Machine Learning Engineer Nanodegree

Capstone Proposal

Masaharu KINOSHITA
email: k.masaharu0219@gmail.com
SNS: LinkedIn
Apr, 2019

Abstract

In this proposal, in order to verify how useful CNN is to solve time-series prediction problem, CNN, LSTM, and CNN+LSTM are build on stock datasets of Google obtained at kaggle. As you know, CNN is mainly used in the field of Image Recognition. CNN, however, is said that it has a potential to solve time-series forecasting problem. In order to show that, CNN, LSTM, and CNN+LSTM models are build on the google stock datasets and their score on the test datasets are compared with benchmark score of RNN, which is often used for time-series data, with MSE.

Agenda

6. Evaluation Metrics

1. Domain Background

What is Algorithm Trading?
In the field of finance, trading using algorithms is said *Algorithm Trading*. This is a method of executing a large order using automated pre-programmed trading instructions accounting for variables such as time, price, and volume.
Traditionally, time series forecasting including Algorithm Trading has been dominated by linear

methods, such as AR, ARIMA etc, because they has well-interpretability on many simpler forecasting problems. They, however, have less predictive ability on more complex problems. On the other hand, deep learning neural networks has potential to automatically learn arbitrary complex mappings from inputs to outputs and support multiple inputs and outputs. Therefore, employing deep learning model is possible to well forecast from its time-series data with advance in computing performance. In fact, it is reported that RNN, LSTM, CNN can be valid approach to predict stock prices.

What is my motivation about Algorithm Trading with deep learning?
My motivation to tackle with this proposal is to acquire skills and tacit knowledge to build network by myself. Now, I'm working as a newly-fladged Data Scientist in Tokyo and I got opportunity to build deep learning models to detect car accident from its acceleration data at the previous project. So, in order to develop my skills more and get opportunity to get more exciting projects at my company, it is necessary to brush up my skills more, especially about deep learning and so on. That's why I'd propose this paper.

2. Problem Statement

In this proposal, usability of deep learning, especially CNN as an feature extractor, is verified. Although CNN is known to be valid in the field of Image Recognition, few use-case of CNN are applied to finance problem, such as stock price predictions. This is because a lot of Algorithm Trading has employed technical index so far. These index, however, are commonly used and developed by humans. So, it can be said that there is some room to improve Trading Algorithm.
In this context, applying CNN to the finance problem and validation of its usefulness is meaningful as CNN has high potential to recognize patterns in given dataset and computational power has advanced so far.
In order to valid the usefulness of CNN, LSTM and CNN+LSTM are compared to the stock price predictions with metrics MSE. In addition to this, RNN is set as base-models. By comparing the four models with MSE, the usefulness of CNN are verified in the stock price problem.

Datasets and Inputs

In this proposal, deep learning models are trained and tested on the stock of Google. The datasets contains the following two csv files.
1. train.csv

- number of rows: 780
- number of columns: 6

2. test.csv

- number of rows: 137
- number of columns: 6

columns

- Date
- Open
- High
- Low
- Volume
- Close (target)

The original datasets are provided at kaggle.

Solution Statement

In this section, a solution to the problem is described. Roughly speaking, a solution consists of two part.
** Preprocessing **
First of all, train and test datasets are split into small datasets according to window length and normalized within window.
Next, train dataset is split into train and validation datasets. The validation dataset is used for decide hyper-parametes, such as number of epochs and layers etc. in order to avoid information leak from test dataset. By doing so, generalization performance of build deep learning models can be evaluated.
** Modeling **
After the above preprocessing, RNN, CNN, LSTM, CNN+LSTM models are build on the preprocessed train datasets with target of Close price. After each model is trained, it is tested on the test dataset with MSE.

Benchmark Model

In this problem, RNN model is build to get base MSE as benchmark model. RNN, one of the famous deep learning models, is often used for time-series forecasting. This is an usual score with conventional method employing deep learning. As mentioned above, the metrics with which the benchmark model is measured is also MSE.

Evaluation Metrics

As mentioned above, MSE is evaluation metrics. Needless to say, less MSE is better for stock price prediction. The reasons of employing MSE in this problem are the followings.

First, the target value, which is daily close stock price, is continuous. So, this is regression problem.
Second, more penalty is added to larger error with MSE compared to MAE by employing squared value.
Therefore, MSE is employed as evaluation metrics.

Project Design

In this final section, a workflow for approaching a solution is summarized. In order to verify the useful of CNN in the stock price prediction problem, the 4 models of RNN, CNN, LSTM, and CNN+LSTM are build on the train and test dataset with metrics MSE. Google stock datasets on kaggle are used. RNN, a benchmark model, and the other models are build on the train datasets from 2014-03-27 to 2017-05-01. After training, their MSE score on the test dataset, from 2017-05-01 to 2017-11-10, are compared with the others.
By doing so, how useful CNN is to time-series forecasting are verified.