: Used the weight matrix W to predict classes for both training and test sets.
- **Accuracy Calculation**:

```
train_accuracy = np.mean(train_predictions == y_train) * 100
test_accuracy = np.mean(test_predictions == y_test) * 100
```

- **Results**:
  - Training Accuracy: Approximately 85.95%
  - Test Accuracy: Approximately 85.39%

2.5 Visualization

- **Correct Predictions**: Displayed 15 images with accurately predicted labels.
- **Optional Visualization**: Presented 15 misclassified images for qualitative analysis.

3 Discussion

3.1 Analysis of Misclassified Examples

The visual examination of misclassified examples revealed that most errors occurred with digits that share similar structures (e.g., '4' and '9', '3' and '5') or digits with high variation in handwriting.

3.2 Question: Why Might the Least Squares Classifier Perform Worse?

- **Explanation**: The least squares method minimizes squared errors, which is more suitable for regression tasks rather than classification. Unlike specialized classifiers such as logistic regression or neural networks, which directly model the probability of class membership, least squares treats classification as a regression problem, which can lead to suboptimal decision boundaries.
- **Other Limitations**:
  - **Sensitivity to Outliers**: Least squares is prone to being influenced by outliers, which can distort classification boundaries.
  - **Linear Assumptions**: It assumes a linear separability that might not hold for complex data patterns like those in handwritten digit recognition.