華南理工大學

South China University of Technology

The Experiment Report of Machine Learning

School: School of Software Engineering

Subject: Software Engineering

Author: Yuming Jiang	Supervisor: Mingkui Tan
Student ID: 202330550601	Grade: Undergraduate

October 24, 2025

Experiment 2: Classification Algorithm Comparison on Breast Cancer and Iris Datasets

Your Name
School of Software Engineering
South China University of Technology

Abstract—This experiment presents a comprehensive comparison of three classification algorithms: Logistic Regression, Linear Support Vector Machine (SVM), and RBF Kernel SVM. The evaluation is conducted on two standard benchmark datasets from the LIBSVM repository: the Breast Cancer dataset (683 samples, 10 features, binary classification) and the Iris dataset (150 samples, 4 features, multi-class classification).

The experimental methodology includes rigorous data preprocessing with feature standardization, hyperparameter optimization using 5-fold cross-validation with grid search, and comprehensive performance evaluation using multiple metrics including accuracy, confusion matrices, and ROC curves. Data visualization techniques including scatter plots and feature distribution histograms provide insights into dataset characteristics.

The results demonstrate that algorithm performance varies by dataset characteristics. On the Breast Cancer dataset, Logistic Regression achieves the highest test accuracy (96.10%) with a cross-validation score of 97.49%. On the Iris dataset, RBF Kernel SVM performs best (95.56%) with a CV score of 97.14%. The study provides practical guidelines for algorithm selection based on dataset linearity, computational constraints, and performance requirements.

I. Introduction

Classification algorithms form the foundation of supervised machine learning, enabling systems to automatically categorize data into predefined classes. The selection of appropriate classification algorithms is critical for real-world applications ranging from medical diagnosis to pattern recognition. While numerous classification methods exist, understanding their comparative performance on different types of data remains essential for practitioners.

This experiment focuses on three fundamental yet powerful classification algorithms: Logistic Regression, Linear SVM, and RBF Kernel SVM. These methods represent different approaches to the classification problem—probabilistic modeling, maximum margin optimization, and non-linear kernel methods, respectively. By evaluating these algorithms on both binary and multi-class classification tasks, we aim to understand their relative strengths and optimal application scenarios.

A. Motivation

The primary motivation stems from practical considerations in algorithm selection. While theoretical foundations provide important insights, empirical performance on real datasets often reveals nuances not apparent from theory alone. The Breast Cancer and Iris datasets, as standard benchmarks in machine learning, provide excellent test cases for comparing linear and non-linear classification methods.

B. Objectives

This experiment aims to:

1) Implement and evaluate three classification algorithms using proper hyperparameter optimization
2) Compare performance on binary classification (Breast Cancer) and multi-class classification (Iris) tasks
3) Analyze the impact of dataset characteristics (linearity, dimensionality, class distribution) on algorithm performance
4) Provide data-driven recommendations for algorithm selection in practical applications
5) Demonstrate comprehensive evaluation methodology including visualization and statistical analysis

II. Methods and Theory

A. Logistic Regression

Logistic Regression is a probabilistic linear classifier that models the posterior probability of class membership using the logistic (sigmoid) function:

$$P(y=1|\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x}+b) = \frac{1}{1+\exp(-\mathbf{w}^T\mathbf{x}-b)} \tag{1}$$

where $\mathbf{w}$ is the weight vector, b is the bias term, and $\sigma(\cdot)$ is the sigmoid function. For multi-class problems, the model is extended using the softmax function:

$$P(y=k|\mathbf{x}) = \frac{\exp(\mathbf{w}_k^T\mathbf{x}+b_k)}{\sum_{j=1}^{K}\exp(\mathbf{w}_j^T\mathbf{x}+b_j)} \tag{2}$$

The parameters are learned by maximizing the log-likelihood with L2 regularization:

$$\max_{\mathbf{w},b} \sum_{i=1}^{n} \log P(y_i|\mathbf{x}_i) - \frac{1}{2C}\|\mathbf{w}\|^2 \tag{3}$$

where C is the inverse regularization strength.

B. Linear Support Vector Machine

Linear SVM finds the maximum margin hyperplane separating different classes. The optimization problem is formulated as:

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \tag{4}$$

subject to:

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, n \tag{5}$$

where ξ_i are slack variables allowing for soft-margin classification, and C controls the trade-off between maximizing the margin and minimizing classification errors.

C. RBF Kernel SVM

For non-linear classification, SVM can be extended using kernel functions. The Radial Basis Function (RBF) kernel implicitly maps data to an infinite-dimensional feature space:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2\right) \tag{6}$$

where γ controls the influence radius of each training example. The dual optimization problem becomes:

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \tag{7}$$

subject to:

$$0 \leq \alpha_i \leq C, \quad \sum_{i=1}^{n}\alpha_i y_i = 0 \tag{8}$$

III. Experiments

A. Datasets

1) Breast Cancer Dataset: The Breast Cancer dataset, obtained from the LIBSVM repository, is a binary classification problem for diagnosing breast tumors as benign or malignant based on cell characteristics.

TABLE I
Breast Cancer Dataset Characteristics

Characteristic	Value
Total samples	683
Training samples	478 (70%)
Testing samples	205 (30%)
Benign (Class 0)	444 (65.0%)
Malignant (Class 1)	239 (35.0%)
Number of features	10
Feature type	Numerical (scaled)
Task type	Binary classification
Data source	LIBSVM repository

The dataset exhibits moderate class imbalance (approximately 2:1 ratio), which is common in medical diagnosis problems where positive cases are less frequent than negative cases.

TABLE II
Iris Dataset Characteristics

Characteristic	Value
Total samples	150
Training samples	105 (70%)
Testing samples	45 (30%)
Setosa (Class 0)	50 (33.3%)
Versicolor (Class 1)	50 (33.3%)
Virginica (Class 2)	50 (33.3%)
Number of features	4
Feature type	Numerical (scaled)
Task type	Multi-class classification
Data source	LIBSVM repository

2) Iris Dataset: The Iris dataset is a classic multi-class classification problem involving three species of iris flowers (Setosa, Versicolor, Virginica) based on sepal and petal measurements.

The dataset is perfectly balanced with equal representation of all three classes, making it ideal for evaluating multi-class classification performance without class imbalance considerations.

B. Implementation Details

1) Environment Configuration:

- Programming Language: Python 3.12.5
- Machine Learning Library: scikit-learn 1.7.2
- Numerical Computing: NumPy 2.3.3
- Visualization: Matplotlib 3.10.6, Seaborn
- Development Environment: Jupyter Notebook
- Random Seed: 42 (for reproducibility)

2) Data Preprocessing: Loading: Datasets were loaded using sklearn.datasets.load_svmlight_file and converted from sparse to dense format.

Label Transformation:

- Breast Cancer: Labels converted from {2, 4} to {0, 1}
- Iris: Labels converted from {1, 2, 3} to {0, 1, 2}

Feature Standardization: Applied StandardScaler to ensure zero mean and unit variance:

$$z = \frac{x - \mu}{\sigma} \tag{9}$$

This preprocessing is crucial for SVM and Logistic Regression, which are sensitive to feature scales.

Data Splitting: Stratified 70-30 train-test split to maintain class distribution.

3) Hyperparameter Optimization: All algorithms were optimized using GridSearchCV with 5-fold cross-validation:

Logistic Regression:

- C: [0.1, 1, 10]
- Penalty: ['l2']
- Solver: ['liblinear', 'lbfgs']
- Max iterations: 1000

Linear SVM:

- C: [0.1, 1, 10]
- Kernel: ['linear']

- Probability: True (for ROC curves)

RBF Kernel SVM:

- C: [1, 10]
- Gamma: ['scale', 0.1]
- Kernel: ['rbf']
- Probability: True

4) Evaluation Metrics:

- Accuracy: Overall classification correctness
- Cross-validation score: Mean accuracy across 5 folds
- Confusion Matrix: Detailed breakdown of predictions
- Precision, Recall, F1-score: Per-class performance metrics
- ROC Curve and AUC: For binary classification (Breast Cancer)

C. Data Visualization

Comprehensive visualization was performed to understand dataset characteristics:

- Scatter plots: Visualizing class separation in 2D feature space
- Feature distribution histograms: Analyzing feature distributions across classes (particularly for Breast Cancer dataset as per experiment requirements)
- Confusion matrices: Heatmaps showing classification results
- ROC curves: Evaluating binary classifier performance

IV. Results and Analysis

A. Breast Cancer Dataset Results

TABLE III
Breast Cancer Dataset Performance

Algorithm	Acc.	CV	Best Params
Logistic Reg.	96.10	97.49	C=0.1
Linear SVM	95.61	97.49	C=0.1
RBF SVM	95.61	96.45	C=1, γ=scale

1) Overall Performance: Logistic Regression achieved the highest test accuracy (96.10%) on the Breast Cancer dataset, demonstrating excellent performance on this linearly separable binary classification task. The high cross-validation score (97.49%) shared by both Logistic Regression and Linear SVM indicates that linear models are well-suited for this dataset.

2) Detailed Classification Metrics: The classification reports reveal strong performance across both classes:

TABLE IV
Breast Cancer - Logistic Regression Detailed Metrics

Class	Precision	Recall	F1-Score	Support
Benign (0)	0.97	0.98	0.97	133
Malignant (1)	0.95	0.92	0.94	72
Macro Avg	0.96	0.95	0.95	205
Weighted Avg	0.96	0.96	0.96	205

3) ROC Curve Analysis: ROC curves were generated for all three algorithms on the Breast Cancer dataset. The Area Under Curve (AUC) scores demonstrate excellent discrimination ability:

- Logistic Regression AUC: 0.99
- Linear SVM AUC: 0.99
- RBF Kernel SVM AUC: 0.99

All three algorithms achieved near-perfect AUC scores, indicating excellent ability to distinguish between benign and malignant tumors across all decision thresholds.

B. Iris Dataset Results

TABLE V
Iris Dataset Performance

Algorithm	Test Acc. (%)	CV Score (%)	Best C	Best γ
Logistic Regression	91.11	98.10	1.0	-
Linear SVM	91.11	98.10	1.0	-
RBF Kernel SVM	95.56	97.14	10.0	0.1

1) Overall Performance: On the Iris dataset, RBF Kernel SVM achieved the highest test accuracy (95.56%), suggesting the presence of non-linear patterns that benefit from kernel methods. The gap between linear methods (91.11%) and RBF SVM (95.56%) indicates that the Iris dataset's decision boundaries are not entirely linear.

2) Confusion Matrix Analysis: The confusion matrices reveal which classes are most easily confused:

RBF Kernel SVM (Best Performer):

- Setosa (Class 0): Perfectly classified (15/15 correct)
- Versicolor (Class 1): 13/15 correct (2 misclassified as Virginica)
- Virginica (Class 2): 15/15 correct

The confusion primarily occurs between Versicolor and Virginica, which is well-documented in machine learning literature as these species have overlapping feature distributions.

C. Comparative Analysis

TABLE VI
Algorithm Performance Across Datasets

Algorithm	Breast Cancer	Iris	Average
Logistic Regression	96.10%	91.11%	93.61%
Linear SVM	95.61%	91.11%	93.36%
RBF Kernel SVM	95.61%	95.56%	95.59%

1) Cross-Dataset Performance: RBF Kernel SVM achieves the highest average accuracy (95.59%) across both datasets, demonstrating its versatility in handling both linear and non-linear patterns.

2) Dataset Characteristics Impact: Breast Cancer (Linear):

- High-dimensional (10 features)
- Moderately imbalanced (2:1 ratio)

- Appears largely linearly separable
- Linear methods (LR, Linear SVM) perform excellently

Iris (Non-linear):

- Low-dimensional (4 features)
- Perfectly balanced classes
- Contains non-linear class boundaries
- RBF Kernel SVM shows clear advantage

3) Hyperparameter Sensitivity: Regularization Parameter C:

- Lower C (0.1-1.0) preferred for Breast Cancer, preventing overfitting
- Higher C (1.0-10.0) selected for Iris, allowing more flexible boundaries

RBF Gamma Parameter:

- 'scale' setting optimal for Breast Cancer
- Lower gamma (0.1) selected for Iris, providing smoother decision boundaries

D. Visualization Insights

1) Feature Distribution Analysis: The feature distribution histograms for the Breast Cancer dataset reveal:

- Clear separation between benign and malignant classes across most features
- Some features show stronger discriminative power than others
- Gaussian-like distributions validating the applicability of linear classifiers

2) Scatter Plot Analysis:

- Breast Cancer: Good linear separability in 2D projections
- Iris: Setosa clearly separated, Versicolor-Virginica overlap visible

V. Discussion

A. Key Findings

1) Dataset Linearity Matters: Linear models excel on the Breast Cancer dataset (linearly separable), while RBF Kernel SVM shows advantage on Iris (non-linear boundaries).
2) Cross-Validation Reliability: High CV scores (97-98%) provide confidence in model generalization, though some variance exists between CV and test performance.
3) Class Imbalance Handling: All algorithms handled the Breast Cancer's class imbalance well without requiring special techniques, likely due to the moderate imbalance ratio (2:1).
4) Multi-class Challenges: The Versicolor-Virginica confusion in Iris dataset highlights the importance of feature engineering and algorithm selection for overlapping classes.
5) Hyperparameter Importance: Grid search with cross-validation proved essential—default parameters would have yielded suboptimal results.

B. Algorithm Selection Guidelines

Choose Logistic Regression when:

- Data is linearly separable
- Probabilistic predictions are needed
- Model interpretability is important
- Training speed is critical

Choose Linear SVM when:

- Maximum margin classification is desired
- Data is high-dimensional
- Robustness to outliers is needed
- Linear separability is expected

Choose RBF Kernel SVM when:

- Non-linear patterns are present
- Highest accuracy is the priority
- Dataset size is moderate (kernel methods scale poorly)
- Computational resources are available

C. Experimental Methodology Strengths

1) Rigorous Hyperparameter Tuning: 5-fold CV with grid search ensures optimal configurations
2) Comprehensive Evaluation: Multiple metrics (accuracy, confusion matrix, ROC, precision/recall) provide complete picture
3) Standard Benchmarks: LIBSVM datasets enable comparison with existing literature
4) Proper Preprocessing: Feature standardization and stratified splitting follow best practices
5) Reproducibility: Fixed random seeds and clear documentation

VI. Conclusion

This experiment successfully compared three classification algorithms across binary and multi-class tasks, revealing important insights about algorithm selection and dataset characteristics.

A. Main Conclusions

1) Best Binary Classifier (Breast Cancer): Logistic Regression (96.10% accuracy, 97.49% CV score) proved optimal for linearly separable binary classification, offering simplicity and interpretability alongside excellent performance.
2) Best Multi-class Classifier (Iris): RBF Kernel SVM (95.56% accuracy) demonstrated clear advantage on non-linear multi-class data, justifying the computational overhead of kernel methods.
3) Most Versatile Algorithm: RBF Kernel SVM achieved highest average performance (95.59%) across both datasets, handling both linear and non-linear patterns effectively.
4) Linear Model Reliability: Linear methods (Logistic Regression, Linear SVM) performed competitively on linearly separable data, offering simpler alternatives when appropriate.

B. Practical Recommendations

1) Start with Linear Models: For initial baseline, use Logistic Regression or Linear SVM—they're fast, interpretable, and often sufficient.
2) Visualize First: Scatter plots and feature distributions help assess linearity before choosing algorithms.
3) Always Use Cross-Validation: 5-fold CV with grid search is essential for reliable hyperparameter selection.
4) Consider Trade-offs: Balance accuracy needs against computational constraints and interpretability requirements.
5) Evaluate Comprehensively: Use multiple metrics (accuracy, confusion matrix, ROC) rather than relying on a single measure.

C. Limitations and Future Work

Limitations:

- Limited to two relatively small datasets
- Did not explore advanced kernels (polynomial, sigmoid)
- No computational time analysis included
- Limited hyperparameter search ranges for efficiency

Future Work:

- Extend to larger, more complex real-world datasets
- Implement ensemble methods (Random Forest, Gradient Boosting)
- Conduct detailed learning curve analysis
- Explore feature selection and dimensionality reduction impact
- Compare with deep learning approaches
- Analyze computational efficiency systematically

D. Final Remarks

This experiment demonstrates that successful machine learning requires more than just applying algorithms—it demands understanding data characteristics, rigorous experimental methodology, and thoughtful algorithm selection. The Breast Cancer and Iris datasets, despite their modest size, provide valuable insights into classifier behavior that generalize to larger problems.

The key takeaway is that no single algorithm is universally best. Logistic Regression excelled on linear data, while RBF Kernel SVM proved superior for non-linear patterns. By following systematic evaluation procedures with proper cross-validation, data visualization, and multiple performance metrics, practitioners can make informed decisions that balance accuracy, interpretability, and computational efficiency.

Acknowledgments

We thank the LIBSVM repository maintainers for providing the benchmark datasets, and the scikit-learn development team for their excellent machine learning library.

References

[1] Chang, C. C., & Lin, C. J. (2011). LIBSVM: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST), 2(3), 1-27.

[2] Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825-2830.

[3] Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

[4] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer.

[5] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.