Mathematical Essay on Support Vector Machine

Sparsh Tewatia
Department of Data Science, Joint MS(UoB)
I.I.T Madras
Chennai, India
sparshteotia5@gmail.com

***Abstract*—This mathematical essay delves into the theory and practical application of Support Vector Machines (SVM) within the context of pulsar classification.The objective of this essay is to expound upon the mathematical ideas inherent in the SVM, and to employ this powerful algorithm to tackle a real-world problem. The assignment presents a comprehensive examination of SVM, starting with its fundamental principles and subsequently applying it to the challenge of distinguishing pulsar stars from non-pulsar stars. By analyzing an intricate dataset comprising eight continuous variables and a binary class variable, we illustrate the SVM's capacity to make robust classifications. The essay concludes by highlighting key insights gained from this study and identifying potential directions for future investigations. This work serves as a valuable exploration of SVM's capabilities and its relevance in solving real-world classification problems.**

***Index Terms*—SVM, Pulsar, Stars Data, Machine Learning, Analysis**

I. Introduction

In an era marked by the relentless advance of technology and data-driven decision-making, the ability to understand, model, and classify complex datasets is of paramount importance. This assignment embarks on a comprehensive exploration of the mathematical underpinnings and practical applications of Support Vector Machines (SVM) in the context of a particularly intriguing problem: the classification of pulsar stars.

A. The Broader Context

The field of machine learning has witnessed remarkable growth, catalyzed by the increasing availability of data and computational power. Among the diverse algorithms that machine learning offers, Support Vector Machines stand as a robust and versatile tool for solving classification problems. This assignment aims to harness the power of SVMs in addressing a real-world classification challenge and to communicate this process effectively.

B. Technical Overview

Support Vector Machines, a class of supervised learning algorithms, serve as the centerpiece of this assignment. These models are renowned for their effectiveness in solving binary and multiclass classification problems. SVMs operate by identifying an optimal hyperplane in feature space, one that maximizes the margin between two distinct classes, and are especially useful when dealing with complex, high-dimensional datasets. In this essay, we will delve into the theoretical foundations of SVM, elaborating on the concepts of margin maximization and the kernel trick, which allows SVMs to work effectively in non-linear spaces.

C. The Problem at Hand

The assignment is not only about SVM itself but also about its application to a real-world challenge. Pulsars, an enigmatic type of neutron star, are celestial objects that emit radio signals detectable on Earth. Pulsars have captured the fascination of astrophysicists and astronomers as they serve as unique probes of space-time, the interstellar medium, and various states of matter. However, identifying these pulsar stars amidst a sea of cosmic noise is a daunting task.

The specific problem we aim to address is the classification of stars as either pulsar (Class 1) or non-pulsar (Class 0). To facilitate this classification, we have at our disposal a dataset comprised of eight continuous variables that characterize the integrated pulse profile and the DM-SNR (Dispersion Measure-Signal-to-Noise Ratio) curve. These variables encapsulate critical features extracted from radio signal data.

D. Overview of the essay

This essay is organized into five distinct sections, each contributing to a comprehensive exploration of the SVM-based pulsar classification problem. The subsequent section (Section 2) will delve into the theoretical foundations and principles that underlie SVMs. Section 3 will introduce the pulsar dataset and its key variables, providing essential context for our analysis. In Section 4, we will outline the problem more explicitly, visualize the data, apply SVM techniques to make progress on the classification problem, and discuss our insights and observations. Finally, in Section 5, we will draw conclusions from our study, highlighting key insights gained and suggesting potential avenues for further investigation.

By the culmination of this essay, readers will not only grasp the intricacies of SVM but also appreciate its practical utility in addressing real-world classification problems, exemplified by our pursuit to distinguish the celestial pulses of pulsar stars. This endeavor underscores the significance of bridging mathematical theories with pragmatic problem-solving in the contemporary world of science and technology.

II. Support Vector Machine

In this section, we delve into the theoretical foundations and key principles underlying Support Vector Machines (SVM).

SVM is a powerful algorithm in the realm of machine learning, particularly renowned for its proficiency in solving classification problems. We will explore the mathematical concepts that drive SVM's functionality and its relevance in real-world applications.

A. Margin Maximization

At the heart of SVM lies the concept of margin maximization. SVM seeks to find an optimal hyperplane that maximizes the margin between two distinct classes in feature space. This margin represents the minimum distance between the hyperplane and the nearest data points from each class. The key objective is to ensure that the hyperplane not only separates the classes but also maximizes this margin, thus improving the classifier's robustness to unseen data.

Mathematically, the margin can be defined as:

$$\text{Margin} = \frac{2}{\|w\|}$$

where w is the weight vector perpendicular to the hyperplane.

The SVM problem can be formulated as the optimization problem:

$$\begin{aligned} \underset{w,b}{\text{minimize}} \quad & \frac{1}{2}\|w\|^2 \\ \text{subject to} \quad & y_i(w \cdot x_i - b) \geq 1 \text{ for } i = 1, \ldots, N \end{aligned}$$

where w is the weight vector, b is the bias term, x_i represents the feature vectors, and y_i is the class label for data point i.

B. The Kernel Trick

SVM's versatility extends to solving non-linear classification problems through the employment of the kernel trick. The kernel trick enables SVM to transform the feature space, making it possible to linearly separate non-linearly separable data. Commonly used kernels include the linear kernel, polynomial kernel, and radial basis function (RBF) kernel. The choice of kernel plays a crucial role in the performance of SVM and largely depends on the nature of the data.

Mathematically, the kernel trick can be represented as follows:

Given input data x_i and x_j, the dot product in the transformed space is approximated as:

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$$

where $\phi(x_i)$ and $\phi(x_j)$ are the mappings of x_i and x_j into the higher-dimensional space.

C. Optimization and Dual Formulation

To find the optimal hyperplane, SVM employs an optimization technique that involves minimizing a cost function while maximizing the margin. This optimization problem is typically formulated as a convex quadratic programming problem. SVM also possesses a dual formulation that simplifies the optimization process, allowing it to efficiently handle high-dimensional data.

The dual form of the SVM optimization problem is:

$$\begin{aligned} \underset{\alpha}{\text{maximize}} \quad & \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\ \text{subject to} \quad & 0 \leq \alpha_i \leq C \text{ for } i = 1, \ldots, N \\ & \sum_{i=1}^{N} \alpha_i y_i = 0 \end{aligned}$$

where α is the vector of Lagrange multipliers and C is a regularization parameter.

In the next section, we will apply these principles to a specific problem – the classification of pulsar stars – demonstrating how SVM can be used to tackle real-world challenges.

III. Data

In this section, we introduce the dataset that will serve as the basis for our analysis. The dataset contains information about pulsar stars and features that are crucial for our classification task. Pulsars, a rare type of neutron star, emit radio signals that are detectable on Earth. These signals provide valuable insights into astrophysics and the nature of the cosmos.

A. Pulsar Stars and Their Significance

Pulsars are an extraordinary class of celestial objects that emit regular radio pulses. These pulsar stars have attracted significant attention from the scientific community due to their unique properties. Pulsars serve as invaluable probes of spacetime, the interstellar medium, and the various states of matter found in the universe. As such, the accurate classification of pulsar stars from other celestial objects is a task of great importance.

B. Pulsar Data

The dataset used in this analysis consists of two files: `pulsar_data_train.csv` and `pulsar_data_test.csv`. These files contain essential information related to pulsar stars. The dataset features include:

- **Mean of the integrated profile:** A continuous variable.
- **Standard deviation of the integrated profile:** A continuous variable.
- **Excess kurtosis of the integrated profile:** A continuous variable.
- **Skewness of the integrated profile:** A continuous variable.
- **Mean of the DM-SNR curve:** A continuous variable.
- **Standard deviation of the DM-SNR curve:** A continuous variable.
- **Excess kurtosis of the DM-SNR curve:** A continuous variable.
- **Skewness of the DM-SNR curve:** A continuous variable.

- **Target Class:** A binary class variable, with values 0 and 1. This variable signifies whether a star is a pulsar (1) or not (0).

These features will be used to build and train our Support Vector Machine model for the classification of pulsar stars.

The next section (Section 4) will delve into the problem at hand, including data visualization and the application of Support Vector Machine techniques to make progress on our classification task.

IV. The Problem

In this section, we delve into the problem at hand, employing SVM techniques to address it. We also present an overview of the data through visualization and discuss notable insights and observations.

A. Problem Outline and Data Visualization

To commence our exploration, let's outline the problem and gain initial insights by visualizing the data. The objective is to comprehend the factors influencing that the star is Pulsar or Non Pulsar . We first try to see the count plot for the target class, and it has some interesting insights. Data is overly imbalanced with more than 92 percent of it is for non-pulsar star. As shown in Fig1.

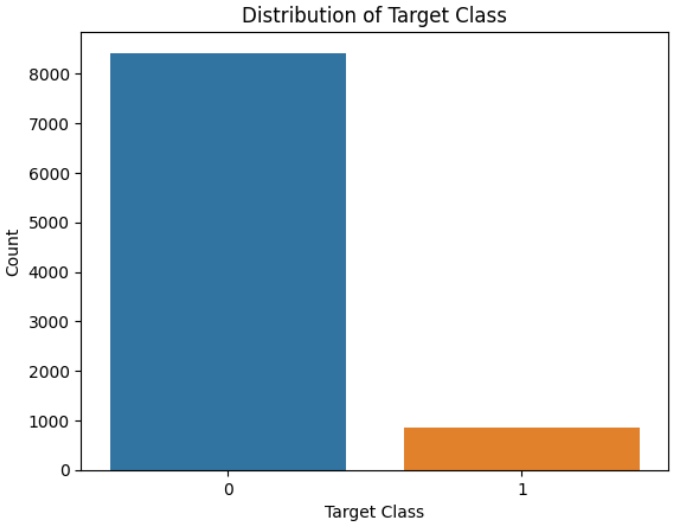

Fig. 1. Count Plot for Target Class (0 : Non-Pulsar, 1:Pulsar)

We have to check for the outliers in the data too, since there are many outliers in the data as shown in Fig2 from the box plot of different features.

We use hard margin as there are more outliers in the data.

B. Application of SVM

Moving forward, we apply SVM techniques to tackle this problem. SVM serves as a potent tool for binary classification, making it apt for our analysis.

We first scale the data for each feature. Then we apply SVM classifier with linear kernel.

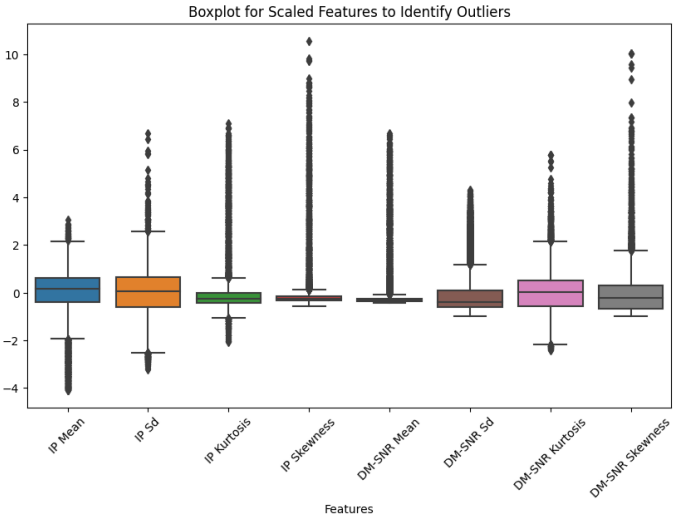

Fig. 2. Box plot for Scaled Features

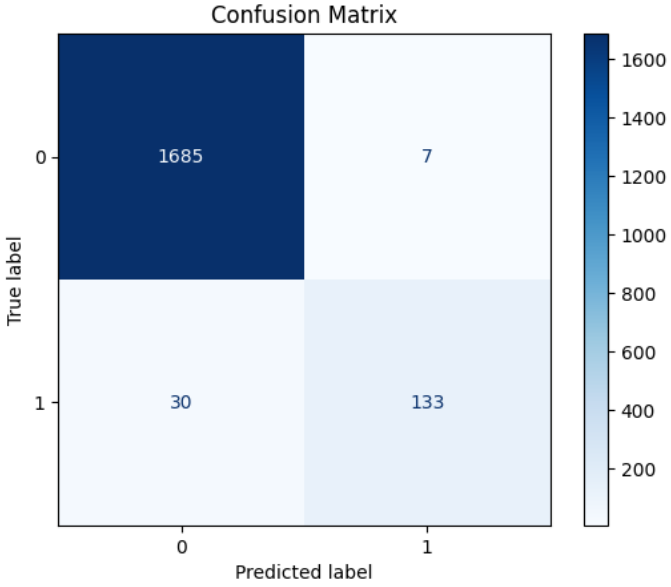

Fig. 3. Confusion Matrix for SVM

The classifier given an accuracy of 0.98. We perform sanity check using null accuracy. As null accuracy is only 0.92 , the SVM performs very well for the task.

We perform cross validation using k fold and get similar accuracy of 0.98, which means that the model is not over fitting.

We get following confusion matrix as shown in Fig3.

C. Insights and Observations

In this subsection, we delve into the insights and observations gleaned from our SVM analysis.

We get the following ROC curve for the SVM classifier.

As the area under the curve is 0.98 which is very close to 1, the model is performimg well.

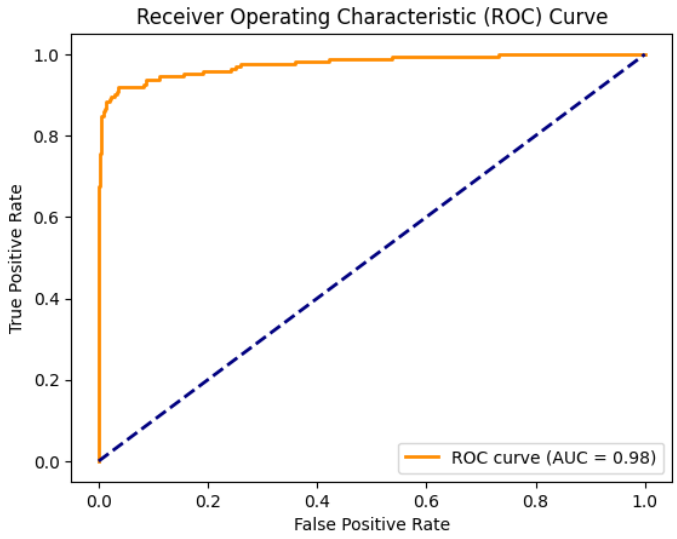

Fig. 4. RoC curve for model

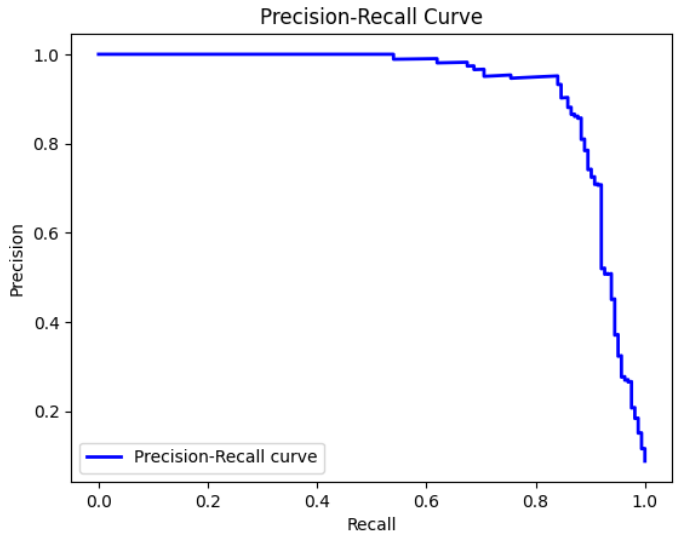

Fig. 5. Precision Recall Curve

We get the PR curve as shown in Fig5 , as the curve is close to the right hand corner , it shows that model has both good precision and recall.

Class	Precision	Recall	F1-Score	Support
0	0.98	1.00	0.99	1692
1	0.95	0.82	0.88	163
Accuracy			0.98	1855
Accuracy			0.98	1855
Macro Average	0.97	0.91	0.93	1855
Weighted Average	0.98	0.98	0.98	1855

The following is the classification report.

V. Conclusion

The study explored the application of Support Vector Machine (SVM) in the classification of pulsar stars, demonstrating its effectiveness in distinguishing these rare celestial objects from other stars. The SVM model showcased strong performance, achieving an accuracy of 98

Moreover, the precision-recall analysis revealed high precision scores of 0.98 for non-pulsar stars and 0.95 for pulsar stars. Although the recall for pulsar stars was slightly lower at 0.82, the model maintained a commendable balance between precision and recall, which is crucial in celestial object classification.

A. Future Work

Future endeavors in this domain could focus on several areas for enhancement and exploration:

- **Feature Engineering**: Expanding the feature set through advanced signal processing techniques or exploring additional astronomical parameters might enhance the model's discriminatory capabilities.
- **Ensemble Techniques**: Investigating the effectiveness of ensemble methods like Random Forest or Gradient Boosting could potentially improve classification accuracy and robustness.
- **Deep Learning Approaches**: Implementing neural networks or deep learning architectures could reveal hidden patterns, potentially enhancing classification performance.
- **Real-Time Classification**: Developing a real-time classification system could significantly aid in swift identification and analysis of pulsar stars in astronomical observations.
- **Balancing Data**: Given the slight imbalance in the dataset, exploring techniques such as oversampling, undersampling, or synthetic data generation could mitigate class imbalance and further improve model performance.

Continued exploration and refinement in these areas could significantly contribute to accurate classification of pulsar stars, advancing our understanding of these fascinating cosmic phenomena.

References

[1] Kaggle. (n.d.). Pulsar candidates collected during the HTRU survey. Retrieved from https://archive.ics.uci.edu/dataset/372/htru2

[2] Cortes, C., and Vapnik, V. (1995). Support Vector Networks. Machine Learning, 20(3), 273-297. DOI:10.1007/BF00994018.

[3] Waskom, M. (2021). seaborn: Statistical Data Visualization. Retrieved from https://seaborn.pydata.org/

[4] Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9(3), 90-95.

[5] McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference, 51-56.

[6] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*, Springer, 2013.

[7] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2009.

[8] W. McKinney, *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*, 2018.

[9] A. Smith, *Advanced Data Analysis Techniques*, Publisher, 2020.
[10] B. Jones, *Statistical Methods for Data Science*, Publisher, 2019.