Choice of Dataset: Stellar Classification Dataset - SDSS17 (Kaggle [1])

Introduction

Stellar Classification is the classification of stellar bodies (galaxies, stars, etc.) based on their spectral characteristics. Effectively, knowing that different light-emitting stellar bodies read different filter values in the electromagnetic spectrum, it is possible to discern them given these filter values. While deterministic methods had been already determined that employ Physics and Electromagnetism to classify stellar bodies, this paper proposes the development of a Machine Learning Model to classify stellar bodies.

The dataset that will be used is the Stellar Classification Dataset uploaded by user "fedesoriano" on Kaggle (Link at top of the document). The dataset provided by the user contains cleaned and compiled data retrieved by the Sloan Digital Sky Surveys [2].

Since we are only interested in predicting the class of a stellar body knowing it's spectral characteristics, we will only require the 5 color filter values (**u**, **g**, **r**, **i**, and **z**), and the class field (either Galaxy, Star, or Quazar) of the stellar objects provided by the dataset. Since we won't require more information, we will pre-process the data to only include our required fields.

Methodology

As we are dealing with predicting the output of a system that deals with 5 numbers as its input, we have an optimization problem where we will optimize (5?) parameters. As such, the optimization could be done using gradient descent, where we would iteratively optimize a vector of parameters. This solution sounds robust as we are dealing with a purely numerical input, and a classification output.

As this model is a classification model, the evaluation metric that applies here is the confusion matrix. After training and validating the data, a lenient lower bound for the overall accuracy of the classifier would be 0.85. Note that this number is only tentative, and that it is subject to change as the project progresses.

Application

The integration of the model will be done in a webapp, where users would input the color filter values of their stellar body through a HTML form text input field and receive the output which would state the class of the body. I am also examining the possibility of letting the user upload a picture of the spectral graph of a stellar body (from which we would extract the required parameters.) Moreover, the webapp would utilize its graphical interface to provide understanding on how spectral classification works and EM waves, and, provides insight on the functioning behind the model itself.

References

[1] Fedesoriano, "Stellar classification dataset - SDSS17," Kaggle, 15-Jan-2022. [Online]. Available: https://www.kaggle.com/fedesoriano/stellar-classification-dataset-sdss17. [Accessed: 21-Feb-2022].

[2] https://arxiv.org/abs/2112.02026