Assignment 6: A mathematical essay on Support Vector Machines

S, Karthik
ME18B149 - IDDD Data Science
Indian Institute of Technology Madras
Chennai, India
me18b149@smail.iitm.ac.in

***Abstract*—This document is an overview of the concept of Support Vector Machines. It gives a general idea of the topic, along with the mathematical aspects, as well as an application involving it based on data from a real-world problem.**

***Index Terms*—Classification, Support Vector Machines, Machine Learning, Data Science**

I. Introduction

Support Vector Machine, abbreviated as SVM can be used for both regression and classification tasks. But, it is widely used in classification objectives. Support vector machine is highly preferred by many as it produces significant accuracy with less computation power.

The specific problem discussed is a classification problem which determines whether the observed star is a pulsar star or not based on different features such as integrated profile and the DM-SNR curve of the observation. The model will be trained using the HTRU2 Dataset. Support Vector Machines will be used to analyze the data and build a model which will then be able to predict the same for unknown data. Various insights and conclusions will be made based on the trends followed by the dataset.

This paper is a case study through which the principles of Support Vector Machines are implemented. It aims to examine which parameters influence whether the star is a Pulsar star or not.

II. Support Vector Machines

A. Model Structure

Support Vector Machine (SVM) is a supervised machine learning algorithm that can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well (Fig 1).

B. Max Margin condition

Margin of a Linear Classifier is defined as the perpendicular distance between the separating hyperplane and the closest data point To separate the two classes of data points, there are many possible hyperplanes that could be chosen. Our objective is to find a plane that has the maximum margin, i.e the maximum distance between data points of both classes (Fig 2). Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence. Max margin not only ensures good classification but also ensures good generalisation.

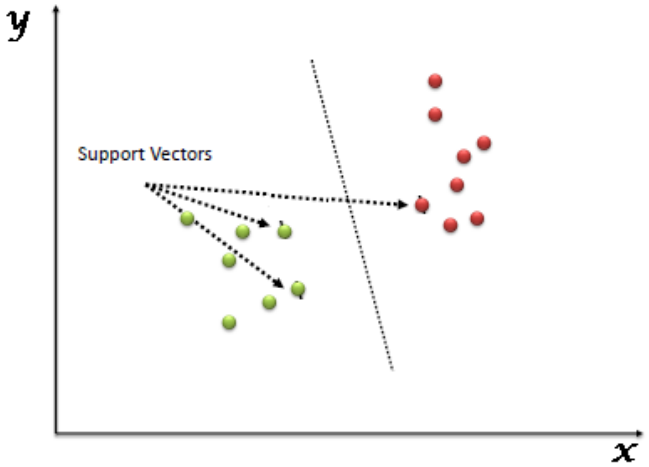

Fig. 1. General structure of a Support Vector Machine

C. Support vectors

Support vectors are data points that are closer to the hyperplane and influence the position and orientation of the hyperplane (Fig 1). Using these support vectors, we maximize the margin of the classifier. Deleting the support vectors will change the position of the hyper-plane. These are the points that help us build our SVM.

D. Cost Function - Separable Case

The equation of separating hyper-plane is given by

$$\boldsymbol{\theta^T x} + \boldsymbol{\theta_0} = 0 \tag{1}$$

Assuming that 2 support vectors $\boldsymbol{x_a}$ and $\boldsymbol{x_b}$ are at a distance of 1 unit from the hyper-plane on either side,

$$\boldsymbol{\theta^T x_a} + \boldsymbol{\theta_0} = 1$$

$$\boldsymbol{\theta^T x_b} + \boldsymbol{\theta_0} = -1$$

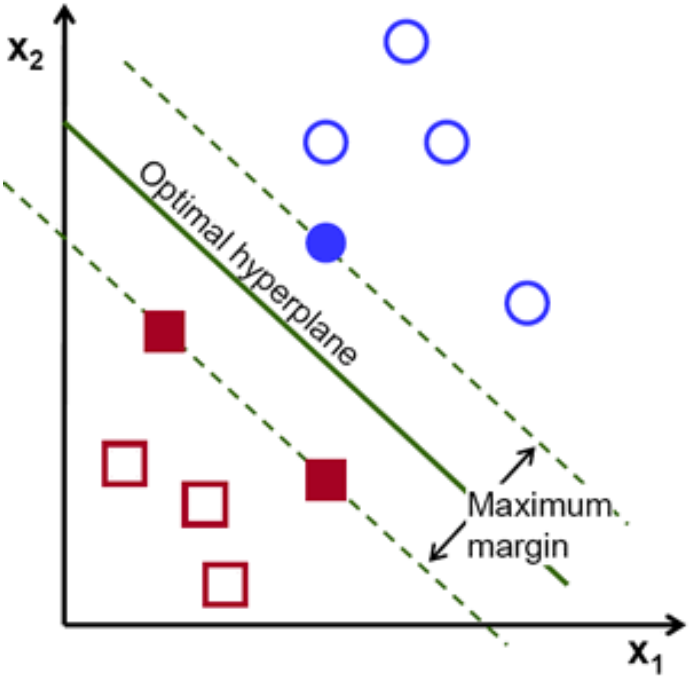

Fig. 2. An example of Max-Margin classifier in 2 dimensions

$$Margin = ||\boldsymbol{x_a} - \boldsymbol{x_b}||_2 = \frac{2}{||\boldsymbol{\theta}||_2} \tag{2}$$

So, the optimization problem can be defined as

$$\max_{\boldsymbol{\theta},\boldsymbol{\theta_0}} \frac{2}{||\boldsymbol{\theta}||_2} \tag{3}$$

such that

$$\boldsymbol{\theta^T x} + \boldsymbol{\theta_0} > 1 \text{ if } y_i = 1$$
$$\boldsymbol{\theta^T x} + \boldsymbol{\theta_0} < 1 \text{ if } y_i = -1$$

Generally, the dual of this problem(minimization) is solved. The solution is given by

$$\boldsymbol{\theta} = \sum_{\forall i} \alpha_i y_i \boldsymbol{x_i} \tag{4}$$

$$\boldsymbol{\theta_0} = y_k - \boldsymbol{\theta^T x_k} \tag{5}$$

where α_i are the Lagrange multipliers which are non-zero for Support vectors, k is Any support vector

E. Cost Function - Non-separable case

In this case, some mis-classification is allowed. The error is measured by introducing slack variables into the optimization cost function. A tunable parameter is also introduced to control the misclassification error to be allowed.

$$\min_{\boldsymbol{\theta},\boldsymbol{\theta_0}} \frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\theta} + C\sum_{\forall i}\xi_i \tag{6}$$

such that

$$y_i(\boldsymbol{\theta^T x_i} + \boldsymbol{\theta_0}) > 1 - \xi_i$$
$$\xi_i > 0$$

III. The Problem

A. Overview and objectives

Pulsars are a rare type of Neutron star that produces radio emissions detectable here on Earth. They are of considerable scientific interest as probes of space-time, the inter-stellar medium, and states of matter. Machine learning tools are now being used to automatically label pulsar candidates to facilitate rapid analysis. The key task is to Predict if a star is a pulsar start or not. Each candidate is described by 8 continuous variables and a single class variable. The first four are simple statistics obtained from the integrated pulse profile (folded profile). This is an array of continuous variables that describe a longitude-resolved version of the signal that has been averaged in both time and frequency. The remaining four variables are similarly obtained from the DM-SNR curve.

B. Reading the data

We will be using python throughout this assignment. The dataset is contained in pulsar_data_train.csv and pulsar_data_test.csv are provided which contains the details of various parameters of come stars which are shown in detail in Fig 3. Since we are creating a model, we would be splitting the training dataset into train and validation sets with a fraction of 10% for validation.

IV. Data cleaning

After reading the data, the next task will be to filter out the attributes or entries that are not required or redundant in this analysis. There are no unfilled values in this dataset which means no entry is being eliminated. Also, all the features are different enough and no feature is redundant in this case which means the dataset will be used completely as it is without any alterations.

V. Exploratory Data Analysis

The complete details of all the attributes are shown in Fig 3.

Variable	Definition	Key
Mean of the integrated profile	-	Continuous
Standard deviation of the integrated profile	-	Continuous
Excess kurtosis of the integrated profile	-	Continuous
Skewness of the integrated profile	-	Continuous
Mean of the DM-SNR curve	-	Continuous
Standard deviation of the DM-SNR curve	-	Continuous
Excess kurtosis of the DM-SNR curve	-	Continuous
Skewness of the DM-SNR curve	-	Continuous
target_class	Class	0, 1

Fig. 3. Exhaustive list of all the attributes in the given dataset

Our first task is to analyze the data and form visible conclusions and trends from the same with the help of plots.

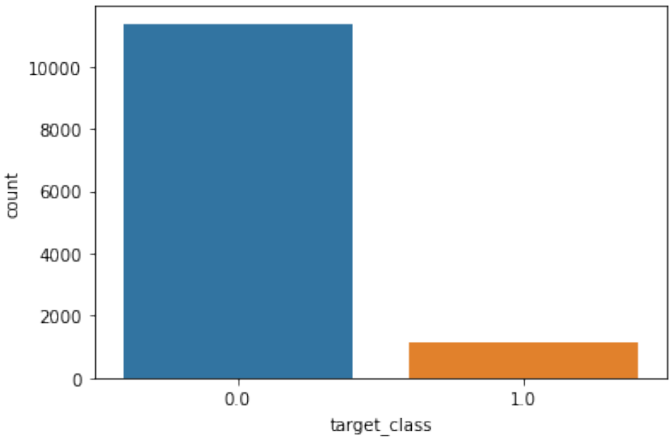

Fig. 4. Bar chart - Distribution among outcomes

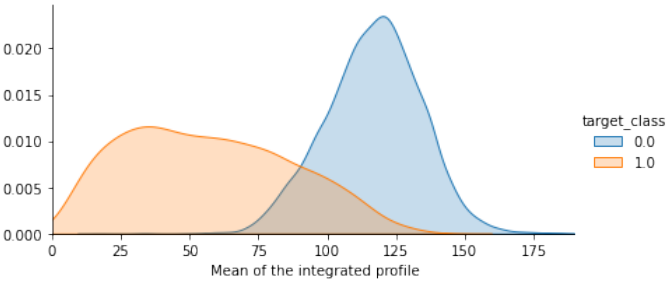

Fig. 5. Distribution of the mean of the integrated profile

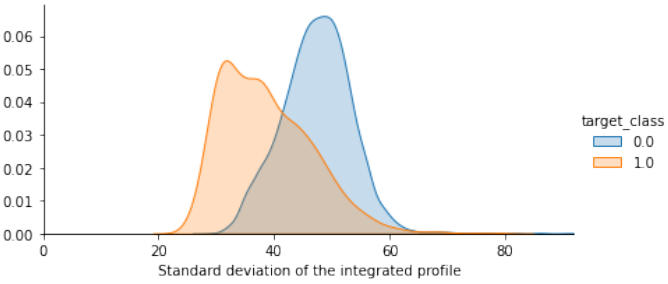

Fig. 6. Distribution of the stddev of the integrated profile

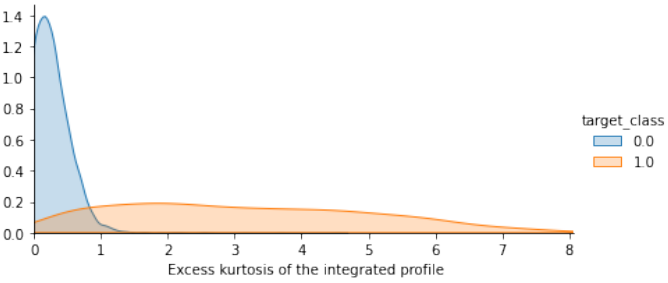

Fig. 7. Distribution of the Excess kurtosis of the integrated profile

From Fig 4, we can see that the dataset is a highly imbalanced and a skewed one as the number of non-pulsar stars are much higher than the pulsar stars.

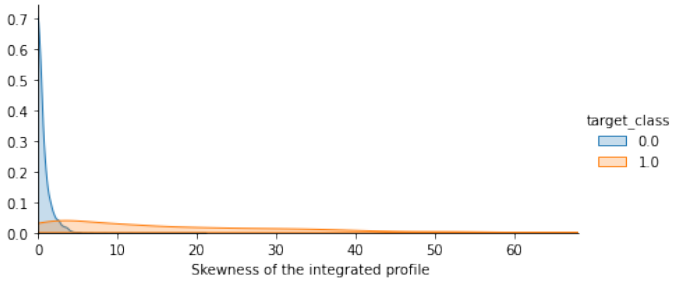

Fig. 8. Distribution of the skewness of the integrated profile

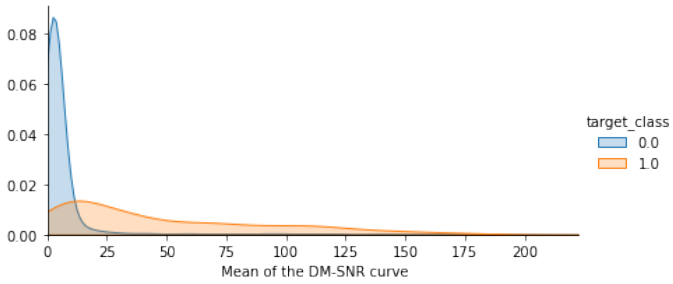

Fig. 9. Distribution of the mean of the DM-SNR curve

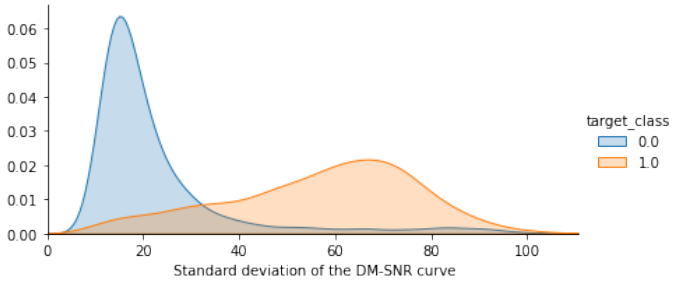

Fig. 10. Distribution of the stddev of the DM-SNR curve

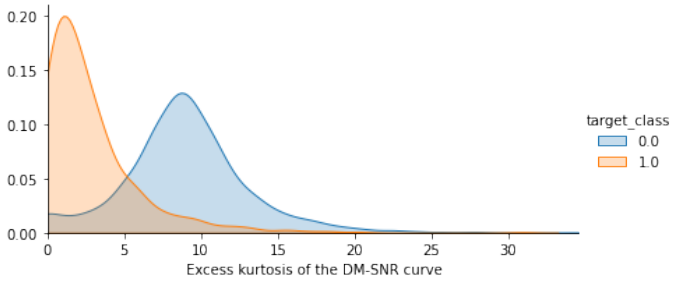

Fig. 11. Distribution of the Excess kurtosis of the DM-SNR curve

Fig 5 to Fig 12 shows the distribution of the features based on the outcomes individually. Some of the parameters from which we can clearly distinguish whether the star is a pulsar star or not are:

- Mean of the integrated profile (Fig 5) is much lower in a Pulsar star

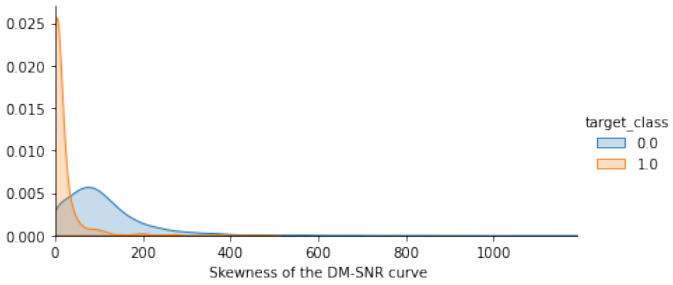

Fig. 12. Distribution of the skewness of the DM-SNR curve

- Standard deviation of the integrated profile (Fig 6) is lower in a pulsar star
- Standard deviation of the DM-SNR (Fig 10) is higher in a pulsar star
- Excess Kurtosis of the DM-SNR (Fig 11) is lower in a pulsar star

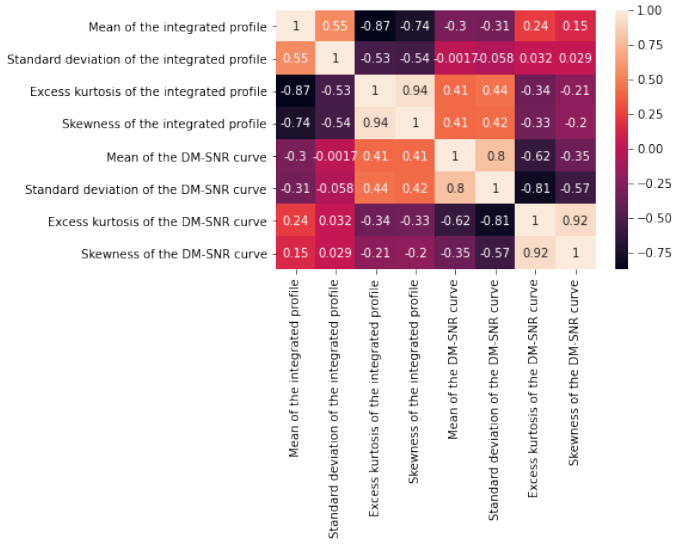

Fig. 13. Correlation matrix

The correlation matrix is plotted in Fig 13. From this, we can conclude that excess kurtosis and skewness are highly positively correlated. Similarly, excess kurtosis and the mean of the integrated profile, excess kurtosis and the standard deviation of the DM-SNR curve are highly negatively correlated.

VI. Data preparation

The given dataset has some missing data as shown in Table I. All the variables are continuous. Hence, the missing variables can be imputed with the median of the distribution.

VII. Model preparation, training and validation

Since the datasets are now model compatible, the next step is to create the model. The training dataset was split into training and validation datasets with a fraction of 10% kept for validation. Validation accuracy, i.e, the fraction of entries for which the predicted and the actual value are the same will be used as the metric to compare between models. 3 parameters (C, gamma and the Kernel type) were tuned for a range a grid search was performed with all these permutations (Table II). From the gridsearch, the best model was determined to be Table III.

TABLE I
Missing values in the dataset

Parameter	Train Dataset	Test Dataset
Mean of the integrated profile	0	0
Standard deviation of the integrated profile	0	0
Excess kurtosis of the integrated profile	1735	767
Skewness of the integrated profile	0	0
Mean of the DM-SNR curve	0	0
Standard deviation of the DM-SNR curve	1178	524
Excess kurtosis of the DM-SNR curve	0	0
Skewness of the DM-SNR curve	625	244
target_class	0	5370

TABLE II
Hyperparameter tuning sample space

S. No.	Parameter	Values
1	C	0.1, 1, 10, 100, 1000
2	gamma	1, 0.1, 0.01, 0.001
3	kernel	Linear, rbf

TABLE III
Best model hyperparameters

S. No.	Parameter	Value
1	C	100
2	gamma	0.1
3	kernel	rbf

A. Final Model

Using the above parameters, an SVM was created and the trained using the training data. The validation accuracy is determined to be **98.56%**.

Also, the confusion matrix which is the plot between the actual outcome and the predicted outcome is plotted for the model and is shown in Fig 14. Only 18 entries incorrectly predicted from the validation dataset which shows that this is a highly accurate model.

VIII. Prediction of unknown data

Using the above model, we can predict whether the observed star is a pulsar star or not for unknown observations. The dataset pulsar_data_test.csv was run through the model and predictions were successfully done. The classification of outcomes is shown in Fig 15.

One observation is that the proportion of pulsar stars are very less in both the training and the test data. This can be due to 2 reasons:

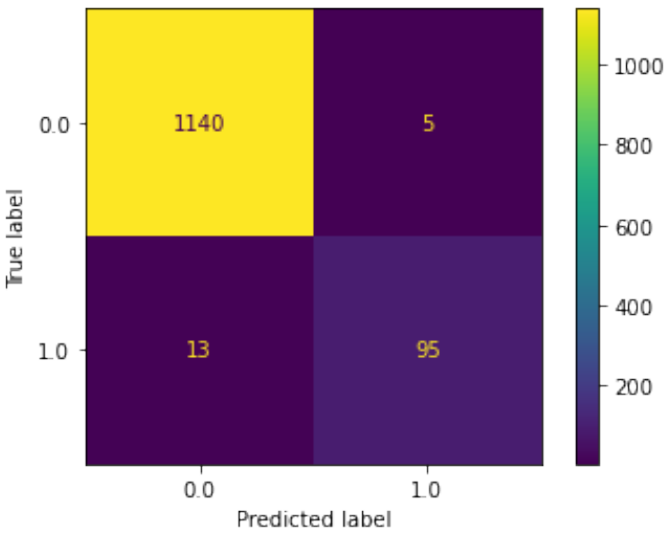

Fig. 14. Confusion matrix - validation data

- Pulsar stars are rare. Hence, the observations are accurate to an extent
- The skewness and the bias in the training data unfairly influenced the test data.

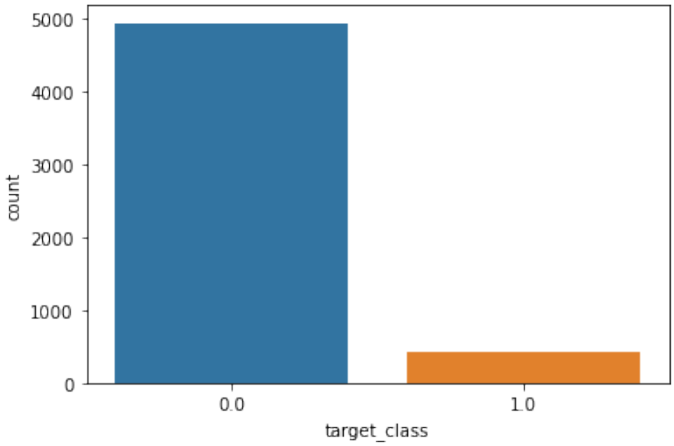

Fig. 15. Bar chart - test data

IX. Conclusion

The effects of various factors which could have affected whether an observed star is a pulsar star or not were analyzed in great detail. Various factors greatly affected the outcome like mean, standard deviation, excess kurtosis and the skewness of both integrated profile and the DM-SNR curve.

Various SVMs models were constructed using grid search on the basis of the known data to determine the right set of hyperparameters which with the best possible model was created with a validation accuracy of 98.56%. Using this model, the confusion matrix was plotted which gave the exact numbers predicted correctly and incorrectly.

This was also used to predict the outcome of a set of data for which it is unknown.

Further avenues could be explored using this data where they can be classified using various other classifiers available to study how the parameters might possibly affect each other and also how they affect the income. Also, implementing techniques more advanced than SVMs might give a model with better accuracy and predictions which would improve on the current model.

References

[1] Dr. Kaushik Mitra, Dr. Ramkrishna Pasumarthy, "Support Vector Machines", EE4708: Data Analytics Laboratory - Week 9
[2] Support Vector Machines — Detailed Overview, Towards Data Science.
[3] Support Vector Machines - https://scikit-learn.org/