Detailed Report on MNIST Image Classification using Logistic Regression

---

1. Introduction

In this project, we implemented a logistic regression model to classify images from the MNIST dataset, which contains handwritten digits ranging from 0 to 9. Logistic regression, although a simple linear classifier, is effective for multiclass classification problems when the dataset has clear patterns and features, as is the case with MNIST.

---

2. Dataset Description

The MNIST dataset is composed of grayscale images of size 28x28 pixels, which represent digits from 0 to 9. The dataset contains 70,000 images in total, split into:

- **Training Set**: 60,000 images
- **Test Set**: 10,000 images

Each image is labeled with a digit (0-9). The goal of the model is to correctly classify unseen images based on these labels.

---

3. Preprocessing Steps

To prepare the data for logistic regression, we performed the following preprocessing steps:

1. **Normalization**: The pixel values were normalized to a range between 0 and 1 by dividing each value by 255.
2. **Flattening**: Each 28x28 image was flattened into a 1D vector of 784 features, since logistic regression expects a 1D input.

Additionally, the dataset was split into training and validation sets, with 80% of the data used for training and 20% for validation.

---

4. Model Architecture

We used logistic regression, which is typically used for binary classification. However, for multiclass problems like MNIST, we extended it to handle multiple classes using the following settings:

- **Solver**: `lbfgs` (an optimizer suitable for multiclass problems)
- **Max Iterations**: 1000 (to ensure convergence)
- **Multinomial Loss**: The model predicts one of several classes.

The softmax function was used to map the output probabilities to the 10 possible digit classes.

---

5. Training the Model

The logistic regression model was trained on the MNIST dataset using the training set. The model learned to map the pixel intensities of each image to the corresponding digit class label. The training took approximately 5 minutes on a modern CPU.

---

6. Model Evaluation and Results

The model was evaluated on both validation and test sets using various metrics, including precision, recall, F1-score, and accuracy.

- **Validation Accuracy**: 0.9224
- **Test Accuracy**: 0.9248

6.1. Validation Set Results

Class	Precision	Recall	F1-score	Support
0	0.96	0.97	0.96	1175
1	0.96	0.97	0.97	1322
2	0.90	0.90	0.90	1174
3	0.91	0.90	0.91	1219
4	0.93	0.94	0.94	1176
5	0.89	0.88	0.89	1104
6	0.95	0.95	0.95	1177
7	0.93	0.93	0.93	1299
8	0.90	0.88	0.89	1160
9	0.91	0.91	0.91	1194
Accuracy	**0.92**	-	-	**12000**
Macro avg	0.92	0.92	0.92	12000
Weighted avg	0.92	0.92	0.92	12000

6.2. Test Set Results

Class	Precision	Recall	F1-score	Support
0	0.96	0.98	0.97	980
1	0.96	0.98	0.97	1135
2	0.91	0.91	0.91	1032
3	0.93	0.89	0.91	1010
4	0.94	0.91	0.93	982
5	0.90	0.87	0.89	892
6	0.94	0.94	0.94	958
7	0.93	0.93	0.93	1028
8	0.88	0.89	0.88	974
9	0.91	0.90	0.91	1009
Accuracy	**0.92**	-	-	**10000**
Macro avg	0.92	0.92	0.92	10000
Weighted avg	0.92	0.92	0.92	10000

---

7. Conclusion

The logistic regression model performed remarkably well on the MNIST dataset, achieving over 92% accuracy on both the validation and test sets. The classification metrics, including precision, recall, and F1-score, indicate that the model is effective across all digit classes.

However, there are some digits, particularly 5, 8, and 9, where the model's precision and recall were slightly lower. This is likely due to similarities in the handwritten forms of these digits, which introduces some ambiguity in classification.

Given the simplicity of logistic regression, the results are impressive, but further improvements could be made by using more advanced models such as Convolutional Neural Networks (CNNs).

---

8. Future Work

To improve the performance further, we could explore:

- **Convolutional Neural Networks (CNNs)**, which are more suited to image data and have been shown to achieve near-perfect accuracy on the MNIST dataset.
- **Data Augmentation**, which can increase the dataset's size and diversity, improving model generalization.
- **Regularization Techniques**, such as L1 or L2 regularization, to prevent overfitting.

This report summarizes the key components and results of the logistic regression model for image classification.