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 Milestone 2

Methodology

The methodologies that we plan to experiment for this project are different types or variations of the tree-based methods. We hope to achieve a balance between the accuracy of the models and the interpretability. On one hand, we seek to provide a model with optimal performance, but we also hope to uncover relationships from the data to serve future astronomical discoveries. The list of methodologies that we intend to adopt are the following:

First we would like to apply different decision trees techniques to classify the astronomical objects. For instance, The **CART model** provides intuitive interpretations as it breaks down the decision variable and cutoff at each step, but the 'greedy' nature might not lead to best predictions as the local splits might not be optimal. Then, The **Random Forest** and **XGBoost** are more robust to input perturbations as they add randomness in the training process from variable selection to splitting, but since the final model is aggregated over different 'trees', they are difficult to unravel and provide intuitions. The **Optimal Decision Tree** is an experimental method that seek to balance between the robustness and interpretability, by re-evaluating an existing decision tree randomly at each node.

We also would like to explore how to use deep learning in this set up by using **RNN** to predict the class of the astronomical object. We will maily focus on implementing **LSTM** (Long-short Term Memory) RNN

Evaluation

The main goal is the accuracy of our predictions and we seek to obtain consistent performance for all categories, which is an important evaluation for the competition. Specifically, we use the following loss function to evaluate our predictions:

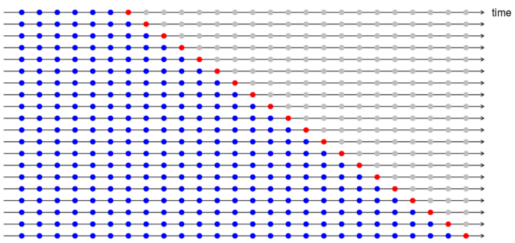
$$L = -\frac{\sum_{j=1}^{M} w_j \sum_{i=1}^{N} \frac{1}{N_j} \tau_{i,j} \ln P_{i,j}}{\sum_{j=1}^{M} w_j}$$

where $\tau_{i,j} = 1$ if the ith object comes from the class j, and N_j is the number of objects in a class j, and w_j are individual weights per class which we decide to equal weight. This designs is aim to give an equal importance to each class we need to classify.

Cross Validation

In time series, we need to avoid using future information to construct the forecast. For validation, we will use the early portion of the sample as training sample, the remainder as the test sample. For cross validation, one possibility is to gradually expand the window for the training set, with the following observation as the test set. One way to see that is do the following[1]:

1. For $i = 1, 2, ..., T - k$, select the observation at $t = k + i$ as the test set. Use the observations from $t = 1$ to $t = k + i - 1$ as the training sample.
2. Fit the model on the training sample and compute the prediction error for observation at $t = k + i$.
3. Repeat Steps 1-2, and compute the overall average cross-validation error.



Division of Labor

All three team members contributed significantly to the milestone. We actually all brainstormed at the same time to start working on the datasets and discussed what methods should we explore. We then agreed on the evaluation score we will choose and how to perform a cross validation for time series process.

[1] https://robjhyndman.com/hyndsight/tscv/