DS-GA 3001: Computational Approaches to NLP

Understanding the Language of Black Holes
Fall 2016

Nora Barry
NYU Center for Data Science
neb330@nyu.edu

Katrina Evtimova
NYU Center for Data Science
kve216@nyu.edu

Advisor:
Dr. Daniela Huppenkothen
Moore-Sloan Data Science Postdoctoral Fellow
NYU Center for Data Science
dh2288@nyu.edu

1 Introduction

In the field of astronomy, light curves represent the light intensity of a celestial object or region as a function of time. Based on the observations of changes in brightness over time, astronomers try to capture recurring patterns and identify different classes of stellar events such as the birth of a supernova. Within the existing literature, manual classification of light signals coming from black holes has been done in [3]. Also, deep learning has been applied to the supervised classification of supernovae [4]. However, in our case, we are provided with light curves data coming from black holes X-ray emissions with no pre-existing labels. The goal of the project is to do an unsupervised identification and classification of different patterns in the emissions, possibly with the use of some of the deep learning techniques covered in this class.

In unsupervised learning, finding a suitable evaluation metric is often a challenge. Usually, the "ground truth" can be used to evaluate a model, ie. using labeled data during testing. However, since black hole X-ray emissions have never been scientifically classified, we will not be able to manually label our data. In other words, there are no labels for these time series, so this model may be the first way of unbiasedly classifying them. So, the evaluation of our model will be exploratory in that we will use visual similarities, among other features to interpret the accuracy of the clusters.

In terms of data, we have access to 13,496 files (amounting to 1.2GB) of time series on light emitted from 29 black holes. The files are encoded in FITS (Flexible Image Transport System) format. In addition to the light curve data, each file contains metadata such as the date and time of the observation and the telescope that was used. The frequency of the observations is at the second level and the series vary in terms of length.

2 Overview

One of the main challenges we will face in applying deep learning techniques to our dataset is the fact that our data are one-dimensional time series. Deep learning models are best at processing and learning from large dimensional datasets and tend to overfit smaller, low dimensional data. As a result, we will have employ multiple methods to process our time series to create a format that's more conducive to the deep learning framework. In particular, we have found a Python library specifically built to extract potentially important features from astronomic time series. This library is called Feature Analysis for Time Series (FATS), and it's able to generate over 40 features such as mean, standard deviation, linear trend and skew from the input time series and error data. We seek to utilize

FATS for both the time series as a whole, and for small windows of the time series, much like how Short Term Fourier Transform is applied. Thus, if we were to employ the latter option, we would have a matrix of feature vectors, and the input would look more like the usual input for a deep learning model. We are also going to consider the standard statistical approaches to processing time series such as Short Term Fourier Transformations, wavelets and PCA. However, we will mainly focus on the FATS option initially.

Once we have found a logical way to represent the time series data, we seek to feed it into a deep learning model to perform unsupervised classification. We have found a framework in [2] called Deep Embedding Clustering (DEC) that seems to suit our interests, and will use parts of that model to perform the clustering. In this setup, the data is first passed through a Stacked Autoencoder (SAE) which essentially de-noises the data and maps it into a feature space. This would be similar to taking the most common words in a corpus and indexing them with integers. Our utilization of this step will likely depend on how we decide to represent the data, which is discussed above. We will then use a Deep Neural Network to learn the embedding for the data output from the SAE (or taken directly from our FATS feature vector if we decide to use that option). After, we seek to use k-means clustering to obtain initial cluster centers, which will help us begin the unsupervised clustering process with a deep learning model. This model will then carry out two steps in parallel. First, it will assign each embedded point to a cluster and second, it will update the embedding by using stochastic gradient descent. In this case, the loss to be minimized is the KL Divergence between the cluster assignments and a target distribution of clusters (which will be a predefined distribution that strengthens the model's predictions). We will stop this iterative process once a small percentage of data points change cluster assignments in a certain iteration.

3 Conclusion

As discussed with the instructors for this class, this project is not completely aligned with the course material and is very open-ended. However, we seek to approach the classification task from a deep learning perspective using the proposed architecture in [2]. In the end, we hope to obtain an automated way of categorizing black holes.

We are planning to follow this tentative schedule:

1. Weeks of Oct 17 and Oct 24: Literature review & pre-processing of the data
2. Weeks of Oct 31 and Nov 7: Extract features using ML/time series techniques
3. Weeks of Nov 14, Nov 21 and Nov 28: Feed features into a DL model and perform classification
4. Weeks of Dec 5 and Dec 12: Wrapping-up (results, analysis, write-up)

References

[1] FATS: Feature Analysis for Time Series - `https://arxiv.org/abs/1506.00010`

[2] Unsupervised Deep Embedding for Clustering Analysis - `https://arxiv.org/abs/1511.06335`

[3] A model-independent analysis of the variability of GRS 1915+105 - `https://arxiv.org/abs/astro-ph/0001103`

[4] Deep Recurrent Neural Networks for Supernovae Classification - `https://arxiv.org/pdf/1606.07442.pdf`