Machine Learning Engineer Nanodegree
Capstone Proposal

Lokesh Soni

October 7th, 2018

Cryptocurrency-predicting

Domain Background

A cryptocurrency (or crypto currency) is a digital asset designed to work as a medium of exchange that uses strong cryptography to secure financial transactions, control the creation of additional units, and verify the transfer of assets. Cryptocurrencies are a kind of alternative currency and digital currency (of which virtual currency is a subset). Cryptocurrencies use decentralized control as opposed to centralized digital currency and central banking systems. The decentralized control of each cryptocurrency works through distributed ledger technology, typically a blockchain, that serves as a public financial transaction database.

Problem Statement

This is a classification problem.Inputs are 60 feature sets of combine price and volume for each coin and the output will be the prediction of the “price” of coin.
The overall goal of the project is to construct a machine learning model that can predict price trends with results superior to that of random selection. As the cryptocurrency has a reputation of being a very speculative investment, driven to a high a degree by the emotional frenzy of amateur investors, we will attempt to discern what impact this has on the market by incorporating various different features and the impact on performance.

Datasets and Inputs

Input data fields will be as follows-

- Open and Close
- High and Low
- Volume data for Bitcoin
- Ethereum, Litecoin and Bitcoin Cash.

For my purposes here, I’m going to only be focused on the Close and Volume columns.The Close column measures the final price at the end of each interval. The Volume column is how much of the asset was traded per each interval, in this case, per 1 minute.In the simplest terms possible, Close is the price of the thing. Volume is how much of thing.

What I want to do is somehow take the close and volume from one coin data, and combine it with the other 3 cryptocurrencies.Next, I need to balance and normalize this data. By balance, I want to make sure the classes have equal amounts when training, so our model doesn't just always predict one class.Then, all I need to do is split the data back to feature sets and labels/targets.

Solution Statement

The solution will be predictions if price will rise or fall. So, I need to take the "prices" of the item I am trying to predict.

Besides that, I will:

1. Balance the dataset between buys and sells.
2. Scale/normalize the data in some way.
3. Create reasonable out of sample data that works with the problem.

I will be going to work on using a recurrent neural network to predict against a time-series dataset, which is going to be cryptocurrency prices and fine tune parameters to get best accuracy.

Benchmark Model

For this problem, the benchmark model will be random forest model. I will try to beat its performance.

Evaluation Metrics

Prediction results are evaluated on the log loss between the predicted values and the original values. My prediction model calculates the rise or fall in the price.

Project Design

Before even start training models, I will first take glimpse of the data see what the shape and is and how they are formatted. Then I will start extract information such as Open, High, Low, Close, Volume data for Bitcoin, Ethereum, Litecoin and Bitcoin Cash. For my purpose, I'll be only focussing on the Close and Volume columns.

The Close column measures the final price at the end of each interval. In this case, these are 1 minute intervals. So, at the end of each minute, what was the price of the asset. The Volume column is how much of the asset was traded per each interval, in this case, per 1 minute.
Then I will balance and normalize this data. By balance, I want to make sure the classes have equal amounts when training, so our model doesn't just always predict one class. Then, all I need to do is split the data back to feature sets and labels/targets.

Now for the model, I am thinking to try a few things like 2 vs 3 layers, 64 vs 128 nodes. I will start with the simple model first, which is sequence_length: 60, futur_period_predict: 3,layers 2 and 64 nodes. After seeing the result, I am going to adjust the parameters with some dropout or batch normalisation and try to get the best result.

References -

https://en.wikipedia.org/wiki/Cryptocurrency

https://people.scs.carleton.ca/~maheshwa/courses/3801/Projects17/crypto-report.pdf

Understanding LSTM Network
(https://colah.github.io/posts/2015-08-Understanding- LSTMs/)

Data Sets
https://pythonprogramming.net/static/downloads/machine-learning-data/crypto_data.zip