ORIE 4741 Project Proposal

Chengnan Xu Tian Yu

September 30 2019

The big data era brings new challenges of signal processing. With sensor and storage technology developing, data collection and storage become cheaper and easier. At the same time, more economic and social activities are mediated by the internet. Therefore, a huge amount of data such as social networking data, online shopping records are generated everyday. These large collections of data pose a problem for traditional data processing algorithms. As the volume of data grows faster than the number of data analysts or theorists, we require automated methods to process and make sense of this data.

Automated machine learning (automated ML) making it easier to build and use machine learning models in the real world by running systematic processes on raw data and selecting models that pull the most relevant information from the data – what is often referred to as "the signal in the noise." Automated machine learning empowers the people not only for expertise in data science, but also those who non-experts. Originally, for a machine learning project, people need to break down into discrete steps like model construction, model development and hyperparameter tuning. For some expertise in data science, they can utilize their strength to model construction. However, for those non-experts, the time consuming pieces of machine learning are hyperparameter tuning and algorithm selection. The job of the automated machine learning is to automate the algorithm selection and hyperparameter tuning tasks to find a promising algorithm with time constraints. Therefore, for some non-machine learning experts, the automated machine learning accelerates their process of doing machine learning research or some related useful applications.

In this project, our work is basically divided into two parts. First, we need to evaluate the existing algorithm e.x. Auto-sklearn, Auto-weka, H2O AutoML, TPOT, Oboe on the benchmark dataset to select a promising algorithm. Next, our task is to understand the math concept of the selected algorithm and implement it in python. We hope to find calculation-expensive part of the original algorithm and optimize the corresponding source code.