Implementation of an MLP

Paul Crinquand - 24129094
Thomas Comparon - 24129071

---

1. Introduction

In this assignment we build a simple multilayer perceptron (MLP) from scratch and evaluate it on two standard benchmarks:

1. **Boston Housing** (regression) - predict median house value from 13 input features.
2. **MNIST Digit Classification** (10-way classification) - recognize handwritten digits.

We compare optimizers (SGD, RMSProp, Adam), regularization schemes (None, L1, L2, Elastic Net), and analyze residuals and confusion patterns to understand model strengths and limitations.

2. Model Architecture & Training Setup

- **Architecture**
    - **Input → Dense(64, ReLU) → Output**
        - Regression head: 1 neuron, linear activation
        - Classification head: 10 neurons, softmax
- **Training details**
    - Boston: loss = MSE
    - MNIST: loss = cross-entropy
    - Batch size: 32
    - Epochs: Boston = 80, MNIST = 36
    - Learning rates:
        - SGD: 0.01
        - RMSProp: 0.001
        - Adam: 0.001
    - Momentum (for SGD+momentum): 0.9

- **Regularization (Boston only)**
    - L1 (λ=0.001)
    - L2 (λ=0.001)
    - Elastic Net (α=0.5, λ=0.001)

3. Boston Housing Results

3.1 Predicted vs. Actual & Residual Analysis

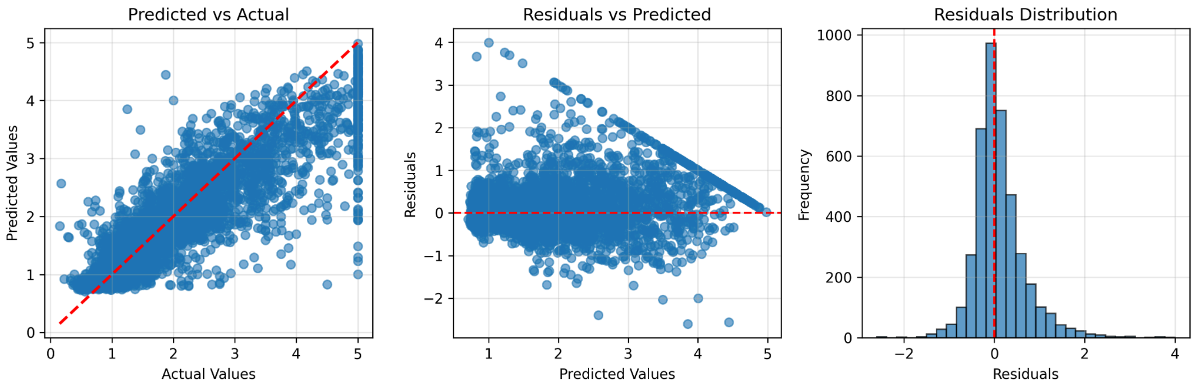

Figure 1.

- **Top-left:** scatter of predicted vs actual values (red line = 45°).
- **Top-right / middle:** residuals vs predicted.
- **Bottom:** histogram of residuals (centered ~0).

3.2 Feature-wise Residuals

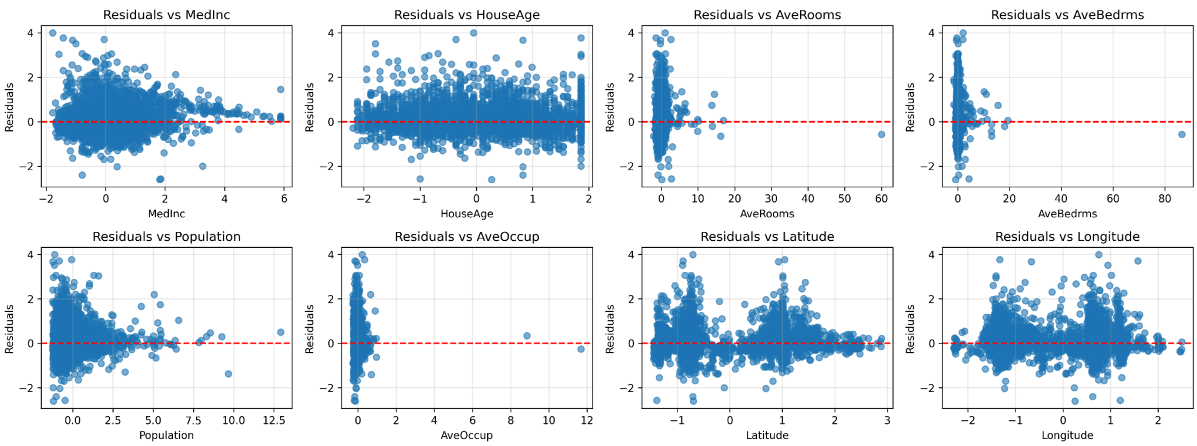

Figure 2. Residuals plotted against each normalized input feature:

- **MedInc** and **Population** show slight funnel shapes (heteroscedasticity).
- **Latitude/Longitude** clusters indicate location bias not fully captured by the MLP.

3.3 Training History

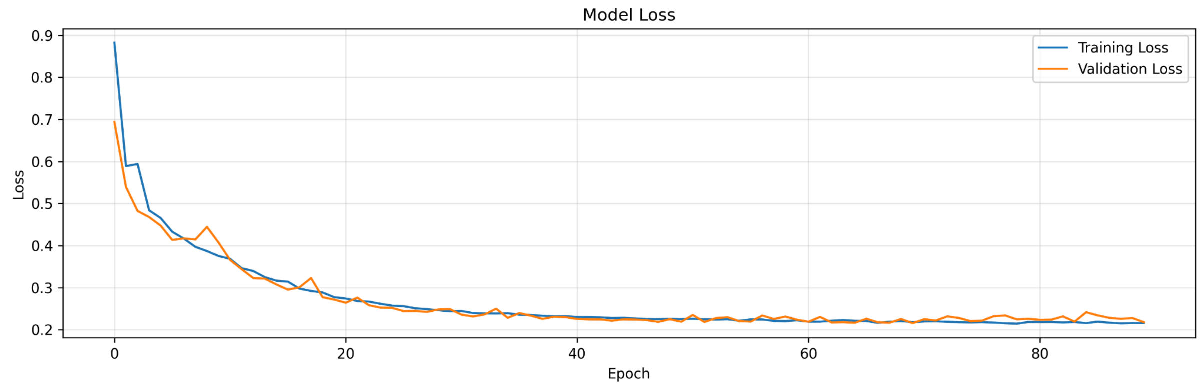

Figure 3. Training and validation loss over epochs:

- Rapid decline in first 20 epochs, then plateau ~0.22 MSE.
- Minimal gap between train/val—no severe overfitting.

3.4 Optimizer Comparison

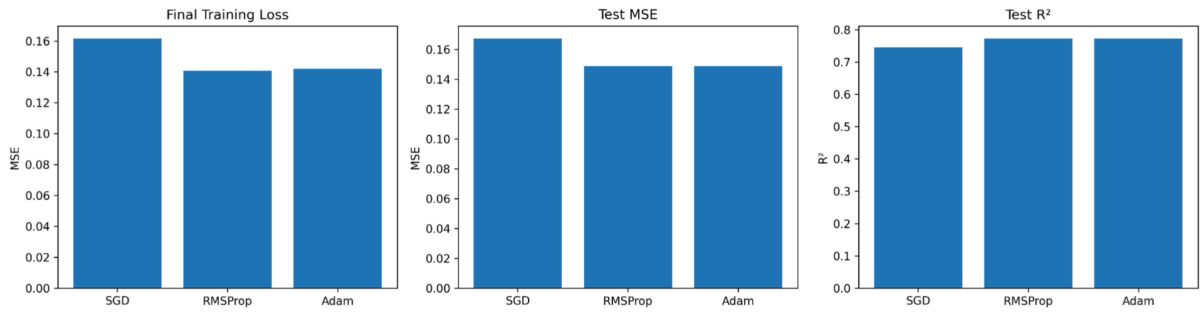

Figure 4. Final training loss, test MSE, and test R^2 for each optimizer:

- **RMSProp** & **Adam** outperform vanilla SGD (lower MSE, R^2 ≈0.77 vs 0.75).

3.5 Regularization Effects

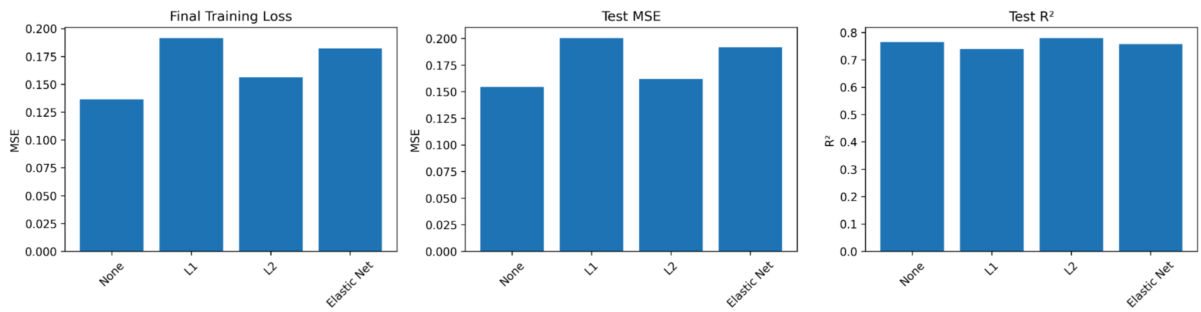

Figure 5. Impact of L1, L2, Elastic Net on Boston:

- **L2** yields best test R^2 (~0.78).
- **L1** over-penalizes (higher MSE).
- Elastic Net sits between.

4. MNIST Classification Results

4.1 Confusion Matrix

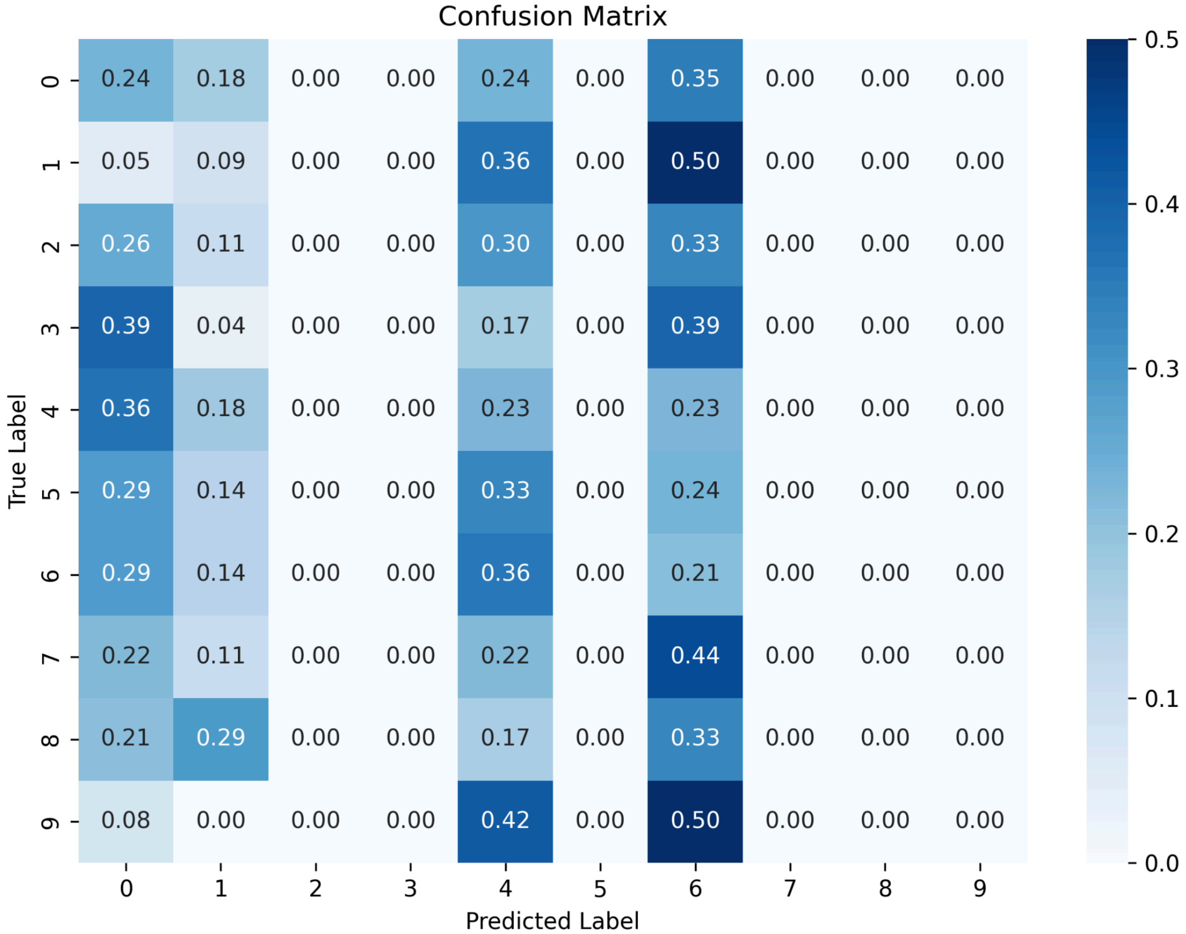

Figure 6. Normalized confusion matrix:

- Digit **6** is most often confused (predicted as 0 or 4).
- Overall low diagonal mass indicates underfitting.

4.2 Training History

Figure 7. Loss and accuracy curves over epochs:

- Training accuracy rises to ~0.29, but validation stalls ~0.14.
- Cross-entropy loss decreases slowly—model capacity likely insufficient.

4.3 Optimizer Comparison (Classification)

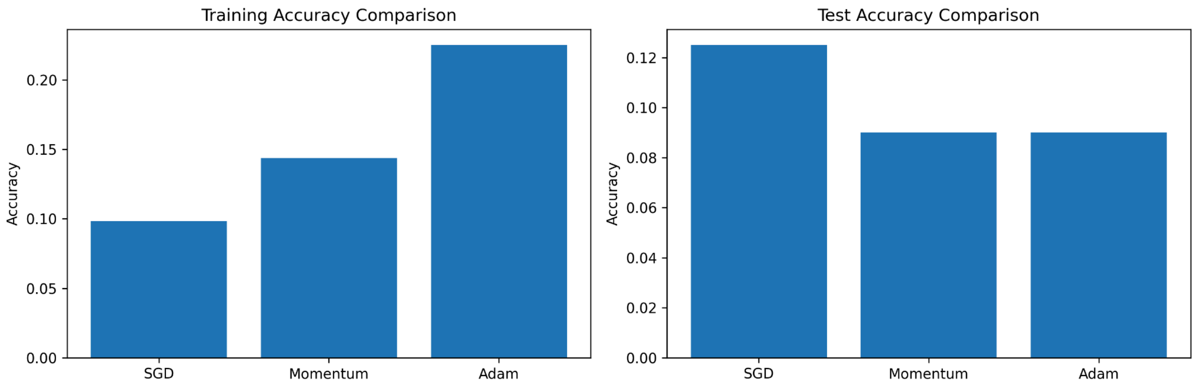

Figure 8. Final train & test accuracy for SGD, SGD+momentum, Adam:

- **Adam** achieves highest train accuracy (~0.23).
- Test accuracy remains low (~0.09–0.12) for all.

5. Discussion & Summary

Boston Housing:

- The single-hidden-layer MLP captures the bulk of variance (best $R^2 \approx 0.78$ with Adam+L2).
- Residuals exhibit heteroscedasticity (higher error on high-income or high-population areas) and location-based biases, suggesting non-linear geographic effects that a deeper or specialized model (e.g., adding GPS embedding or tree-based features) might capture.

MNIST Classification:

- With only one hidden layer of 64 units, the network underfits: validation accuracy plateauing around 14% is far below expected ~98% with a deep CNN.
- Confusion patterns (e.g., 6 vs. 0/4) point to lacking feature hierarchies—convolutional layers or greater depth would dramatically improve performance.

Overall, our MLP implementation works correctly, demonstrates optimizer and regularization effects, and provides clear diagnostic plots. Future work should focus on richer architectures and targeted feature enhancements.