Support Vector Machines for Credit Default Prediction: A Comprehensive Study of Kernel Functions and Optimization Techniques

Anonymous Author(s)
Affiliation
Address
email

Abstract

This study presents an analysis of support vector machines (SVMs) applied to the UCI Default of Credit Card Clients dataset. The project explores the impact of different kernel functions and optimization techniques on classification performance. SVMs with linear, radial basis function (RBF), polynomial, and sigmoid kernels are evaluated. A grid search approach is used for hyperparameter optimization, and the decision boundary is visualized using Principal Component Analysis (PCA). The results are evaluated using accuracy, confusion matrices, and computational efficiency, demonstrating the utility of SVMs for binary classification tasks on tabular data.

1 Introduction

Support vector machines represent a powerful supervised machine learning approach that has demonstrated robust performance in a wide range of classification tasks. This project investigates the application of SVMs to the UCI Default of Credit Card Clients dataset, with a specific focus on assessing the impact of kernel functions and optimization techniques on model performance. The study also examines how hyperparameter tuning can enhance SVM efficiency and accuracy, while PCA is employed to visualize the decision boundary for improved interpretability. The primary objective of this study is to assess the efficacy of SVMs for predicting credit card default and to analyze the computational trade-offs associated with various kernel functions and optimization strategies.

2 Dataset

The UCI Default of Credit Card Clients dataset consists of demographic, financial, and historical payment information for 30,000 individuals. Features include demographic variables such as age, sex, education, and marital status, as well as financial metrics like credit limit, bill statements, and payment history. The target variable is binary, indicating whether an individual defaulted on their credit card payment in the subsequent month. Preprocessing steps involved handling missing data, downsampling to achieve a balanced dataset, and one-hot encoding categorical variables for compatibility with SVMs. The dataset was further divided into training and testing subsets, with continuous features standardized to have a mean of zero and a standard deviation of one.

3 Methodology

The study began with training a preliminary SVM using the radial basis function (RBF) kernel and default hyperparameters. The performance of the model was evaluated using confusion matrices and

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

accuracy scores. Subsequently, a hyperparameter tuning process was conducted using grid search cross-validation, optimizing the regularization parameter C and kernel coefficient γ. To facilitate visualization, PCA was employed to reduce the feature space to two dimensions, allowing for the approximation of the decision boundary. The grid search explored a range of values for C and γ, and the best-performing parameters were used to train the final model.

4 Results

The performance of the SVM was evaluated using several metrics, including accuracy and confusion matrices. The initial model with default hyperparameters achieved an accuracy of approximately 91.44%. Hyperparameter tuning through grid search identified optimal values of $C = 1$ and $\gamma = 0.01$. The optimized model demonstrated improved classification performance, as evidenced by a more balanced confusion matrix. PCA was used to visualize the decision boundary, revealing the separation of the two classes in the reduced feature space. While the PCA-transformed data provided valuable insights, it was noted that the approximation was limited due to the relatively low explained variance captured by the first two principal components.

5 Discussion

The results indicate that support vector machines are effective for binary classification tasks, particularly when applied to tabular data such as the credit card default dataset. The choice of kernel function significantly influenced model performance, with the RBF kernel demonstrating superior accuracy compared to linear, polynomial, and sigmoid kernels. Hyperparameter optimization through grid search further enhanced the model's classification ability, although the improvements were incremental. The use of PCA for visualization highlighted the potential for dimensionality reduction to aid in interpretability, though it also underscored the limitations of such approximations in capturing the full complexity of the data.

6 Conclusion

This study demonstrates the utility of SVMs for credit card default prediction, highlighting the importance of kernel selection and hyperparameter tuning. The findings suggest that SVMs with RBF kernels, coupled with grid search optimization, offer a robust solution for binary classification problems in financial datasets. Future work could explore the integration of other dimensionality reduction techniques, the application of SVMs to larger datasets, and the evaluation of additional kernel functions to further enhance model performance and interpretability.

References

- Chang, C. and Lin, C. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3), 27.
- UCI Machine Learning Repository: Default of Credit Card Clients Dataset. Available at `https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients`.
- Cortes, C., and Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 20(3), 273–297.

Acknowledgments

This research was conducted as part of the Advanced Machine Learning course, with guidance and resources provided by the Department of Computer Science. The UCI repository is acknowledged for providing the dataset used in this study.