Project Proposal

The project I would like to propose is a machine learning model that's goal is to predict stock returns. Most basic models of this type analyze previous stock returns in order to predict future returns, and although this method does provide useful analysis, the stock market itself is a highly complex and chaotic system that is far from completely understood. As an addition to the most basic methods used to model the market I propose that in this project we include relevant textual news sources to help better capture the less quantitative elements of the market. I will be using both structured and unstructured data to train this machine learning model on. The structured data will be the end of day stock prices for the specific stock under analysis, and the unstructured data will be the raw text from the news articles relevant to the stock under analysis. More specifically the news source I will be using are the SEC company filings, of the stock under analysis. This decision was made because these filings can be correlated in time with the end of day stock prices, and contain relevant information about the health of the company under analysis. It also common knowledge that many investors look to these documents as a source of information to aid in their decision making. At its core this project is a supervised machine learning problem because the ground truth information on the past prices of stocks are available. Because our target variables are future stock returns, our project will involve some sort of regression modeling because our target variables are continuous variables. The features we will be trying to correlate with our targets, are the previous week's stock returns of a given target stock return, along with a vector representation of an SEC filing that was filed within the previous week. The vector embedding of the SEC filing will be determined using deep learning NLP techniques. We decided on this because deep learning text embedding techniques are currently state of the art. The full model will also involve deep learning techniques, because deep learning techniques tend to produce more accurate results at the cost of interpretability. We decided this sacrifice in interpretability was worthwhile because the goal of this project is to (hopefully) produce accurate results to facilitate in trading, and thus any gains in accuracy will be worth more than gains in interpretability. Also the stock market itself, is already a highly complex dynamic system that has not lent itself easily to interpretation for many years, thus more interpretable models will have a harder time accurately modeling the ebbs and flows of the market. Since this project will inevitably involve deep learning techniques, one of the computational resources required will be a GPU. This is because deep learning techniques involve the tuning of hundreds of thousands of parameters, and the only way to get this done on time scales that are not weeks to months is if GPU calculations are being used. Since the two data types I will be using are times series data of text files and stock prices, the memory requirements needed should be much less when compared to the memory requirements needed to analyze video footage or images. The dataset should be able to fit into a couple gigabytes of memory. Most of the heavy lifting will be done by the GPU this implies that the CPU will only need to be fast enough to not cause a bottleneck when transferring the data from disk to GPU. Therefore the CPU's speed is not of priority. All in all, the final deliverable of this project will be a trained model, along with a REST API that will allow a user to query to model to make a prediction for future stock returns.