

Assignment 6: Data Lab, Support Vector Machine

- Due by [Friday 26th November, 2021 by 5pm IST](#).
- To be submitted to the following email address: office.of.gr@gmail.com
- The subject of the email should be: [Assignment Number \[6\]: Data Lab, 2021](#)
- Please clearly mention your name and roll number.
- Submit your work as a single pdf file. Additional material, code, etc can/should also be submitted, but there should be atleast 1 pdf, which has the entire assignment.
- Wherever there is code, in the assignments, the code should be well documented and easy to understand / follow.

The objective of the assignments is three fold. One is to be able to develop expertise in writing and communicating about technical topics. This will be done by using the IEEE conference style format for all assignments. The other is to explain, in your own way, the mathematical ideas that are embedded within the technical topic of interest. For example, in this case it is support vector machine. The third is to use the topic, in this case of support vector machine, to understand a problem from the real world. So in a sense the objective is to write what one may call a mathematical essay on support vector machine.

Title could be: Assignment 6: a mathematical essay on support vector machine.

Abstract: Give a brief overview of your assignment.

Author: Name, Department, Institution, Email

Section 1: Introduction

In this section, the 1st paragraph should be on a broad overview of the topic. The 2nd paragraph should be an overview of the technical aspects (i.e. in this case it is a support vector machine). The 3rd paragraph should be about the problem that you are aiming to solve/understand using support vector machine. Finally, the 4th paragraph should give an overview of the paper.

Section 2: Support Vector Machine

This section should outline the key principles underlying support vector machine.

Section 3: Data

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. The dataset is contained in pulsar_data_train.csv and pulsar_data_test.csv and the variables are summarised below:

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

Section 4: The problem

- (a) Outline the problem, and plot/visualise the data.
- (b) Make progress on the problem, by applying the techniques of support vector machine to the problem at hand.
- (c) Discuss any insights and observations.

Section 5: Conclusions

Write about 1 paragraph on the key insights that were obtained from your study and also outline any further avenues of possible investigation.

References

Please put in all the references that you have used during the assignment. The format should be the same as the IEEE conference format.