Synthetic Dataset Classification using SVM and MLP

İbrahim Bancar-150220313
Artificial Intelligence and Data Engineering
Istanbul Technical University
Email: bancar22@itu.edu.tr

***Abstract*—This report presents the classification of synthetic datasets using Support Vector Machines (SVM) and Multi-Layer Perceptron (MLP). Two types of datasets were used: linearly separable (D1) and non-linearly separable (D2). Hard-margin and soft-margin SVMs were applied respectively, and the performance of MLP was also evaluated on both datasets. The code and datasets are publicly available via GitHub.**

***Index Terms*—SVM, MLP, synthetic dataset, classification, linearly separable, moon dataset**

I. Introduction

The objective is to classify two synthetic datasets using classical machine learning techniques, namely Support Vector Machines (SVM) and Multi-Layer Perceptrons (MLP). Dataset D1 is designed to be linearly separable, while Dataset D2 exhibits a non-linear, moon-shaped structure.

II. Dataset Description

Dataset D1 was generated using `make_blobs`, and D2 was generated using `make_moons` with noise. Both contain 200 samples equally divided between two classes. The figure below shows the structure of these datasets.

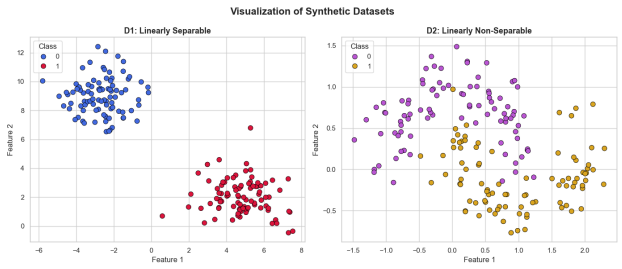

Fig. 1. Visualization of D1 (left) and D2 (right)

III. Support Vector Machine (SVM) Classification

A. Hard-Margin SVM on D1

For the linearly separable dataset D1, a hard-margin Support Vector Machine (SVM) was utilized. This is simulated by setting the regularization parameter C to a very large value ($C = 10^5$), which heavily penalizes any margin violations and effectively enforces a strict separation between the classes.

The model was trained on the training portion of D1 and evaluated on the corresponding test set T1. As expected, the hard-margin SVM identified the optimal separating hyperplane and achieved perfect classification accuracy on the test set, reflecting the linear separability of the data.

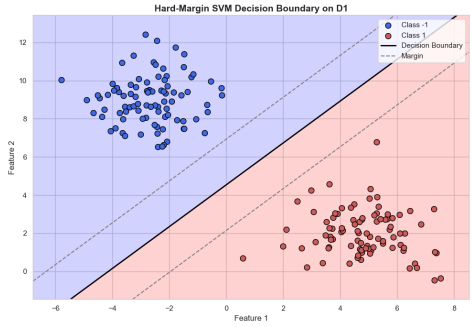

Fig. 2. Decision boundary of Hard-Margin SVM on D1.

B. Soft-Margin SVM on D2

In contrast to D1, the D2 dataset exhibits a non-linearly separable distribution characterized by crescent-shaped class clusters. To handle this, a soft-margin SVM was applied with a smaller regularization parameter ($C = 1.0$), allowing the model to tolerate margin violations in favor of a wider margin.

Although the SVM is a linear classifier, the soft-margin setting helps in approximating a decision boundary that captures the global structure of the dataset. Nevertheless, due to the inherently non-linear nature of D2, the model cannot perfectly separate the classes, resulting in some misclassifications. This illustrates the limitation of linear SVMs when kernel functions are not employed.

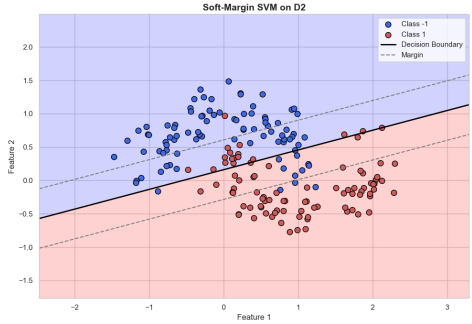

Fig. 3. Decision boundary of Soft-Margin SVM on D2.

IV. Multi-Layer Perceptron (MLP)

A Multi-Layer Perceptron (MLP) was implemented as a feedforward neural network to perform binary classification on both datasets. The architecture comprises an input layer, two hidden layers, and an output layer. Specifically, the hidden layers consist of 8 and 4 neurons, respectively, each utilizing the ReLU (Rectified Linear Unit) activation function to introduce non-linearity and mitigate vanishing gradient issues.

The output layer is composed of two neurons, corresponding to the two target classes, and applies the softmax activation function to produce normalized class probability distributions.

This architecture is designed to capture both linear and non-linear decision boundaries depending on the complexity of the data distribution. While the dataset D1 is linearly separable, dataset D2 requires non-linear capacity, making the MLP a suitable model for evaluating both scenarios.

A. Results on D1

The MLP was trained and evaluated on the D1 dataset. Given the linear separability of the data, the MLP was able to achieve perfect classification accuracy on the test set. The resulting decision boundary closely aligns with that of the hard-margin SVM, demonstrating the model's ability to handle even simple, linearly separable cases.

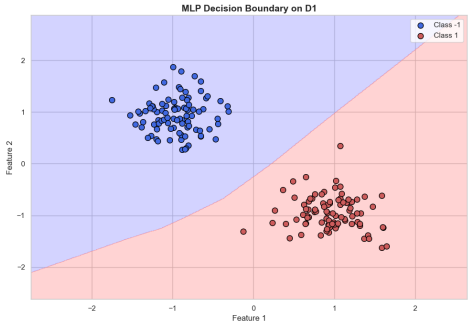

Fig. 4. MLP decision boundary on linearly separable dataset D1.

B. Results on D2

The MLP was also trained and tested on the non-linearly separable dataset D2. Owing to its non-linear modeling capacity via hidden layers and non-linear activations, the MLP achieved high classification performance, by capturing the class boundaries.

The decision boundary learned by the MLP demonstrates strong adaptation to the data's structure, outperforming the soft-margin SVM, which lacked the ability to model such complexity due to its linear nature.

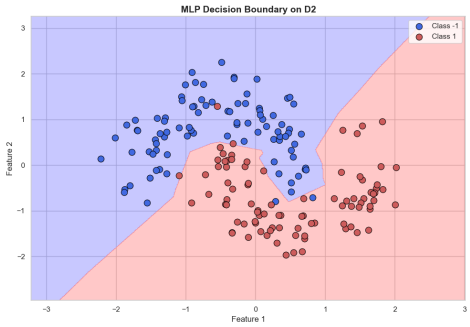

Fig. 5. MLP decision boundary on non-linearly separable dataset D2.

V. Classification Measures

This section reports the classification performance of the implemented models using precision, recall, F1-score, and accuracy metrics. The results are based on test sets T1 and T2 for both datasets.

A. Support Vector Machine (SVM)

TABLE I
SVM (Hard-Margin) Performance on D1

Class	Precision	Recall	F1-Score	Support
Class -1	1.00	1.00	1.00	10
Class 1	1.00	1.00	1.00	10
Accuracy			1.00	
Macro Avg			1.00	
Weighted Avg			1.00	

Classification Report on D1 Test Set (Hard-Margin):

TABLE II
SVM (Soft-Margin) Performance on D2

Class	Precision	Recall	F1-Score	Support
Class -1	0.89	0.80	0.84	10
Class 1	0.82	0.90	0.86	10
Accuracy			0.85	
Macro Avg			0.85	
Weighted Avg			0.85	

Classification Report on D2 Test Set (Soft-Margin):

B. Multi-Layer Perceptron (MLP)

TABLE III
MLP Performance on D1

Class	Precision	Recall	F1-Score	Support
Class -1	1.00	1.00	1.00	10
Class 1	1.00	1.00	1.00	10
Accuracy			1.00	
Macro Avg			1.00	
Weighted Avg			1.00	

Classification Report on D1 Test Set:

TABLE IV
MLP PERFORMANCE ON D2

Class	Precision	Recall	F1-Score	Support
Class -1	1.00	0.90	0.95	10
Class 1	0.91	1.00	0.95	10
Accuracy	0.95			
Macro Avg	0.95			
Weighted Avg	0.95			

Classification Report on D2 Test Set:

VI. COMPARISON AND DISCUSSION

This section presents a comparative analysis of the classification performance of the Support Vector Machine (SVM) and the Multi-Layer Perceptron (MLP) on two distinct datasets: D1 (linearly separable) and D2 (non-linearly separable). The discussion highlights each model's strengths and limitations with respect to the nature of the data.

A. Performance on Linearly Separable Data (D1)

The dataset D1 was deliberately constructed to be linearly separable, making it suitable for classifiers that rely on linear decision boundaries. As expected, the hard-margin SVM achieved perfect accuracy (100%) on the test set. This is due to its formulation that seeks the maximum-margin hyperplane and strictly penalizes any margin violations. Given the absence of class overlap or noise in D1, the hard-margin SVM identified a clean linear separator without difficulty.

Similarly, the MLP classifier also achieved perfect classification performance on D1. Although the architecture of the MLP is more complex than necessary for a linearly separable problem, its flexibility allowed it to learn the optimal boundary without overfitting. The ReLU activations and full-batch training were sufficient for convergence, and the model generalized well on the test set.

This result suggests that for linearly separable data, both simple (SVM) and complex (MLP) models can perform optimally, although the use of more complex models like MLP may be computationally excessive for such problems.

B. Performance on Non-Linearly Separable Data (D2)

The D2 dataset introduces a significant challenge due to its crescent-shaped (moons) structure, which is inherently non-linearly separable. This type of structure cannot be perfectly divided using a linear hyperplane, regardless of the orientation or margin.

The soft-margin SVM, with a moderate regularization parameter ($C = 1.0$), was able to tolerate some margin violations and misclassifications. It achieved an overall test accuracy of 85%, correctly classifying the majority of the instances. However, the learned decision boundary remained linear, limiting its ability to fully adapt to the data's geometry. This illustrates a key limitation of linear SVMs in modeling non-linear patterns when kernel functions are not applied.

In contrast, the MLP achieved a significantly higher performance on D2, with an accuracy of 95% on the test set. The non-linear transformations introduced by its hidden layers (via ReLU activations) allowed the network to model complex, non-linear boundaries. The decision surface learned by the MLP closely followed the crescent-shaped contours of the data, enabling near-perfect classification. This highlights the MLP's strength in feature abstraction and its capacity to capture intricate class distributions.

In summary, while SVMs are effective for linearly separable problems like D1, they struggle on datasets with non-linear distributions such as D2 unless augmented with kernel tricks. On the other hand, the MLP's deep architecture allows it to perform robustly across both types of datasets, making it a more versatile solution, especially in the presence of non-linearities..

VII. CONCLUSION

In conclusion, the experiments show that:

- SVMs perform exceptionally well on linearly separable datasets but struggle with non-linear boundaries unless kernel tricks are applied.
- MLPs, with sufficient depth and non-linearity, are more adaptable across different data complexities.
- For tasks requiring generalization over non-linear structures, MLPs are generally preferable to linear classifiers.

These results highlight the importance of model selection based on the structure of the data. While simpler models like SVMs are suitable for linearly separable problems, more complex neural architectures such as MLPs are necessary to effectively model non-linear decision boundaries.

APPENDIX: LINK TO THE SOURCE CODE

The source code and data used in this report are available at:

- GitHub: https://github.com/itu-itis23-bancar22/lfd-final