6.867 Machine Learning (Fall 2018)
Prof.: Suvrit Sra
TA: Alex Turner

Youssef Berrada (ID: 911844154)
Georgy Guryev (ID: 924667514)
Sheng Yao (ID: 916855968)

Project proposal: PLAsTiCC Astronomical Classification

Introduction

The project we choose to take on is the The Photometric LSST Astronomical Time-Series Classification Challenge funded by National Science Foundation. The challenge involves classifying astronomical sources that vary with time into different classes, scaling from a small training set to a very large test set of the type the LSST will discover. The data of the astronomical sources is simulated in anticipation of what would actually be observed by LSST. We hope that by experimenting different classification techniques we can provide a superior classifier for labeling the astronomical sources and empower the scientific endeavor of the LSST project.

The Dataset

The dataset we use was collected from numerous members of the astronomical community to provide models of astronomical transients and variables. The dataset includes information on multiple astronomical objects. The data can be found here:

https://www.kaggle.com/c/PLAsTiCC-2018/data

The data comes in two forms

- Metadata Objects: header files that contain summary (astronomical) information about the objects. Each objects is uniquely identified by an integer and provide information about the astronomical object.
- Time Series: light-curve data for each object consisting of a time series of fluxes in six filters, including flux uncertainties.

We will provide with insight and prediction for each astronomical element, based on its characteristic from the first set of data, and use the time series to make our prediction. this work will involve analytical tool detailed below.

Analysis Techniques

This project will help us understand classification methods for multi-class classifications. We would like to first understand our dataset by cleaning our data using different techniques. After understanding our features, we will explore normalizing features across time without introducing forward-looking biases. This project will also be the opportunity for us to a variety of tree-based classification methods. The main classifiers we intend to experiment with include decision tree, random forest, gradient boosting and optimized decision tree. We hope that by comparing these models we can provide a reliable classifier for the task.

Impact of The Project

The project aims to tackle a specific challenge, mainly, how well can we classify objects in the sky that vary in brightness from simulated LSST time-series data, with all its challenges of non-representativity? We hope that by adapting existing classifier to a time-series setting, we can provide a classifier that tailored to the nature of the task. The classifier will be used to support the actual endeavor of LSST by helping us to categorize the astronomical objects observed, once the training and evaluation are completed on the simulated data.

Expected Output

- As part of the Kaggle competition, we need to submit a matrix of probabilistic classifications where the sum of probabilities across all classes per row (object) is unity.
- As part of the project for the 6.867 class, we will provide a survey of methodology we went over as well as some performance metrics to compare them and understand the pros and cons for each of them. As outlined above, we will discuss how methods like optimal trees might outperform in such a setting and provide with an analytic edge.