Technical Report: Comparative Analysis of Linear and Non-Linear Classifiers for Handwritten Digit Recognition

Authors	ID
Youssef Mohamed Amin	202304504

Abstract

This study evaluates the performance and computational trade-offs of two distinct supervised machine learning classifiers, **Logistic Regression** (a linear model) and the **Support Vector Machine (SVM)** utilizing a Radial Basis Function (RBF) kernel (a non-linear model), applied to the task of handwritten digit recognition. Data preprocessing involved normalization and standardization. The Logistic Regression model achieved a superior predictive accuracy of **0.9750**. This significantly exceeds the SVM model's accuracy of **0.9356**, which was trained on a restricted data subset. The results indicate that inadequate training data volume can negate the theoretical advantage of non-linear kernel methods, making the simpler Logistic Regression model more effective under the specified training constraints.

1. Introduction

Handwritten digit recognition, primarily benchmarked against the MNIST dataset, is fundamental in pattern recognition and machine learning. This report presents a formal comparison between the established linear model, Logistic Regression (LR), and a robust non-linear approach, the Support Vector Machine (SVM) with an RBF kernel. The objective is to quantify the accuracy differential between these two algorithmic families, analyze their respective computational demands, and determine the most effective classifier for the specific challenges posed by the 28×28 pixel image data under limited training conditions.

2. Methodology

2.1. Data Acquisition and Preprocessing

The `mnist_784` dataset, comprising 70,000 images, was the source of data.

1. **Normalization:** Raw pixel values, ranging $[0, 255]$, were normalized to the interval $[0.0, 1.0]$ to stabilize model optimization.
2. **Data Partition:** The dataset was partitioned into 80% training and 20% testing subsets (`test_size=0.2`), ensuring consistent validation on unseen data.

3. **Standardization:** For the SVM, features were standardized using `StandardScaler` (zero mean, unit variance). This step is mathematically prerequisite for SVM convergence and performance.

2.2. Classifier Implementation

All models were implemented using the `scikit-learn` library.

Classifier	Type	Kernel / Solver	Training Strategy
Logistic Regression (`log_clf`)	Linear	`lbfgs`	Trained on $\mathbf{100\%}$ of the normalized training set.
Support Vector Machine (`svm_clf`)	Non-linear	RBF (`rbf`)	Trained on a standardized $\mathbf{10,000 - \text{sample}}$ subset due to cubic complexity ($\mathcal{O}(n^3)$).

3. Results

The models were evaluated on the reserved 20% test set. The primary metric for comparison was the Classification Accuracy, which measures the proportion of correctly predicted instances.

Model	Overall Classification Accuracy
Logistic Regression (LR)	$\mathbf{0.9750}$
Support Vector Machine (SVM)	$\mathbf{0.9356}$
Performance Gain (SVM over LR)	$\mathbf{-0.0394}$

3.1. Detailed Performance Metrics

The Classification Reports provide granular per-class metrics (Precision, Recall, F1-Score), offering deeper insight into the models' performance on individual digits.

Logistic Regression Classification Report

Digit	Precision	Recall	F1-Score	Support
0	0.98	0.99	0.98	1402
1	0.99	0.99	0.99	1607
2	0.97	0.96	0.96	1369
3	0.97	0.95	0.96	1440

4	0.97	0.98	0.97	1385
5	0.97	0.95	0.96	1262
6	0.98	0.98	0.98	1378
7	0.97	0.97	0.97	1441
8	0.96	0.95	0.95	1373
9	0.95	0.96	0.95	1393

Micro-Average Accuracy: 0.9750

Support Vector Machine (SVM) Classification Report

Digit	Precision	Recall	F1-Score	Support
0	0.96	0.95	0.95	1402
1	0.99	0.98	0.99	1607
2	0.92	0.92	0.92	1369
3	0.91	0.88	0.89	1440
4	0.94	0.94	0.94	1385
5	0.89	0.91	0.90	1262
6	0.95	0.98	0.96	1378
7	0.96	0.95	0.95	1441
8	0.90	0.89	0.90	1373
9	0.92	0.92	0.92	1393

Micro-Average Accuracy: 0.9356

4. Discussion

4.1. Performance Validation

The results indicate a significant divergence from typical expectations: the **Logistic Regression (LR)** model demonstrated superior classification accuracy (**0.9750**), surpassing the Support Vector

Machine (SVM) model ($\mathbf{0.9356}$) by a margin of $\mathbf{3.94}$ percentage points. This reversal is primarily attributed to the training strategy necessitated by the SVM's high computational cost.

4.2. Detailed Error Analysis

The detailed reports (Section 3.1) reveal that the SVM suffered a disproportionate loss of performance on specific, structurally ambiguous digits:

- **Digit 5:** The SVM's F1-Score ($\mathbf{0.90}$) is significantly lower than LR's ($\mathbf{0.96}$), indicating frequent confusion with other digits (likely 3s, 6s, or 8s) due to inadequate training samples to define its complex boundary.
- **Digit 3 and 8:** Both models show lower performance on these highly curvilinear and similar digits, but the SVM's F1-Scores ($\mathbf{0.89}$ and $\mathbf{0.90}$ respectively) are notably worse than the LR's ($\mathbf{0.96}$ and $\mathbf{0.95}$).

4.3. Impact of Data Volume on SVM Performance

While SVMs with non-linear kernels are theoretically better suited for complex image data, their performance is highly sensitive to sufficient training data. The LR model was trained on the **full** training set ($\sim 56,000$ samples), allowing it to effectively map linear boundaries across the entire feature space. Conversely, the SVM was restricted to a **10,000-sample subset**. This limitation prevented the SVM from adequately learning the complex, non-linear feature space of the full MNIST dataset, leading to a substantial decrease in generalization accuracy on the test set.

4.4. Resource Consumption and Model Selection

The analysis underscores a critical decision point in resource-constrained machine learning: the computational constraint on the SVM led directly to inferior performance. This suggests that for high-dimensional classification tasks, when the entire training set cannot be leveraged by the SVM, a well-optimized linear model like Logistic Regression may yield superior results.

Shutterstock

5. Conclusion

The comparative analysis confirms that the Logistic Regression model is the superior classifier in this execution, achieving $\mathbf{0.9750}$ accuracy versus the SVM's $\mathbf{0.9356}$. This outcome is strongly correlated with the computational necessity to train the SVM on only $10,000$ samples, while the LR model utilized the full $\sim 56,000$ training samples. Future work should focus on implementing dimensionality reduction techniques, such as Principal Component Analysis (PCA), to reduce the feature space complexity, potentially enabling the SVM to train efficiently on the full dataset without sacrificing its non-linear modeling power.

6. References

- Scikit-learn Developers. (n.d.). `sklearn.linear_model.LogisticRegression`.
- Scikit-learn Developers. (n.d.). `sklearn.svm.SVC`.

- Scikit-learn Developers. (n.d.). **`sklearn.preprocessing.StandardScaler`**.
- Deng, L. (2012). The MNIST Database of Handwritten Digit Images. *IEEE Signal Processing*