Naimish Mani B
naimish240@gmail.com

Dimensionality Reduction for Studying Diffuse Circumgalactic Medium

9th April, 2021

Note to the Mentors

Thanks for taking a look at this! I think this would be a really interesting project and I'd love to work with everyone to make it happen!

ABSTRACT

This project aims to make use of Machine Learning-based dimensionality reduction techniques on simulated quasar absorption spectra to better understand the Diffuse CircumGalactic Medium (CGM). This project is to be carried out with the mentors Jeremy Bailin (University of Alabama), Jacob Morgan (University of Alabama), Sergei Gleyzer (University of Alabama), Jianghao Huyan (University of South Carolina) and Varsha Kulkarni (University of South Carolina) through the organisation "ML4SCI" as a part of Google's Summer of Code 2021.

PERSONAL INTRODUCTION

Hi! I am Naimish (*nei-me-sh*), a student pursuing my undergraduate degree in Mechanical Engineering at SASTRA Deemed to be University, Thanjavur India. I'm currently serving as the head of the Web Development cluster of my university's Google Developer Student's Clubs chapter. I also work actively with the Machine Learning team to train other students on campus. I am also currently serving as the President of the university's space club (Stellaria-SASTRA).

At Stellaria, I am leading a team of 4 to work on the Kepler K2 dataset for detecting exoplanets. Last year, I worked on a pet project to generate new anime character faces using Autoencoders, initially by randomly sampling from the latent space and then by finding points that generate more realistic faces through Principal Component Analysis. It was a fun project which taught me the fundamentals of how Convolutional Neural Networks, Autoencoders and PCA work.

I have been passionate about satellite technology ever since I was a kid. One of my lifelong ambitions is to build a communication satellite constellation to provide rural and remote India with access to a reliable means of communication during times of natural disasters. I found a team of passionate and motivated people too, and we will be working on it once we graduate.

I have worked with TensorFlow, NumPy and Pandas, and am comfortable with Jupyter Notebooks. I look forward to working on this project to broaden my horizons and gain research experience working with the distinguished mentors.

ABOUT THE CGM

The CircumGalactic Medium (CGM) broadly refers to the gas/plasma found in the region surrounding a galaxy beyond the galactic disk but within the galaxy's virial radius. Studying the CGM helps us understand how galaxies form, and hence broadens our understanding of the universe. In the mid 1950s, Guido Münch observed neutral sodium and singly ionised calcium in the spectra of hot stars at high Galactic latitudes. Lyman Spitzer interpreted this to be evidence for the presence of a hot, diffuse gas not planar to the Milky Way. This gave way to the idea of a "Galactic Corona", which metamorphosed into the foundation for the CGM theory.

The CGM is currently studied through the following methods:

1. Transverse Absorption-Line Studies
2. Stacking Analysis
3. Down-the-Barrel spectroscopy
4. Hydrodynamic Simulations

Of these methods, the method of simulation is of particular interest for us, as we are working with spectra generated by means of Monte Carlo Simulations.

(Ref. : https://arxiv.org/pdf/1709.09180.pdf)

OBJECTIVE

Through this project, I look forward to learning how spectroscopic data is generated through Monte Carlo simulations, test Linear Discriminant Analysis as a dimensionality reduction technique, and build Neural Autoencoders for dimensionality reduction. These have been summarised as a list of goals below.

Also, throughout the 3rd evaluation task, I only worked with 21 of the 28 columns. I look forward to learning about how the final 7 columns were generated and what they signify.

Dimensionality reduction, in this context, is being performed to aid classification efforts. Hence, I believe that learning features and not reconstruction is of higher priority. If we are able to build an AutoEncoder that "learns" features which aid the classification task, but at the cost of highly-accurate reconstructions, I am assuming it is fine for our task.

GOALS

1. Learn about the CGM and how quasar absorption spectra helps us study it.
2. Test LDA, PCA and t-SNE as viable methods of dimensionality reduction on the dataset.
3. Build and test Autoencoders of different architectures to gauge their performances and help us identify the optimal network for the task.
4. Perform PCA, LDA, t-SNE on the latent vector and study the efficiency of the network.
5. Build a Python module to aid this task.
6. Learn how the spectroscopic data was generated through Monte Carlo Simulations.

METHODS

For Task 3 (B), I tried implementing PCA and Neural Autoencoders for dimensionality reduction. But PCA was only performed on a limited sample of the dataset (less than 5% of the whole dataset), and the Autoencoder tested was also a very basic one. So, the next logical steps would be to re-run PCA and also test LDA and t-SNE on the entire dataset to see if there's any difference, and also build deeper Autoencoders and test the performance of various architectures. Random Forests also offer potential for dimensionality reduction, but it did not yield a promising result in the evaluation test. Adjusting the depth and number of nodes is something to look into, though.

The Autoencoders will be built using TensorFlow and tested through Jupyter notebooks. Personally I prefer making standalone Python scripts in tandem with Jupyter notebooks, and would look forward to doing the same, as it enables smoother collaborations. Subsequently, I look forward to publishing a library on PyPi for this project.

PROPOSED TIMELINE

Note : The timeline below has been designed to accommodate my academic schedule, and is flexible.

Learning about the dataset (May 17 to May 28)

During these 2 weeks, I look forward to learning about how the study of Quasar absorption spectra is used to study the CGM, learn about how the dataset was generated, what the two classes in the task signify, and the significance of the final 7 columns. I also look forward to developing a good rapport with my mentors. This and the subsequent phase (Discussing benchmark metrics) are to take place during the "Community Bonding" period of the project. As

GSoC recommends students start coding after the Community Bonding phase, I have hence incorporated the same.

Discuss benchmark metrics (May 29 to June 6)

This week, I look forward to discussing the metrics which we will be using to quantify the effectiveness of the dimensionality reduction algorithm. I.e., if we will be using a classifier to quantify the performance as we did in Task 3, then deciding the algorithm used by the classifier and its architecture, etc for the same. This will be used as a baseline test for all future tests.

Test traditional dimensionality methods on the whole dataset (June 7 to June 13)

During these 7 days, I will implement and test LDA, PCA, Random Forest and t-SNE as dimensionality reduction techniques on the entire dataset and hence evaluating their performances using the prior decided upon benchmarking metrics. LDA works by analysing the class-wise relations of the dataset, PCA identifies the principal components from the eigenvectors, and t-SNE is used to give us a feel for how the data is distributed in higher dimensions through stochastic methods. Random Forests' can be used as feature selectors, and their performance is also something we can test out during this phase. All these functions are available in sklearn already, and hence I will make use of the given modules. As all of these algorithms are resource intensive, I have allotted a few extra days to ensure the processing runs smoothly on my end.

Developing custom classifiers to work with the reduced data from traditional methods (June 14 to June 24)

For this week, I plan on implementing and testing custom networks / algorithms to act as case-specific classifiers, hence evaluating their new performances with these new context-dependent benchmarks. Since we cannot reasonably guess without prior testing which algorithms will work better than others, we will run through the battery of tests (Logistic Regression, RandomForest, Multi Layer Perceptron and XGBoost, just to name a few) to identify the best performing compression technique and the classifier which we use to evaluate its performance.

Build and prototype Autoencoders (June 25 to July 8)

During these two weeks, I look forward to implementing and testing various autoencoder architectures on a limited version of the dataset as a form of "wide search" to find good prospective architectures for testing on the full dataset. I also plan on employing PCA / LDA / t-SNE on the autoencoder's latent vectors to interpret the AutoEncoder's "learnings". As we will

be training multiple deep neural networks (a very-compute intensive task), I have been very conservative with the dates for this part.

Prepare Report for Phase 1 Evaluation (July 9th to July 11th)

For this week, I plan on preparing a comprehensive report of the work done thus far with the feedback and help of mentors for the first phase of evaluations.

Phase 1 Evaluations (July 12 to July 16)

These days have been reserved for addressing and responding to any last-minute updates/fixes for Phase 1 Evaluations. If time permits, though, I will start working on the next task immediately.

Training and testing the Auto Encoders (July 17 to Aug 1)

During these 16 days I look forward to training and testing the autoencoders on the full version of the dataset to get a much more comprehensive understanding of the networks and also evaluate their performances with custom classifiers. Since it will be a very resource-intensive task, I have been conservative with the number of days it'll take.

Writing a module to perform dimensionality reduction on input data using the algorithms developed so far (Aug 2 to Aug 9)

For this week, I plan on converting all techniques and algorithms developed thus far into a script that can be called to perform dimensionality reduction on the input data using any of the algorithms developed for use in future projects.

Prepare report for final submissions (Aug 10 to Aug 16)

During the final week, I plan on preparing a comprehensive report outlining all the work done as part of this project with constant feedback from the mentors for the final evaluations, while also addressing any last-minute unforeseen circumstances.

FUTURE WORK

The above-mentioned goals and timeline only address the dimensionality reduction aspect of studying the Diffuse CGM. But it was mentioned on the project page that the quasar's absorption spectrum data can be used to determine the gas composition, temperature and density as well. So I would like to study these aspects of the dataset as well, and work together to help build algorithms and libraries for the same. Apart from this, I am sure we will find other potential future works as we carry out progress on the project.