Machine Learning for Investing Behaviours on Cryptocurrencies

Aims: Generate insight of popular cryptocurrencies in time-series by establishing correlation between them with unsupervised learning algorithms such as K-Means Clustering and applying deep learning algorithms like Long Short-Term Memory under Recurrent Neural Network to predict the price variation on the next day.

Background: Cryptocurrencies like Bitcoin and Ripple are becoming popular in these years. They can be obtained through mining or transactions. Every individual cryptocurrency can be transferred directly using a public key in digital wallets. Their price is mostly tied to supply/demand and hard to be interfered by governments. In addition, cryptocurrencies have some dependencies because some of them like Dogecoin can only be bought by some major cryptocurrencies such as Bitcoin and Ethereum in cryptocurrency exchanges.

There are lots of historical resources on the internet that can be obtained by an API, downloading csv files or crawling web pages directly. The easiest way is to access through an API and integrate with the analysing backend. One of the best APIs for cryptocurrency is https://min-api.cryptocompare.com/ which is free for personal purposes.

K-Means Clustering is a popular unsupervised learning algorithm for grouping unlabelled data. Deep Learning becomes very popular in recent years because of better computation on machines and larger scale of data, and it can predict data more precisely even though the difficulty of explanation. Recurrent Neural Network is a multi-layer neural network with time-series in Deep Learning, and Long Short-Term Memory is one of the most popular one in it due to its capability of eliminating vanishing/exploding gradient problems.

An implementation of modern algorithms combining nearly pure supply/demand problem is highly possible to capture investors' investing behaviours. If patterns of investing behaviours are found, the most/least profitable products can be predicted accordingly and be used for investment.

Early Deliverables

1. A report giving an overview of Clustering.
2. A report describing Recurrent Neural Network and Long Short-Term Memory.
3. Precise specification of the methodology and the shape of inputs being used on the algorithms after discussion with the advisor.
4. Proof of concept programme: K-Means Clustering and Long Short-Term Memory applied to a small number of datasets.

Final Deliverables

1. An automatic mechanism pulling data from the API to the local PostgreSQL database in Docker using Java with Maven/Gradle.
2. Building prediction models on the previous programme with Deeplearning4j.
3. Integration with Spark to accelerate the speed of analysis.
4. Training models with the real data from the local database.

5. Examine the prediction accuracy by conducting backtesting through time.
6. Final report will discuss techniques, findings, how to improve the result and some meaningful visualisations.

Suggested Extensions

- Building a local web app displaying data and visualising predictions.
- Transforming the programme to near real-time data extraction and analysis.
- Executing text mining to detect investors' sentiment from forums such as Reddit and 4chan/biz, then put it as factors into the prediction models to enhance the accuracy.

Reading

Adam Gibson and Josh Patterson, Deep Learning - A Practitioner's Approach, 2017
Research papers.

Prerequisites: Java, Maven/Gradle, Deeplearning4j, Spark, PostgreSQL and Docker. Javascript/Typescript, React, Redux, HTML and CSS/SCSS if extensions are implemented. Good knowledge of calculus, especially differentiation.