Machine Learning Engineer Nanodegree

Capstone Proposal

Zilvinas Marozas
June 19th, 2018

Proposal

Domain Background

Predicting financial instruments prices using machine learning has been attempted by many researchers with varying degree of success. Predicting Commodity Futures market is very important not only for speculators, money managers, etc. It is vital for many companies (food producers, oil refineries, farmers) to predict these prices with certain degree of accuracy since very survival of the company and many jobs could depend on this prediction. Therefore, Commodities Futures market is very important to US economy and is highly regulated by US government. The Commitment of Traders (COT) Report is conducted by the Commodity Futures Trading Commission (CFTC) detailing the open interest in each futures and options on commodities markets containing 20 or more traders holding position sizes large enough to meet the CFTC's reporting level. The purpose of this report is to provide traders with transparency in regards to the open interest in various futures markets and the sizes of those positions for different groups of traders. Open interest is the total number of open futures or options positions that are not closed or delivered on in a particular day.

The COT report is separated into five categories to differentiate the **types of traders** who have positions in a market, and three sub-categories to differentiate the **types of positions**. The COT Report also breaks down four **statistical** categories.

Trader Categories

- **Processors/Users** – Processors or users are traders who use the futures markets as a hedge to the physical commodity in the cash market. These traders include producers of the commodity, consume mass quantities of the commodity or trade the commodity in the cash market.

- **Swap Dealers** – Swap dealers, as defined by the CFTC, is "an entity that deals primarily in swaps for a commodity and uses the futures markets to manage or hedge the risk associated with those swaps transactions." In other words, swap dealers are typically traders who take the other side of trades for hedgers and large speculators.

- **Managed Money** – These traders are Commodity Trading Advisors (CTAs), Commodity Pool Operators, or hedge funds that hold large speculative positions in the futures.

- **Other Reportable** – Traders outside of the three categories listed above that have substantial positions, as defined by the CFTC, in a market.

- **Non-Reportable** – Traders outside of the first three categories listed above that have small positions, as defined by the CFTC, in a market.

Position Categories

- **Long** – This is the number of traders that are long the market.
- **Short** – This is the number of traders that are short the market.
- **Spreading** – This represents the number of offsetting positions in a particular trader category.

Problem Statement

In this project we will try to predict closing weekly price of Corn Commodity Futures. In order to perform this prediction we will create a dataset that includes weekly Corn Futures closing prices as well as Long Open Interest and Short Open Interest of Processors/Users(sometimes they are called Commercials) from COT reports and by using this dataset we will try to predict next week's prices. Why do I think this method will work. I do believe that using only historical prices alone is not sufficient to predict prices of commodities and especially it is very hard to predict turning points, when commodity reverses the previous trend and starts moving to the opposite direction. Therefore, researcher needs to use "Fundamentals data" if she wants to predict these prices with some degree of consistency. However, obtaining "Grains Fundamentals Data" (area planted, area harvested, exports, etc.) is difficult and expensive and it will make machine learning model much more complicated. My hypothesis is that we can substitute "Fundamentals Data" with data from COT report by tracking open interest(long open interest and short open interest) of Processors/Users. This group of

traders represents well-funded enterprises that have "deep pockets" in order to conduct research and also have inside information since they are typically large users of grains.

Datasets and Inputs

Up until recently it was rather difficult and expensive to obtain consistent futures data across exchanges in frequently updated manner. That has all changed recently with the release of Quandl. It's a fantastic source of free financial data. They have a substantial library of daily futures prices in some cases going back to the 1950s!

We will use Quandl for both Corn Futures prices as well as for Commitments of Traders (COT) data even though this data could be obtained directly from CFTC web site (https://www.cftc.gov/MarketReports/CommitmentsofTraders/index.htm).

Quandl already scrubbed and cleaned COT data and provides it as easily downloadable CSV file.

Here are descriptions and links to these datasets:

Historical Futures Prices: Corn Futures, Continuous Contract #1. Non-adjusted price based on spot-month continuous contract calculations. Raw data from CME:

https://www.quandl.com/data/CHRIS/CME_C1-Corn-Futures-Continuous-Contract-1-C1-Front-Month

Commitment of Traders - CORN (CBT) - Futures Only (002602)

https://www.quandl.com/data/CFTC/002602_F_ALL-Commitment-of-Traders-CORN-CBT-Futures-Only-002602

Solution Statement

In this project we will try to predict sequence. Sequence prediction is different to other types of supervised learning problems. The sequence imposes an order on the observations that must be preserved when training models and making predictions. The Long Short-Term Memory network or LSTM is a recurrent neural network that can learn and forecast long sequences. It is internally composed by two kind of units, hidden units, which as in CNNs contain a hidden representation of the inputs, and gated units,

which control the amount of information that flows from the inputs, to the hidden units and to the outputs.

Benchmark Model

A benchmark in forecast performance provides a point of comparison If a model achieves performance at or below the benchmark, the technique should be fixed or abandoned. Three properties of a good technique for making a benchmark forecast are:

- Simple: A method that requires little or no training or intelligence.
- Fast: A method that is fast to implement and computationally trivial to make a prediction.
- Repeatable: A method that is deterministic, meaning that it produces an expected output given the same input.

A common algorithm used in establishing a baseline performance for time series forecasting is the persistence algorithm. The persistence algorithm uses the value at the current time step (t) to predict the expected outcome at the next time step (t+1). This satisfies the three above conditions for a baseline forecast. Due to simplicity of benchmark model it will take into account only the price and will disregard other feature like Open Interest from COT report.

Evaluation Metrics

After the model is fit, we can forecast for the entire test dataset. We combine the forecast with the validation dataset and invert the scaling. We also invert scaling on the validation dataset with the expected weekly corn futures closing prices.

With forecasts and actual values in their original scale, we can then calculate an error score for the model. In this case, we calculate the Root Mean Squared Error (RMSE) that gives error in the same units as the variable itself.

Benchmark model we will evaluate on the validation dataset. We do this using the walk-forward validation method. In essence, we step through the validation dataset time step by time step and get predictions. Once predictions are made for each time step in the validation dataset, they are compared to the expected values and a Root Mean Squared Error (RMSE) score is calculated. Once we have RMSE for both models we will compare them and decide whether or not LSTM model is good choice for our problem.

Project Design

The first step is to prepare the dataset for the LSTM. This involves framing the dataset as a supervised learning problem and normalizing the input variables. We will frame the supervised learning problem as predicting the Corn Futures price for the current week given the price, Long Open Interest and Short Open Interest of Processors/Users at the prior week.

Data Preparation

We will obtain data from Quandl. Since, we only have COT weekly data we will resample the frequency of the Corn futures pricing data and turn it into weekly data.

Then we will combine two data sets into one data set and drop unnecessary columns

Data Normalization

Normalization is a rescaling of the data from the original range so that all values are within the range of 0 and 1. Normalization requires that we are able to accurately estimate the minimum and maximum observable values. The good thing about commodities prices that they tend to fluctuate within certain range during the long run(may trend up or tend down for short term or medium term) , where other securities (e.g. stocks) tend to appreciate over time and can experience large upward trends, where maximum price could be very difficult to estimate.

Data Splitting

We will split data into three datasets

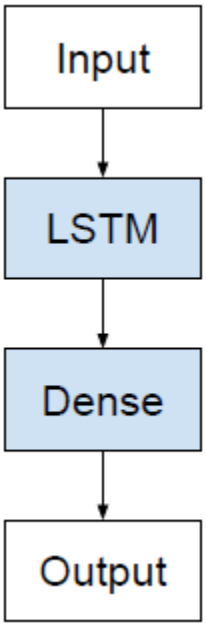
Training data: 06/16/2006- 12/31/2016

Testing data: 01/01/2017-12/31/2017

Validation data 01/01/2018 -

Model

We will use relatively simple LSTM model.



We will use the Mean Absolute Error (MAE) loss function and the efficient Adam version of stochastic gradient descent.

References

1) Williams, L. (2013). Trade stocks and commodities with the insiders. Hoboken, N.J.: Wiley. http://a.co/6qkXmqN
2) https://machinelearningmastery.com/multivariate-time-series-forecasting-lstms-keras/
3) Galit Shmueli and Kenneth Lichtendahl, Practical Time Series Forecasting with R: A Hands-On Guide, 2016. http://amzn.to/2k3QpuV
4) https://machinelearningmastery.com/lstms-with-python/