Assignment-5: A Mathematical Essay on Support Vector Machine

Shania Mitra
Roll Number: CH18B067
Chemical Engineering
IIT Madras, Chennai
ch18b067@smail.iitm.ac.in

***Abstract*—In this study, we predict whether a star is a pulsar or not based on statistical properties of the integrated profile and DM-SNR curve. A support vector classifier is used to model the importance of these factors and predict the label of the star.**

***Index Terms*—Support Vector Classifier, Visualization, Pulsar prediction**

I. Introduction

Pulsars are a rare type of Neutron star that produce 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.

SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes.

In this study, we use Support Vector Machines to model the category of stars based on statistical properties of the integrated profile and DM-SNR curve.

We begin by gathering, cleaning and preparing the data, following which we perform exploratory analysis. Finally, we build statistical models and perform visualizations to provide quantitative and visual evidence of the relationships observed. In the next section, we highlight the key principles underlying Support Vector classifier. In section 3, we discuss the insight and observations drawn from the data and the models. Finally, in section 4 we outline the salient features of the study and present further avenues of possible investigation.

II. Support Vector

A support-vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks like outliers detection.Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin, the lower the generalization error of the classifier

We are given a training dataset of n points of the form

$$\left(\mathbf{x}_1, y_1\right), \ldots, \left(\mathbf{x}_n, y_n\right)$$

where the y_i are either 1 or -1, each indicating the class to which the point $\mathbf{x}_i$ belongs. Each $\mathbf{x}_i$ is a p dimensional real vector. We want to find the "maximum-margin hyperplane" that divides the group of points $\mathbf{x}_i$ for which $y_i = 1$ from the group of points for which $y_i = -1$, which is defined so that the distance between the hyperplane and the nearest point $\mathbf{x}_i$ from either group is maximized. Any hyperplane can be written as the set of points $\mathbf{x}$ satisfying

$$\mathbf{w}^T\mathbf{x} - b = 0$$

Geometrically, the distance between these two hyperplanes is $\frac{2}{\|\mathbf{w}\|}$[17] so to maximize the distance between the planes we want to minimize $\|\mathbf{w}\|$. The distance is computed using the distance from a point to a plane equation. We also have to prevent data points from falling into the margin, we add the following constraint: for each i either

$$\mathbf{w}^T\mathbf{x}_i - b \geq 1, \text{ if } y_i = 1$$

or

$$\mathbf{w}^T\mathbf{x}_i - b \leq -1, \text{ if } y_i = -1$$

These constraints state that each data point must lie on the correct side of the margin. This can be rewritten as

$$y_i\left(\mathbf{w}^T\mathbf{x}_i - b\right) \geq 1, \quad \text{for all } 1 \leq i \leq n$$

We can put this together to get the optimization problem: "Minimize $\|\mathbf{w}\|$ subject to $y_i\left(\mathbf{w}^T\mathbf{x}_i - b\right) \geq 1$ for $i = 1, \ldots, n$." The $\mathbf{w}$ and b that solve this problem determine our classifier, $\mathbf{x} \mapsto \operatorname{sgn}\left(\mathbf{w}^T\mathbf{x} - b\right)$ where $\operatorname{sgn}(\cdot)$ is the sign function. An important consequence of this geometric description is that the max-margin hyperplane is completely determined by those $\vec{x}_i$ that lie nearest to it. These $\mathbf{x}_i$ are called support vectors. Soft-margin [edit] To extend SVM to cases in which the data are not linearly separable, the hinge loss function is helpful

$$\max\left(0, 1 - y_i\left(\mathbf{w}^T\mathbf{x}_i - b\right)\right)$$

Note that y_i is the i-th target (i.e., in this case, 1 or -1), and $\mathbf{w}^T\mathbf{x}_i - b$ is the i-th output. This function is zero if the constraint in (1) is satisfied, in other words, if $\mathbf{x}_i$ lies on the correct side of the margin. For data on the wrong side of the margin, the function's value is proportional to the distance from the margin. The goal of the optimization then is to minimize

$$\lambda\|\mathbf{w}\|^2+\left[\frac{1}{n}\sum_{i=1}^{n}\max\left(0,1-y_i\left(\mathbf{w}^T\mathbf{x}_i-b\right)\right)\right]$$

where the parameter $\lambda > 0$ determines the trade-off between increasing the margin size and ensuring that the $\mathbf{x}_i$ lie on the correct side of the margin. Thus, for sufficiently small values of λ, it will behave similar to the hard-margin SVM, if the input data are linearly classifiable, but will still learn if a classification rule is viable or not. (This parameter λ is also called C, e.g. in LIBSVM.)

III. The Problem

In this study, we estimate the safety of a car based on factors such as buying price, maintenance price, capacity, etc.

A. Data Preparation

The training dataset used in this study consists of 12528 stars. 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.

- Mean of the integrated profile
- Standard deviation of the integrated profile
- Excess kurtosis of the integrated profile
- Skewness of the integrated profile
- Mean of the DM-SNR curve
- Standard deviation of the DM-SNR curve
- Excess kurtosis of the DM-SNR curve
- Skewness of the DM-SNR curve
- Target class - 0, 1

The aim is to predict the multiclass feature Target.

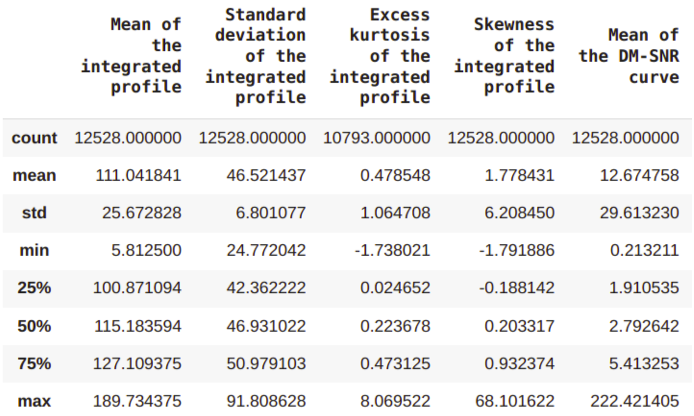
B. Exploratory Analysis

In this section we look at the trends between various features.

In Table I we have a description of the features and the count, min, mean, median etc. of the first few features.

In figure 1, we plot the distributions of a features based on target class. In each of these, we observe that the distributions of "not pulsar" and "pulsar" have different means and variances. The pulsar stars tend to have a higher variance and smaller peak.

In figure 2 we plot scatter plots of various properties of the DM-SNR and Integrated profile coloured by the class of the star. We can clearly see that the classes form distinct clusters in the plot that can be distinguished from one another. Further, we can see a clear trend in the skew vs kurtosis graph [Fig. 2c]. Non-pulsars tend to have low or negative kurtosis and a small skew relatively independent of eachother while pulsars tend to have high skew and kurtosis, approximately linearly dependent on eachother. Further, in Fig 2.c, d, we can clearly see that there is a linear line of separation.



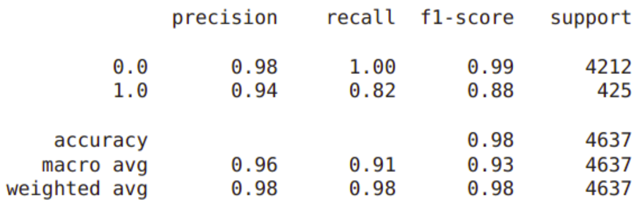
	Mean of the integrated profile	Standard deviation of the integrated profile	Excess kurtosis of the integrated profile	Skewness of the integrated profile	Mean of the DM-SNR curve
count	12528.000000	12528.000000	10793.000000	12528.000000	12528.000000
mean	111.041841	46.521437	0.478548	1.778431	12.674758
std	25.672828	6.801077	1.064708	6.208450	29.613230
min	5.812500	24.772042	-1.738021	-1.791886	0.213211
25%	100.871094	42.362222	0.024652	-0.188142	1.910535
50%	115.183594	46.931022	0.223678	0.203317	2.792642
75%	127.109375	50.979103	0.473125	0.932374	5.413253
max	189.734375	91.808628	8.069522	68.101622	222.421405

TABLE I
Description of each feature

In figure 3, we plot the heatmap of correlation among the features. It can be observed that the target class is highly positively correlated to the skew and kurtosis of the integrated profile and negatively correlated to the mean and standard deviation of the integrated profile. It also exhibits a similar trend with the properties from the DM-SNR curve, but the correlations are weaker.

C. Support Vector Machine

To train the model and gauge its performance, the dataset was split into train and test sets with an 80-20 split. An SVM Model with a linear kernel is used since from the visualizations in section 3.2, we observed that the classes are easily separable. Following are the metrics achieved by the SVM model:



```
              precision    recall  f1-score   support

         0.0       0.98      1.00      0.99      4212
         1.0       0.94      0.82      0.88       425

    accuracy                           0.98      4637
   macro avg       0.96      0.91      0.93      4637
weighted avg       0.98      0.98      0.98      4637
```

In Figure 4, we can see the confusion matrix of the predictions by the model on the validation dataset. We can see that there are only 96 misclassifications among the 4637 stars, with the model confusing some pulsars to be non-pulsars due to the imbalance in the dataset.

IV. Conclusions

In this study, we observe the factors that decide whether a star is a pulsar or not. We observed that in case of the integrated profile as well as the DM-SNR curve, pulsars tend to have higher skew and kurtosis and lower mean and standard deviation. In future, class imbalance handling techniques could be implemented.

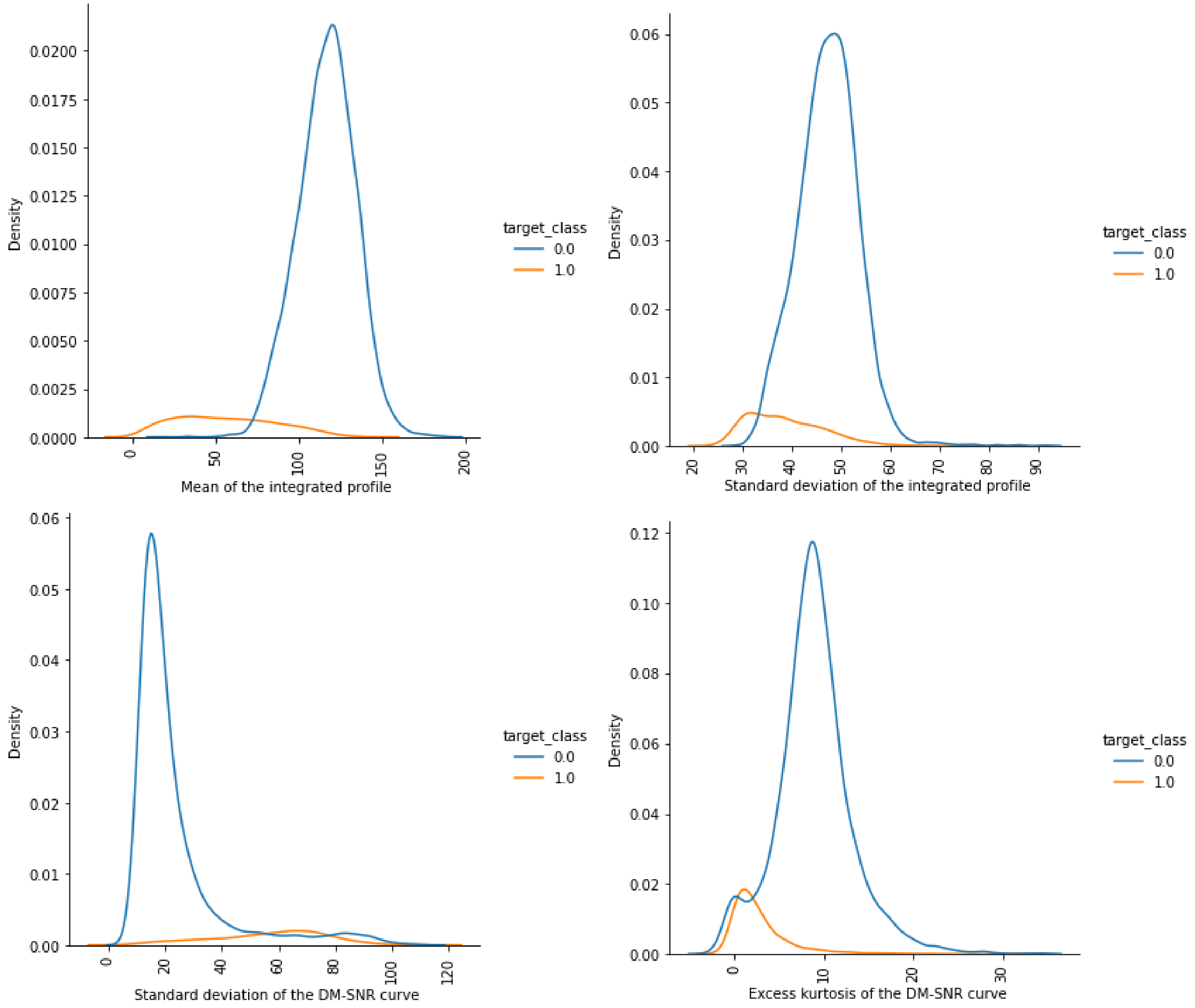


Fig. 1. Distribution of (a) Mean of integrated profile (b) Std of integrated profile (c) Std of DM-SNR (d) Excess kurtosis of DM-SNR for pulsars (orange) and other stars (blue) separately.

REFERENCES

[1] Wikipedia contributors. “Support-Vector Machine.” Wikipedia, 24 Nov. 2021, en.wikipedia.org/wiki/Support-vector_machine.

[2] Gandhi, Rohith. “Support Vector Machine — Introduction to Machine Learning Algorithms.” Medium, 5 July 2018, towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47.

[3] “Sklearn SVM (Support Vector Machines) with Python.” DataCamp Community, www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python. Accessed 28 Nov. 2021.

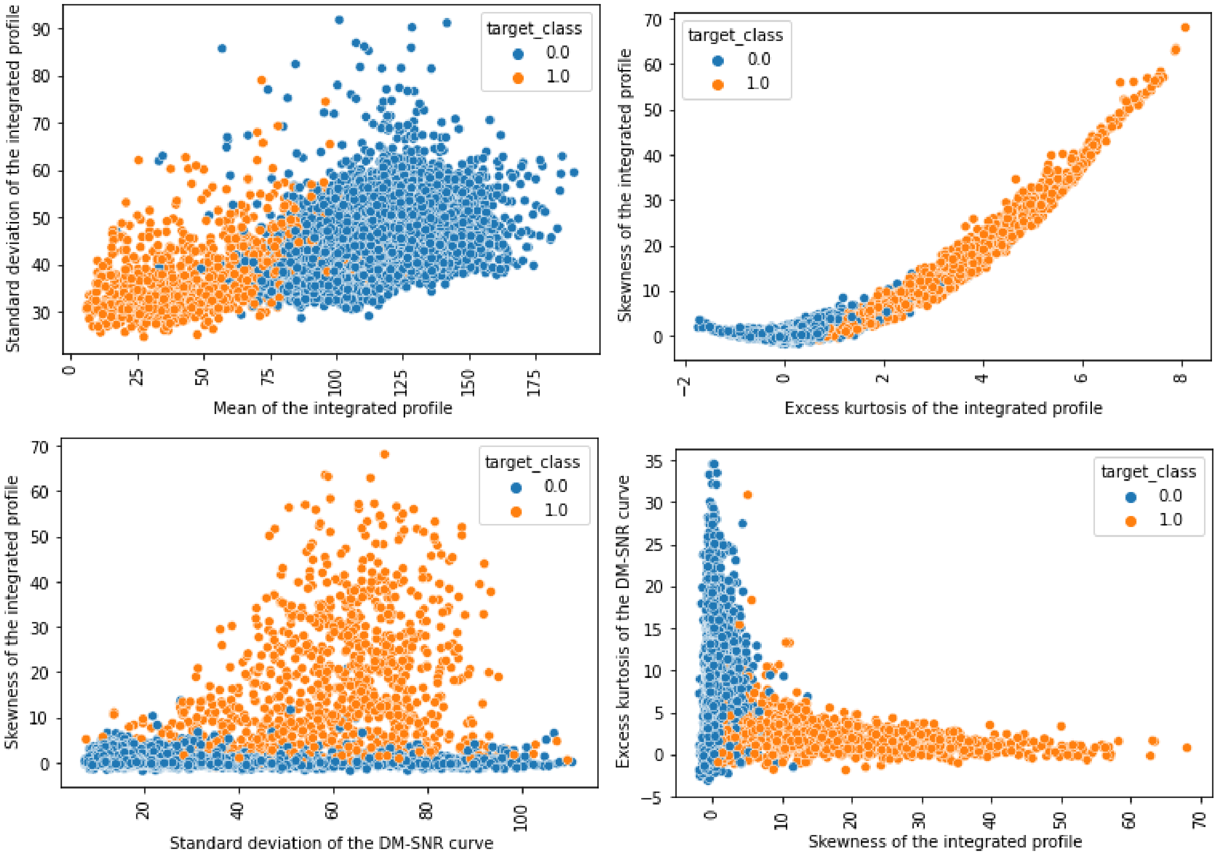


Fig. 2. Scatterplot of (a) Std of integrated profile (IP) vs mean of IP (b) skewness vs kurtosis of IP (c) Skewness of IP vs Std of DM-SNR (d) Kurtosis of DM-SNR vs skewness of IP

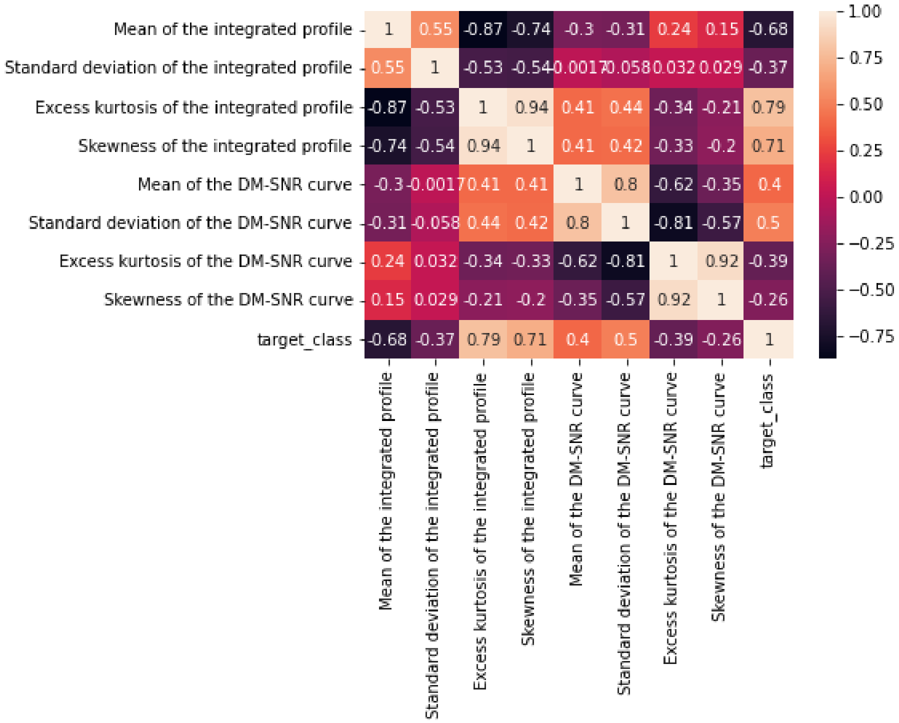


Fig. 3. Count plot of Number of persons split by Target

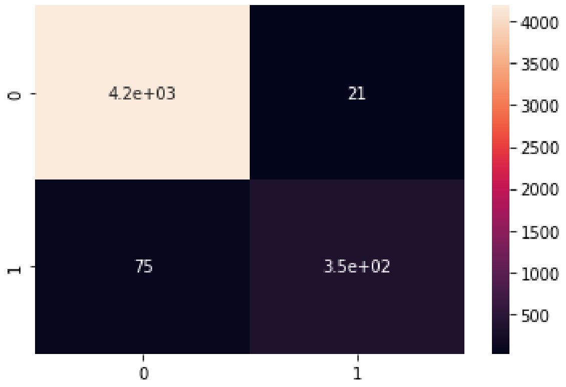


Fig. 4. Feature importances learnt by the Random Forest model