Variational Autoencoders on Astronomical Catalogs

Ash Karale and Ariella Atencio

Machine learning is growing as one of the essential skillsets that a professional astronomer should possess. This is due to the fact that the astronomy community must deal with the ever-increasing amount of big data. For instance, the upcoming Legacy Survey of Space and Time will survey the universe and produce 20 Terabytes of data per night. It will be impossible for any human on the planet to analyze this amount of data. Therefore, machine learning proves as the essential pathway for big data analysis. A subset of machine learning is deep learning. Deep Learning employs neural networks for object classification and uses a chain of hidden layers to encode and decode data. This project will thus utilize a type of neural network called a "variational autoencoder" to classify quasi-stellar objects from other astronomical sources.

To understand variational autoencoders, we must first discuss neural networks. Neural networks are computer programs that imitate the behavior of the human brain. They recognize patterns and solve problems using those patterns. The neural network used in this project is a variational autoencoder. A variational autoencoder's encoding distribution is regularized during training to guarantee that its latent space has good properties to generate new data. This variational autoencoder will be used on data generated from DELVE, VHS, CatWISE, and GAIA. All four of these surveys are in different geographic locations taking images of different parts of the sky focusing on different portions of the electromagnetic spectrum. The information from these four surveys is combined to get a catalog of data made into one fits file. Looking at the image data with the naked eye doesn't provide much information, technology is needed to distinguish quasars from stars and galaxies. In the past, this analysis was done only using spectra, but this resulted in a lot of time and effort. Machine learning classification, on the other hand, is far more efficient requiring less time and effort. For instance, images are classified as point sources or extended sources with the help of classification algorithms. A point source can be thought of as a pin hole through which sunlight is entering, as a point source of light. An extended source has an angular size greater than the resolution of the instrument used to observe it, so it doesn't appear as a point with no size. In most cases, Quasars and galaxies appear as extended sources whereas stars and planets appear to be point sources.

This classification process begins via importing all the necessary modules into the jupyter notebook. The programming language used in Python 3, and the main framework used for processing our neural network model is PyTorch. Next, the data is acquired and preprocessed to check for errors and null values. Once the data set is ready, the autoencoder can begin to build its model and define an algorithm. The algorithm must then be trained on training data and checked against known 'test' data. At this point, much of the model code has been completed. Then the data will be imported into the defined model. The variational autoencoder will begin running on the training data for a number of epochs. Furthermore, the model will be validated using test data

to test the accuracy of the model. This will be done by running accuracy scores and plotting the confusion matrix.