Machine Learning Engineer Nanodegree Capstone Report

Daniel Chao Zhou

31 December 2019

1 Definition

1.1 Project Overview

The stock market prediction has been identified as a significant practical problem in the economic field. Trading algorithms rather than humans performed over 80% of trading in the stock market and the FOREX market. In the crypto-currency market, algorithmic trading is also a hot topic among investors. However, timely and accurate prediction of the market is generally regarded as one of the most challenging problems, since the environment is profoundly affected by volatile political-economic factors, such as legislative acts, political unrest, and mass irrational panic.

There are many studies regarding algorithmic trading in financial markets based on machine learning, where recurrent neural network (RNN) and reinforcement learning (RL) are being popular in recent years. In this study, a Bitcoin price predictor based on long short-term memory (LSTM, a variant of RNN) is presented.

1.2 Problem Statement

Given a time-series data of Bitcoin price with each time step indicating one minute, the goal is to build a price predictor for the price of the next minute.

In this study, data of the BitMEX's XBTZ19 contract, which is a Bitcoin futures contract expiring in December 2019, is used to train and test the predictor. The dataset is a table that each row indicates one minute and each column indicates a specific data such as open-high-low-close prices, volume, and volume-weighted average price (VWAP). The dataset is formulated as $\{x_t | t = 1, 2, \ldots, T\}$, where x_t is a vector of the data in minute t, such that

$$x_t := \begin{bmatrix} \text{open price of time } t \\ \text{high price of time } t \\ \text{low price of time } t \\ \text{close price of time } t \\ \text{volume of time } t \\ \text{VWAP of time } t \\ \vdots \end{bmatrix}.$$

Given data of the time range with length of N and end at time step t, viz $\{x_{i-N+1}, \ldots, x_i\}$, the predictor estimates the VWAP of the next time step $y_i = v_{i+1}$, where v_{i+1} is an item of vector x_{i+1}. The predictor can be seen as a function $\hat{y}_i = f(x_{i-N+1} \ldots x_i)$ mapping from the history trading data to the next VWAP.

Two predictors, an LSTM model as solution and a linear model as benchmark will be built and compared with the metrics as discussed as follows.

1.3 Metrics

The mean squared error (MSE) between labels y and predictions $\hat{y}$ will be used to evaluate the performance of both of the benchmark model and the solution model. For a given N and a time-series test dataset, all consecutive sub-sequence of the time-series with length N will be used and equally contribute to the final MSE.

2 Analysis