Zijun Hao(zh2282@columbia.edu) & Qing Zhou(qz2273)
Lecturer: Ching-Yung Lin
E6895 Advanced Big Data Analytics
03/22/2017
Market Intelligence Analysis: Bitcoin Price Analysis

1.Introduction

Our goal is to analyze the history data of bitcoin, and predict its price in the future. Considering that everyday's bitcoin price is affected by the events and bitcoin price in the days before, we use Long Short Term Memory(LSTM) networks as the training tool of our model. As for the history data of bitcoin, we pay attention to the Bitcoin data as well as the social media data related to bitcoin. As the final outcome of our analysis, a graph indicating the price trend of bitcoin given by our network will be given, and we will evaluate the outcome with several fitting criteria.

1.1 Bitcoin

Bitcoin is a cryptocurrency and a payment system invented by Satoshi Nakamoto. Bitcoin can be obtained by mining and transactions[1].

2.Related Work

Shah et al. implemented a latent source model as developed by Chen et al. to predict the price of Bitcoin.
Greaves et al. analysed the Bitcoin Blockchain to predict the price of Bitcoin using SVM and ANN.
McNally, Sean (2016) Predicting the price of Bitcoin using Machine Learning.

3.Methodology

3.1.Bitcoin data preparation and processing

(1)Data structure
The Bitcoin data can be divided into 3 areas. First, we would like to know how the Bitcoin market behaved in the past, which is intuitively closely related to and visually displayed in the price fluctuation. Bitcoin activity, In addition, is another closely related factor that involves the Bitcoin's transactions and trade. What's more, we should not neglect the impact of the production, or say the driving force of Bitcoin, which is similar to the circumstance that we consider GDP when analyzing stock data. Considering the economic basis of Bitcoin and the related factors above, we extract 15 features from the primary data as the elements of feature vector.
Date: a year-month-day format date, representing that all the following features are related to that certain day.
Total Bitcoins: the historical total number of bitcoins that have been mined.
Bitcoin Market Capitalization: the total number of bitcoins in circulation.

Bitcoin Number of Unique Bitcoin Addresses Used: number of unique bitcoin addresses used per day.
Bitcoin USD Exchange Trade Volume: the USD trade volume from the top exchanges.
Bitcoin Number of Transactions: total number of unique bitcoin transactions per day
Cumulative Number of Bitcoin Transactions
Bitcoin Average Transaction Confirmation Time: average time taken for transactions to be accepted into a block.
Bitcoin Total Transaction Fees: the total BTC value of transaction fees miners earn per day.
Bitcoin Cost Per Transaction: miners' revenue divided by the number of transactions
Bitcoin Blockchain Size: the total size of all block headers and transactions.
Bitcoin Miners Revenue: (number of bitcoins mined per day + transaction fees)* market price.
bitcoin Hash Rate: the estimated number of giga hashes per second the bitcoin network is performing.
Bitcoin Network Deficit: difference between transaction fees and cost of bitcoin mining.
Bitcoin Days Destroyed: "Days destroyed" is a statistical measure of the adoption of Bitcoin; it tries to encompass both the depth (size) and the width (spread) of transaction activity in Bitcoin[2].
In addition, considering the prominent performance of processed data in analyzing the data better than raw data[3], we added three more features as technical indicators.
Adaptive Moving Average(AMA):used for constructing a moving average with low sensitivity to price series noises and is characterized by the minimal lag for trend detection[4].
Simple Moving Average(SMA):the unweighted mean of the previous n data.[5]
Standard Deviations
As our target is to build the neural network and predict the Bitcoin price, we set our ground truth to be the price of the next day's price. It's as long as the feature vector number.
(2)Dataset and source
We got data via two methods: API and open source dataset.
The feature vectors of our neural network cover from 01/03/2009 to 03/22/2017, and each vector represents one single day, and construct a 2999*17 matrix.

4. Social Media

4.1 source

Apart from the financial data, we also put social media into consideration. For the reason that bitcoin price is strongly related to people's attention, social media is a significant source to forecast bitcoin price. For large social media platform like Twitter or Facebook, it' hard to get the tweets from the past few years, instead, we

can only get access to the real-time sent tweets. However, for some bitcoin forums, we can get all the discussions from any time through crawling. As a result, we can train the model by adding another dimension of data includes the social media factors through forum discussions, and make the prediction using the real-time social media data.

4.2 evaluation

The way that we evaluate social media is sentiment analysis. We use a package named Textblob to do the sentiment analysis. The textblob sentiment module contains two sentiment analysis implementations, PatternAnalyzer and NaiveBayesAnalyzer. We choose the latter one which is an NLTK classifier trained on a movie reviews corpus. The input is a sentence and output is a positive parameter and a negative parameter which indicate the degree of positive sentiment and negative sentiment. The result is the positive parameter divides the negative parameter. If the result is larger than 1, then we think this sentence represents a positive attitude while the result is smaller than 1 represents a negative attitude. For example, if we type the sentence as sentence='the bitcoin will grow', the result will be

```
Sentiment(classification='pos', p_pos=0.6336912790450064, p_neg=0.3663087209549934)
The sentiment result:1.729937735014682
```

4.3 weights of different sources

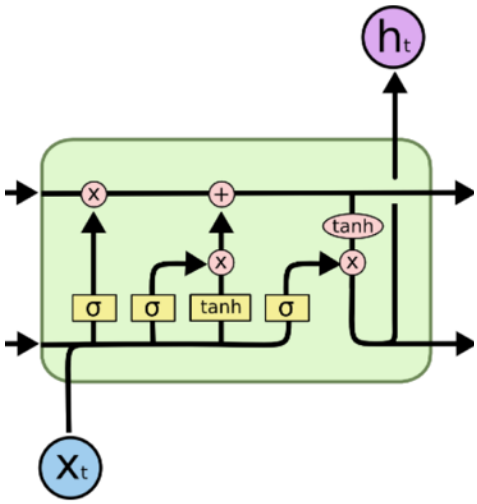
Also, we can grant different social media platform with different weight, for example we can give bitcoin forum larger weight for the reason that it's more professional than twitter or Facebook.

5. Recurrent Neural Network Methodology

We use the method of recurrent neural network to do the model training and prediction. The data we use are the normalized financial data plus social media data, and the ground truth is daily bitcoin price. We use the data of past 30 days to do the training, validation and test. Especially, when we focus on more sophisticated units that implement a gating mechanism, such as a long short-term memory (LSTM) unit and a recently proposed gated recurrent unit (GRU).

5.1 Long-Short Term Memory

Long Short Term Memory networks – usually just called "LSTMs" – are a special kind of RNN, capable of learning long-term dependencies. They were introduced by Hochreiter & Schmidhuber (1997), and were refined and popularized by many people in following work.1 They work tremendously well on a large variety of problems, and are now widely used.



5.1.1 Tensorflw

TensorFlow is an open source software library for machine learning across a range of tasks, and developed by Google to meet their needs for systems capable of building and training neural networks to detect and decipher patterns and correlations, analogous to the learning and reasoning which humans use. We use TFlearn, a deep learning library featuring a high-level API for Tensorflow which supports LSTM.

Reference

[1]wiki-Bitcoin:https://en.wikipedia.org/wiki/Bitcoin
[2]quandl-bitcoin-data:https://www.quandl.com/collections/markets/bitcoin-data
[3]McNally, Sean. Predicting the price of Bitcoin using Machine Learning. Diss. Dublin, National College of Ireland, 2016.
[4]https://www.metatrader5.com/en/terminal/help/indicators/trend_indicators/ama
[5]https://en.wikipedia.org/wiki/Moving_average#Simple_moving_average
[6]https://blockchain.info
[7]http://www.coindesk.com
[8]https://github.com/brandonrobertz/BitcoinTradingAlgorithmToolkit/blob/master/README.md
[9]https://forum.bitcoin.com
[10]https://bitcointalk.org
[11]https://en.wikipedia.org/wiki/Bitcoin