Machine Learning Engineer Nanodegree

Capstone Proposal

David Pham

November 26th, 2017

Proposal

Investment and Trading Capstone Project: building an algorithm for representation of financial assets.

Domain Background

According to the link in the proposal about the investment and trading capstone project

> Investment firms, hedge funds and even individuals have been using financial models to better understand market behavior and make profitable investments and trades. A wealth of information is available in the form of historical stock prices and company performance data, suitable for machine learning algorithms to process.

Predicting stock prices is difficult task but the usual goal is to diversify our risks and try to get the best return possible. Understanding the decision of the our algorithm is also paramount and it would be interesting if we could create algorithms that could discover or retrieve structures in the financial instruments.

This subject is interesting because it would allow to study algorithms to classify financial with minimal information and maybe discover clusters only with the price movements.

Problem Statement

In this project, we would focus to create an algorithm whose input will end of day prices (EoD) of US equities and output some representation of each of the inputs. The algorithm could be extendable to financial futures instruments as long as we can keep end of day prices.

In a technical jargon, the world of financial assets could be assigned to a id, and we want to apply deep learning techniques to classify the assets and maybe discover hidden structures. This is a unsupervised learning problem, but we might have some intuition in some group of data.

Datasets and Inputs

As suggested from Udacity, we would like to use the data provided from Quandl. They have an open data providing the required data with the Wiki EOD Stock Prices.

Solution Statement

We might have the following sub problems to solves:

1. We apply some transformation to the data by computing the Z-scores of the day to day price difference. For this, we would need to compute a sample mean, and sample deviation and the exact procedure should be determined. This works as a standardizing steps.
2. Samples correlations will be computed (this could be computationally intensives with a lot of data).
3. An neural network will try to find clusters with the new transformed data.

The deep learning algorithm will be inspired from the Word2Vec from Google for word embedding.

Benchmark Model

We will use the correlation matrix and Hierarchial clustering, also see scikit, is a good benchmark model with maximum distance between groups as linkage criteria.

Evaluation Metrics

We will use the silhouette score as evaluation metrics. The silhouette score measures how appropriately the data have been clustered.

Project Design

First part will be to download and understanding the data. A module should be written to record all the transformations and make it reproducible. Then data will be cached to avoid doing the same computation over and over.

A second step would be implemented the benchmark model to set our expectations. Then, we would implemented the neural network (in keras and tensorflow). Tensorboard should be then used to inspect the embeddings and we should try to make exports in csv and maybe to polish the visualizations..

Finally, we should try several hyper parameters to observe any difference, report the differences.