Lab Report: Support Vector Machine (SVM) Classification

Arham Jain, 23/CS/073

1. Introduction

This experiment focuses on training and evaluating different Support Vector Machine (SVM) models using various kernel functions. The goal is to understand how kernel choice and hyperparameters such as C and γaffect the model's decision boundary, accuracy, and generalization performance. We compare linear, RBF, polynomial, and the best-tuned model from GridSearchCV.

2. Model Comparison

Table shows the validation accuracy for four SVM models trained and tested on the scaled dataset.

Table : Validation Accuracy of Different SVM Models

Model	Validation Accuracy
SVC(kernel='linear',C=1)	0.842857
SVC(kernel='ref',default params)	0.908571
SVC(kernel='poly',degree =3)	0.854286
GridSearch CV Best Model	0.946667

3. Decision Boundary Visualization

Figure shows the decision boundaries for three models: the linear SVM, the default RBF kernel, and the best-tuned RBF model found via GridSearchCV. These plots illustrate how different kernels produce different decision regions for the same dataset.

4. Final Performance Evaluation

The final tuned SVM model was evaluated on the 30% hold-out validation set. The classification report and confusion matrix are presented below.

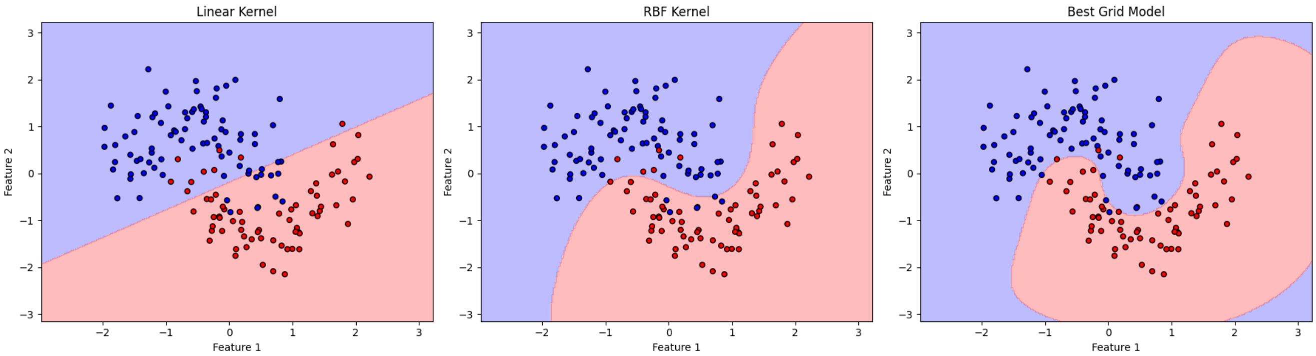

Figure : Comparison of decision boundaries for Linear, RBF, and Best RBF models.

Classification Report

	precision	recall	f1-score	support
0	0.85	1.00	0.92	168
1	1.00	0.84	0.91	182
accuracy			0.92	350
macro avg	0.93	0.92	0.92	350
weighted avg	0.93	0.92	0.92	350

Confusion Matrix

Predicted 0	Predicted 1
168	0
29	153

5. Conclusion and Discussion

Why did the linear SVM fail, and why did the RBF kernel succeed?

The linear SVM failed to achieve perfect accuracy because the dataset is not linearly separable. The classes overlap in certain regions of the feature space, and a straight-line (hyperplane) boundary cannot fully separate them. The RBF (Radial Basis Function) kernel, on the other hand, maps the input data into a higher-dimensional space, allowing for non-linear boundaries that can capture complex patterns in the data. This flexibility enabled the RBF kernel to correctly classify most of the samples and achieve higher accuracy.

What did the GridSearchCV tell you? What were the best C and gamma?

The GridSearchCV identified the optimal hyperparameters as:

$$C = 10,\ \gamma = 1,\ \text{kernel} = \text{'rbf'}$$

These values strike a balance between bias and variance, providing a model that generalizes well to unseen data while maintaining high training accuracy.

Effect of extreme hyperparameters

- **When γ is too high (e.g., 1000):** The model becomes overly sensitive to individual data points. It creates very tight, complex decision boundaries that perfectly fit the training data but fail to generalize — a clear case of overfitting.
- **When C is too low (e.g., 0.01):** The model becomes too tolerant of misclassifications, resulting in a wider margin but poorer separation between classes. This leads to underfitting and lower accuracy, as the decision boundary becomes overly smooth.

Overall Conclusion

The experiment demonstrated that kernel choice and hyperparameter tuning are critical in SVM performance. The linear kernel performed adequately on simpler data, but the RBF kernel captured more complex structures, resulting in higher accuracy. Grid- SearchCV provided an automated way to identify optimal C and γ values, leading to the best-performing model with a validation accuracy of 94.67%. Understanding these relationships helps in selecting the right SVM configuration for real-world datasets.