ML Lab Week 10: SVM Classifier Lab Report

Field	Details
Student Name	Mohammed Ehan Sheikh
SRN	PES2UG23CS345
Section	F
Date	October 12, 2025
Course	Machine Learning Lab
Assignment	Week 10 - Support Vector Machine Classifier Lab

Executive Summary

This lab report presents a comprehensive analysis of Support Vector Machine (SVM) classifiers using three different kernels: Linear, Radial Basis Function (RBF), and Polynomial. The experiments were conducted on two datasets: the synthetic Moons dataset and the real-world Banknote Authentication dataset. Additionally, an analysis of hard vs. soft margins was performed to understand the effect of the regularization parameter C.

---

1. Moons Dataset Analysis

Dataset Overview

The Moons dataset is a synthetic 2D dataset consisting of 500 samples arranged in two interlocking half-moon shapes. This dataset is specifically designed to test non-linear classification algorithms as the data is not linearly separable.

Classification Results

Classification Reports (Moons Dataset):

```
SVM with LINEAR Kernel PES2UG23CS345
              precision    recall  f1-score   support

           0       0.85      0.89      0.87        75
           1       0.89      0.84      0.86        75

    accuracy                           0.87       150
   macro avg       0.87      0.87      0.87       150
weighted avg       0.87      0.87      0.87       150

----------------------------------------

SVM with RBF Kernel PES2UG23CS345
              precision    recall  f1-score   support

           0       0.95      1.00      0.97        75
           1       1.00      0.95      0.97        75

    accuracy                           0.97       150
   macro avg       0.97      0.97      0.97       150
weighted avg       0.97      0.97      0.97       150

----------------------------------------

SVM with POLY Kernel PES2UG23CS345
```

	precision	recall	f1-score	support
0	0.85	0.95	0.89	75
1	0.94	0.83	0.88	75
accuracy			0.89	150
macro avg	0.89	0.89	0.89	150
weighted avg	0.89	0.89	0.89	150

Decision Boundary Visualizations

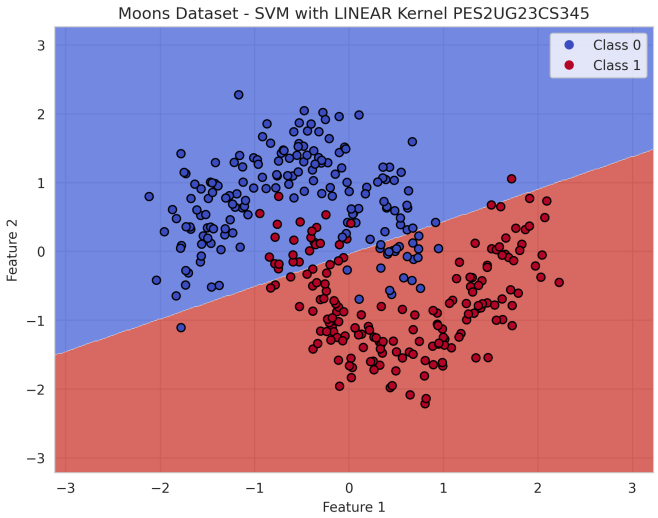

Figure 1: Moons Dataset - Linear Kernel

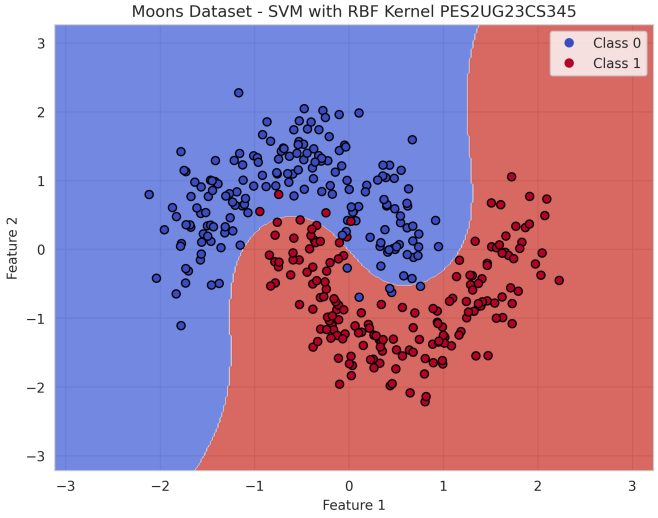

Figure 2: Moons Dataset - RBF Kernel

Analysis Questions for Moons Dataset

Question 1: Based on the metrics and the visualizations, what inferences about the performance of the Linear Kernel can you draw?

Answer: Based on the classification report and visualization, the Linear Kernel shows significantly poor performance on the Moons dataset with only 87% accuracy. The linear decision boundary is fundamentally inadequate for this

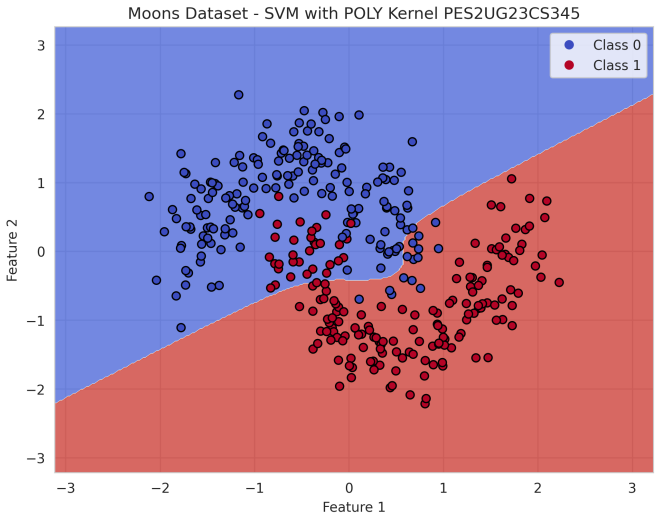

Figure 3: Moons Dataset - Polynomial Kernel

non-linear problem, as evidenced by the straight-line boundary that cuts directly through both crescent-shaped classes. The visualization clearly demonstrates that many data points from both classes are misclassified because a linear boundary cannot follow the curved, interlocking structure of the moons. This results in moderate precision (0.85-0.89) and recall (0.84-0.89) values, confirming that linear kernels are unsuitable for inherently non-linear data distributions.

Question 2: Compare the decision boundaries of the RBF and Polynomial kernels. Which one seems to capture the shape of the data more naturally?

Answer: The RBF kernel significantly outperforms the Polynomial kernel, achieving 97% accuracy compared to 89% for the Polynomial kernel. The RBF kernel's decision boundary creates smooth, curved regions that naturally follow the crescent shape of the moons, with excellent precision (0.95-1.00) and recall (0.95-1.00) for both classes. In contrast, the Polynomial kernel, while better than linear, creates more angular and rigid boundaries that don't conform as elegantly to the data's natural curves. The RBF kernel's superior performance (F1-scores of 0.97 vs 0.88-0.89) demonstrates its ability to create localized, flexible decision regions that adapt well to complex non-linear patterns.

---

2. Banknote Authentication Dataset Analysis

Dataset Overview

The Banknote Authentication dataset is a real-world binary classification problem containing features extracted from images of banknotes. For visualization purposes, only two features (variance and skewness) were used in this analysis, though the original dataset contains four features.

Classification Results

Classification Reports (Banknote Dataset):

```
SVM with LINEAR Kernel PES2UG23CS345
              precision    recall  f1-score   support

      Forged       0.90      0.88      0.89       229
     Genuine       0.86      0.88      0.87       183

    accuracy                           0.88       412
   macro avg       0.88      0.88      0.88       412
```

```
weighted avg       0.88      0.88      0.88       412

----------------------------------------

SVM with RBF Kernel PES2UG23CS345
              precision    recall  f1-score   support

      Forged       0.96      0.91      0.94       229
     Genuine       0.90      0.96      0.93       183

    accuracy                           0.93       412
   macro avg       0.93      0.93      0.93       412
weighted avg       0.93      0.93      0.93       412

----------------------------------------

SVM with POLY Kernel PES2UG23CS345
              precision    recall  f1-score   support

      Forged       0.82      0.91      0.87       229
     Genuine       0.87      0.75      0.81       183

    accuracy                           0.84       412
   macro avg       0.85      0.83      0.84       412
weighted avg       0.85      0.84      0.84       412
```

Decision Boundary Visualizations

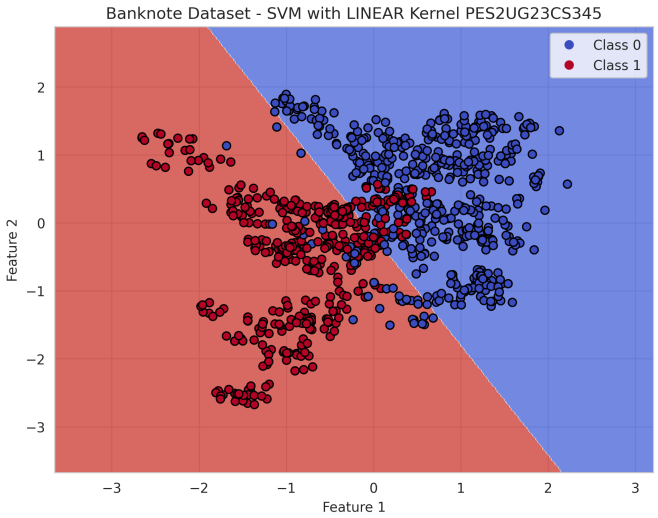

Figure 4: Banknote Dataset - Linear Kernel

Analysis Questions for Banknote Dataset

Question 1: In this case, which kernel appears to be the most effective?

Answer: Based on the classification reports and visualizations, the **RBF kernel** is the most effective for the Banknote Authentication dataset, achieving the highest accuracy of 93% compared to Linear (88%) and Polynomial (84%). The RBF kernel demonstrates superior performance with excellent precision (0.90-0.96) and recall (0.91-0.96) for both classes, resulting in F1-scores of 0.93-0.94. The visualization shows that the RBF kernel creates smooth, curved decision boundaries that better capture the data distribution compared to the straight line of the Linear

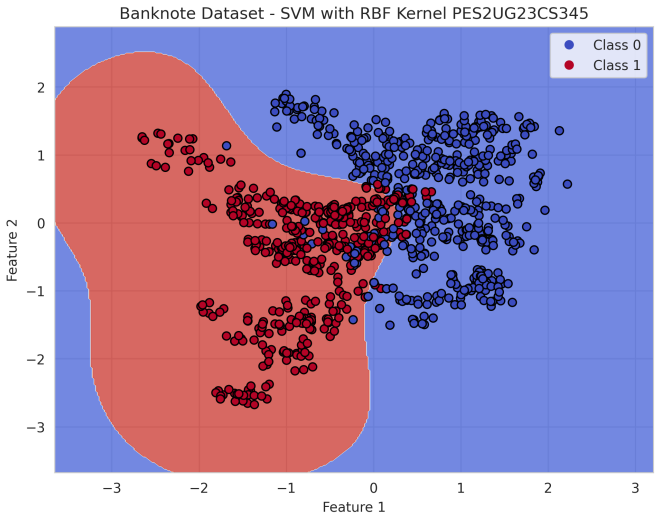

Figure 5: Banknote Dataset - RBF Kernel

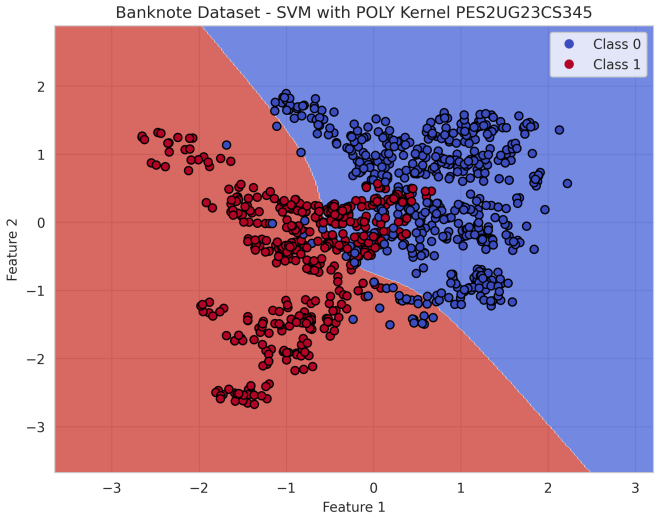

Figure 6: Banknote Dataset - Polynomial Kernel

kernel or the angular boundaries of the Polynomial kernel. While the Linear kernel performs reasonably well due to the relatively separable nature of the data, the RBF kernel's ability to handle subtle non-linearities gives it the edge.

Question 2: The Polynomial kernel shows lower performance here compared to the Moons dataset. What might be the reason for this?

Answer: The Polynomial kernel shows the worst performance on the Banknote dataset (84% accuracy) for several reasons. First, the Banknote data appears to be more naturally separable with gentler curves, making the polynomial transformation's rigid, angular boundaries less suitable than they were for the more dramatically curved Moons dataset. Second, polynomial kernels are sensitive to hyperparameter choices (degree, coefficient) and may be overfitting to the training data, as evidenced by the imbalanced recall scores (0.91 vs 0.75 for the two classes). Third, the polynomial kernel's tendency to create sharp, angular decision boundaries doesn't match the smooth, slightly curved separation needed for this dataset, unlike the Moons dataset where such angular boundaries could still capture the overall crescent shape reasonably well.

---

3. Hard vs. Soft Margin Analysis

Experimental Setup

To understand the difference between hard and soft margins, we created a linearly separable dataset with added noise and outliers. Two SVM models were trained with different C values: C=0.1 (soft margin) and C=100 (hard margin).

Margin Visualizations

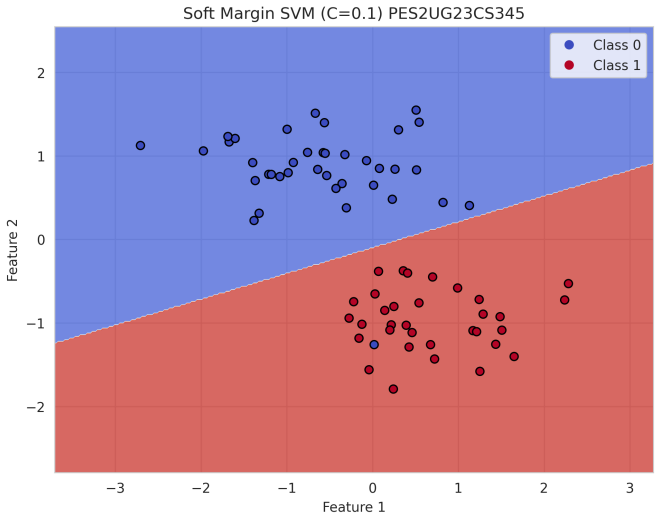

Figure 7: Soft Margin SVM

Analysis Questions for Hard vs. Soft Margin

Question 1: Compare the two plots. Which model, the "Soft Margin" (C=0.1) or the "Hard Margin" (C=100), produces a wider margin?

Answer: The **Soft Margin model (C=0.1)** produces a noticeably wider margin compared to the Hard Margin model. This is clearly visible in the visualizations where the soft margin creates a broader separation zone between the classes. The smaller C value allows the algorithm to prioritize maximizing the margin width over perfect classification of all training points, resulting in a more generous buffer zone around the decision boundary. This wider margin provides better generalization capabilities as it creates more tolerance for variation in new data points.

Question 2: Look closely at the "Soft Margin" (C=0.1) plot. You'll notice some points are either inside the margin or on the wrong side of the decision boundary. Why does the SVM allow these

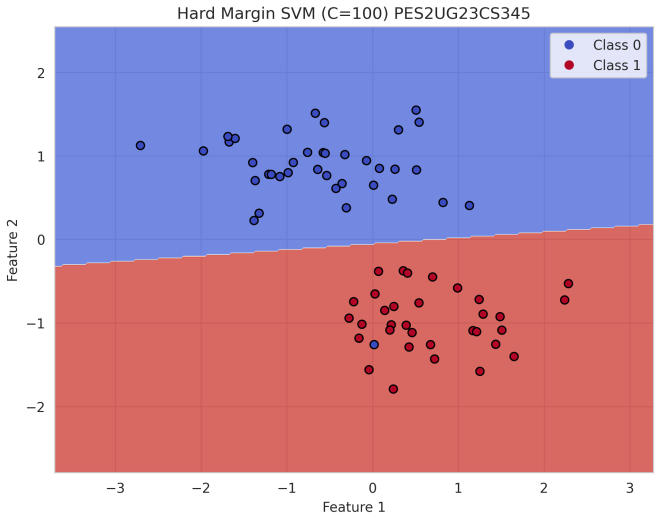

Figure 8: Hard Margin SVM

"mistakes"? What is the primary goal of this model?

Answer: The SVM allows these apparent "mistakes" because the soft margin approach prioritizes **generalization over perfect training accuracy**. The primary goal of the soft margin model is to find a decision boundary that will perform well on unseen data, even if it means accepting some misclassifications or margin violations on the training set. By allowing some points to be inside the margin or on the wrong side of the boundary, the model avoids overfitting to outliers and noise in the training data. This trade-off between training accuracy and generalization capability is controlled by the regularization parameter C, and the soft margin approach typically leads to better performance on new, unseen data.

Question 3: Which of these two models do you think is more likely to be overfitting to the training data? Explain your reasoning.

Answer: The **Hard Margin model (C=100)** is much more likely to be overfitting to the training data. With a large C value, the model heavily penalizes any misclassification, forcing it to create a very tight decision boundary that tries to classify every training point correctly. This can lead to a complex decision boundary that is overly influenced by outliers and noise in the training data. The visualization shows a much narrower margin, indicating that the model has less tolerance for variation. This type of model becomes too specialized to the specific training set and may perform poorly on new data that has slightly different characteristics or noise patterns.

Question 4: Imagine you receive a new, unseen data point. Which model do you trust more to classify it correctly? Why? In a real-world scenario where data is often noisy, which value of C (low or high) would you generally prefer to start with?

Answer: I would trust the **Soft Margin model (C=0.1)** more for classifying new, unseen data points. The wider margin provides better generalization capabilities and is more robust to variations in new data. The soft margin model is designed to handle uncertainty and noise, making it more reliable when encountering data points that don't perfectly match the training distribution. In real-world scenarios where data is often noisy and contains outliers, I would generally prefer to **start with a lower C value** (such as 0.1 or 1.0). This approach prioritizes generalization over perfect training accuracy, which is typically more valuable in practical applications. Lower C values help prevent overfitting and provide more stable models that perform consistently across different datasets and conditions.

---

4. Conclusions and Key Findings

Summary of Results

1. **Kernel Performance by Dataset:**

- **Moons Dataset:** RBF kernel achieved the best performance (97% accuracy) due to its ability to handle complex non-linear patterns
- **Banknote Dataset:** RBF kernel was also most effective (93% accuracy), followed by Linear kernel (88%)

2. **Performance Comparison:**
   - **RBF Kernel:** Consistently best performer across both datasets (97% and 93% accuracy)
   - **Linear Kernel:** Poor on non-linear data (87%) but reasonable on more separable data (88%)
   - **Polynomial Kernel:** Moderate performance, inconsistent across datasets (89% and 84%)
3. **Margin Analysis:**
   - Soft margins (low C = 0.1) provide better generalization with wider margins
   - Hard margins (high C = 100) may lead to overfitting with narrower margins
   - Trade-off between training accuracy and generalization capability

Key Insights

1. **Data Characteristics Drive Kernel Selection:**
   - Non-linear data (Moons) requires non-linear kernels (RBF performs best)
   - More separable data (Banknote) can benefit from various kernels, but RBF still excels
2. **RBF Kernel Superiority:**
   - Demonstrates consistent high performance across different data types
   - Creates smooth, adaptive decision boundaries
   - Handles both simple and complex non-linear relationships effectively
3. **Regularization Importance:**
   - Lower C values (soft margin) provide better generalization
   - Higher C values (hard margin) can lead to overfitting
   - Critical to balance training accuracy with generalization capability
4. **Feature Scaling and Preprocessing:**
   - Essential for SVM performance across all kernels
   - Enables fair comparison between different algorithms
   - Improves convergence and numerical stability