Capstone Proposal: Predict Stock Market Return
(Machine Learning Engineer Nanodegree)

Introduction

The proposed problem for this capstone is derived from Kaggle, the Winton Stock Market Challenge [1]. From Wikipedia [2] – "Kaggle is platform for predictive modelling and analytics competitions on which companies and researchers post their data and statisticians and data miners from all over the world compete to produce the best models".

The main objective in choosing this competition is to gain more understanding on time-series related problem and also to be able to solve it using recurrent neural networks – a domain of which the author has not had many opportunities to work with.

Problem Statement

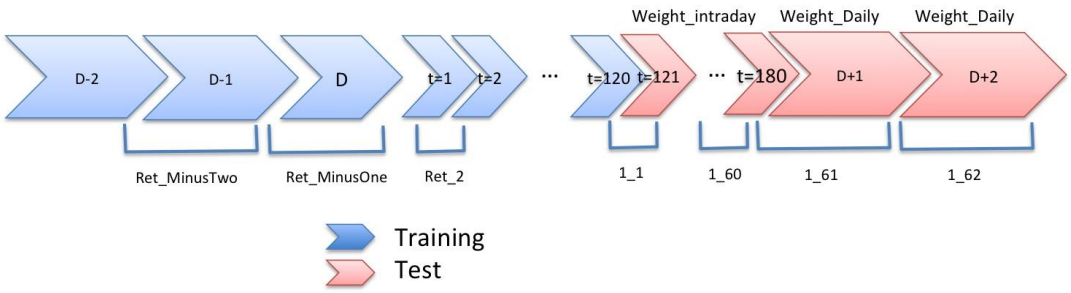
The competition requires that competitors determine the intra and end of day returns of a stock market, given historical stock performance and a host of other features which are masked.

Traditionally, these are the same set of problems being worked on by quantitative analysts in financial firms on a day to day basis – thus it reflects the problem normally faced by the industry.

Datasets

To solve the given problem of predicting the future, we are given historical data of the past few days. The datasets are available from Kaggle's website [3].

We are given 5 days in total, days D-2, D-1, D, D+1, and D+2 respectively.



The data in D-2, D-1 and part of day D will be used for training, while the rest of remaining days are to be used as test.

Return data are available only for training datasets (D-2,D-1 and part of day D).

In total there are 25 features provided for the whole 5 day duration to be used in modelling.

Below are description on the files available for the problem.

- train.csv:
    - Feature_1 - Feature_25
    - Ret_MinusTwo, Ret_MinusOne
    - Ret_2 - Ret_120
    - Ret_121 - Ret_180: **target variables**
    - Ret_PlusOne, Ret_PlusTwo: **target variables**
    - Weight_Intraday, Weight_Daily
- test.csv:
    - Feature_1 - Feature_25
    - Ret_MinusTwo, Ret_MinusOne
    - Ret_2 - Ret_120
- sample_submission.csv

The last file mentioned earlier is to be used as reference for the output file that will need to submitted to Kaggle for scoring purposes.

Below are descriptions of the features available in the train and test dataset.

1. **Feature_1 to Feature_25:** different features relevant to prediction
2. **Ret_MinusTwo:** this is the return from the close of trading on day D-2 to the close of trading on day D-1 (i.e. 1 day)
3. **Ret_MinusOne:** this is the return from the close of trading on day D-1 to the point at which the intraday returns start on day D (approximately 1/2 day)
4. **Ret_2 to Ret_120**: these are returns over approximately one minute on day D. Ret_2 is the return between t=1 and t=2.
5. **Ret_121 to Ret_180**: intraday returns over approximately one minute on day D. ***target variables.**
6. **Ret_PlusOne**: this is the return from the time Ret_180 is measured on day D to the close of trading on day D+1. (approximately 1 day). ***target variables.**
7. **Ret_PlusTwo**: this is the return from the close of trading on day D+1 to the close of trading on day D+2 (i.e. 1 day) ***target variables.**
8. **Weight_Intraday**: weight used to evaluate intraday return predictions Ret 121 to 180
9. **Weight_Daily**: weight used to evaluate daily return predictions (Ret_PlusOne and Ret_PlusTwo).

Solution Statement

The solution will contain the predicted results (ie return) for each of the stocks provided in the dataset. The derived solution will be submitted to Kaggle, as it will evaluated and scored when compared against the hold-out dataset.

Benchmark Criteria

For benchmarking purposes, we compare our model performance against the scenario of zero prediction. A zero prediction is a situation when returns have no changes over time period.

Evaluation Metrics

Each submission in Kaggle will be evaluated using the Weighted Mean Absolute Error. As such, we will also be using the same metrics to evaluate our model while offline. Each model predicted return is compared with the actual return. The formula is then

$$WMAE = \frac{1}{n}\sum_{i=1}^{n} w_i \cdot |y_i - \hat{y}_i|,$$

where w_i is the weight associated with the return, Weight_Intraday, Weight_Daily for intraday and daily returns, i_i, y_i is the predicted return, $\hat{y}_i$ is the actual return, and n is the number of predictions.

The weights for the training set are given in the training data, while the weights for the test set are unknown.

Method and Design

The modelling methodology used will be based on the following workflow:

1. Data preparation and general cleansing
2. Feature selection/Dimension reduction
3. Split data into train and validation
4. Model building
5. Evaluation and optimization

Since the data is in time series, one of the classifiers that we could take is by using linear regression. However, without prior knowledge of the features – this could limit the performance. As such the proposed algorithm to be used is recurrent neural networks, so that we can identify hidden features that are important, while also being able to work with time series. [4]

References

[1]: https://www.kaggle.com/c/the-winton-stock-market-challenge

[2]: https://en.wikipedia.org/wiki/Kaggle

[3]: https://www.kaggle.com/c/the-winton-stock-market-challenge/data

[4]: http://machinelearningmastery.com/time-series-prediction-lstm-recurrent-neural-networks-python-keras/