Capstone Proposal

Dennis P. F.
December 11th, 2016

Proposal

Domain Background

Stock price prediction is a very basic and important operation in financial entities like hedge funds, mutual funds and exchange traded funds. Performance of these financial institutions depend highly on the degree of success of this prediction. Many works in the past have used machine learning techniques like Linear Discriminant Analysis(LDA), K-nearest neighbor(KNN), Gaussian Process, Support Vector Machines(SVM), Neural Networks(ANN), Random Forest and Naive Bayes[1][2]. Although minimizing the error of prediction is important, the confidence score of the prediction made is also important. The speed of making the prediction is a critical factor in successfully leveraging the prediction made in the market.

Problem Statement

This work attempts to make a 5-day forecast of adjusted-close[3] prices of N current stocks of S&P 500 Index which were first added before 2006-01-01 purely based on Technical Analysis features like past m day's Simple Moving Average(SMA), Rolling Standard Deviation(RSD), Bollinger-Bands(BB)[4] and Momentum of the adjusted close prices and also based on the daily Volume of each stock. Mathematically the aim can be written as :

Estimate the functions $\mathbf{F}_i$ for all i such that

$$\mathbf{P}_{\mathbf{i\text{-}Forecast}} = \mathbf{F}_i(\mathbf{SMA}_i, \mathbf{RSD}_i, \mathbf{BBscore}_i, \mathbf{Mscore}_i, \mathbf{Volume}_i)$$

where

$$\mathbf{P}_{\mathbf{i\text{-}Forecast}} = [P_{i,t+1}, P_{i,t+2}, P_{i,t+3}, P_{i,t+4}, P_{i,t+5}]$$

$$\mathbf{SMA}_i = [SMA_{i,t}, SMA_{i,t-1}, SMA_{i,t-2}, \ldots SMA_{i,t-m-1}]$$

$$\mathbf{RSD}_i = [RSD_{i,t}, RSD_{i,t-1}, RSD_{i,t-2}, \ldots RSD_{i,t-m-1}]$$

$$\mathbf{BBscore}_i = [BBscore_{i,t}, BBscore_{i,t-1}, BBscore_{i,t-2}, \ldots BBscore_{i,t-m-1}]$$

$$\mathbf{Mscore}_i = [Mscore_{i,t}, Mscore_{i,t-1}, Mscore_{i,t-2}, \ldots Mscore_{i,t-m-1}]$$

Volume$_i$ = [Volume$_{i,t}$, Volume$_{i,t-1}$, Volume$_{i,t-2}$, ... Volume$_{i,t-m-1}$]

$P_{i,t}$ is the t^{th} day's adjusted-close price of i^{th} stock of the selected S&P 500 list sorted alphabetically by ticker.

$SMA_{i,t}$ is given by $\frac{1}{3}\sum_{j=t-2}^{t} P_{i,j}$ which is the 3-day moving average of i^{th} stock of S&P 500.

$RSD_{i,t}$ is given by $\sqrt{\frac{1}{3}\sum_{j=t-2}^{t} (P_{i,j} - SMA_{i,t})^2}$ which is the 3-day rolling standard deviation of i^{th} stock of S&P 500.

$BBscore_{i,t}$ is given by $\frac{P_{i,t} - SMA_{i,t}}{2\,RSD_{i,t}}$ which is the Bollinger band score of i^{th} stock of S&P 500 as introduced in [5] .

$Mscore_{i,t}$ is given by $\frac{P_{i,t}}{P_{i,t-3}}$ which is the momentum score for i^{th} stock of S&P 500 as introduced in [5].

$Volume_{i,t}$ is the t^{th} day's volume of the i^{th} stock of S&P 500.

Datasets and Inputs

The project needs a list of companies(tickers) in S&P 500 index that were first added before 2006-01-01. This information can be programmatically obtained from [6]. Now that the list of tickers are retrieved, for each ticker in the list, the daily adjusted closing prices and volume for the date range [2006-01-01 to 2016-12-10] can be obtained programmatically in csv format using the URL template:

```
http://chart.finance.yahoo.com/table.csv?
s=<TICKER>&a=0&b=1&c=2006&d=11&e=10&f=2016&g=d&ignore=.csv
```

For each ticker there would be 2755 rows as there are 2755 business days in the selected date range. As per the problem statement, the features required are SMA, RSD, BBscore, Mscore and Volume of which Volume can be used directly from the downloaded data.

The rest of the features can be calculated from adjusted closing price present in the downloaded data files.

Notations used

$X_{i,p:q}$ is used to represent the features of i^{th} stock from timestamp p to q-1, the features being, $SMA_{i,p:q}$, $RSD_{i,p:q}$, $BBscore_{i,p:q}$, $Mscore_{i,p:q}$, $Volume_{i,p:q}$ and $Y_{i,p:q}$ denote $P_{i,p:q}$, the target adjusted closing prices for stock i.

Total number of points in the time-series is denoted as S.

Number days to look back for predicting 5 future dates is denoted as m.

Number of features per day is denoted as n. Here n = 5 as there are 5 features namely SMA, RSD, BBscore, Mscore and Volume per day.

Number of stocks selected from S&P 500 is denoted as N.

Benchmark Model

Linear regression is used to create the benchmark model. The problem to be solved is a classic regression style problem in machine learning where a continuous number(adjusted closing price of future date) is required to be predicted from a set of features and Linear regression is the most simplest non-trivial choice. Since the problem to be solved needs to predict adjusted close price for 5 future dates for all of the selected N stocks, 5*N linear regressors are needed.

Due to the time-series nature of the data, evaluation of the model is done in a number of "trials". Each trial consists of certain number of train examples and a single test example.

A general Q^{th}trial consists of:

(S-Q-m-9) train points : {

$$(X_{i,0 : m}, Y_{i,m : m+5}),$$
$$(X_{i,1 : m+1}, Y_{i,m+1 : m+5+1}),$$
$$(X_{i,2 : m+2}, Y_{i,m+2 : m+5+2}),$$
$$\ldots$$
$$(X_{i,S-Q-m-10 : S-Q-10}, Y_{i,S-Q-10 : S-Q-5})$$

}

and the single test point is

$$(X_{i,S-Q-m-5 : S-Q-5}, Y_{i,S-Q-5 : S-Q})$$

For each trial Q, the model is trained on S-Q-m-9 train points and evaluated on the trial's sole test point. The total number of trials is limited by the minimum number of training points

required for the last trial. Since a single regressor in the model has m*n features, minimum number of train samples is set as 3*m*n, which implies that there can be a maximum of S − m − 9 − 3*m*n trials. For S = 2755, m = 5, n = 5, the maximum number of trials is 2666.

Since there are 5*N classic regressors in the model, there are 5*N predicted values per each evaluation of the model. The evaluation metric is calculated separately for each of 5*N regressors, then averaged to get a single number.

Test performance of the model = $\frac{1}{5N}\sum_{i=0}^{N-1}\sum_{j=0}^{4} EvalMetric_Q(\hat{Y}_{i,j,Q}, Y_{i,j,Q})$

where j is the future day index, i is the index of stock and EvalMetric() is the evaluation metric used. $\hat{Y}$ and Y are the predicted target values and ground truths respectively.

Solution Statement

The project aims to solve the problem described in the problem statement in four steps *for each trial* :

1. Learn 5*N AdaBoost Regressors with the benchmark model as the base weak learner using validation data points to create model-1.

2. Learn 5*N AdaBoost Regressors with KNN regressor as the base learner to create model-2.

3. Learn 5*N Linear Regressors that takes the outputs of model-1 and model-2 as input and ground truth as target.

4. Tune parameters of model-1 (learning rate) and model-2 (learning rate, number of neighbors) using error metric of evaluation of model-3 on validation data points. The output of the model-3 is considered as final and is used to evaluate on test data.

Evaluation Metrics

A suitable evaluation metric for the project is R^2 score or coefficient of determination[7] since it is a regression problem. It reflects how well future samples are likely to be predicted by the model. It produces a score in the range (-Inf, 1], score = 1 indicate best performance and closer to zero and below indicate poor performance. A model that just returns the mean of target variables as output will have 0 R^2 score. R^2 evaluation metric takes estimated target samples $\hat{Y}$ and corresponding ground truths Y and computes the score as :

$$R^2(\hat{Y}, Y) = 1 - \frac{\sum_{k=0}^{M-1}(Y_k - \hat{Y}_k)^2}{\sum_{k=0}^{M-1}(Y_k - \bar{Y}_k)^2}$$

Where M = number of samples of Y and $\hat{Y}$ and $\bar{Y}$ is the mean of Y over its samples.

Project Design

After getting data as described in Datasets and Inputs section, the features are normalized using

$$normalizedFeature = \frac{feature - mean(feature)}{standardDeviation(feature)}$$

Then different trials are done on the normalized data. Each trial consists of a single test data point and many train points obtained from data with timestamps older than the test point. The train and test data for a general Qth trial is given by :

(S-Q-m-9) train points : {

$$(X_{i,0:m}, Y_{i,m:m+5}),$$
$$(X_{i,1:m+1}, Y_{i,m+1:m+5+1}),$$
$$(X_{i,2:m+2}, Y_{i,m+2:m+5+2}),$$
$$\ldots$$
$$(X_{i,S-Q-m-10:S-Q-10}, Y_{i,S-Q-10:S-Q-5})$$

}

denoted compactly as $(Xtrain_i^{(Q)}, Ytrain_i^{(Q)})$ with number of points denoted by $Ntrain_Q$.

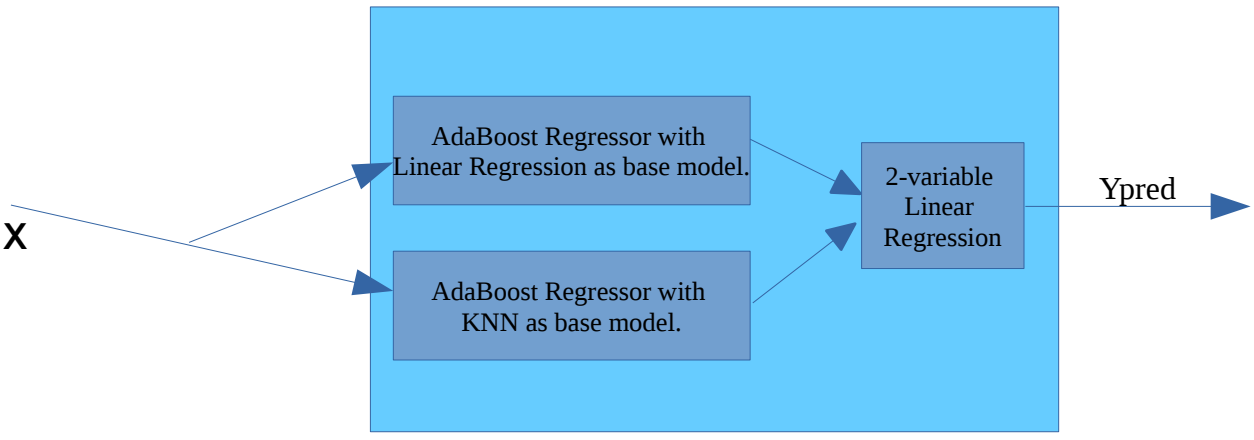
and the single test point is

$(X_{i,S-Q-m-5:S-Q-5}, Y_{i,S-Q-5:S-Q})$ compactly written as $(Xtest_i^{(Q)}, Ytest_i^{(Q)})$

The maximum value of the index Q is set as Q_{max} = S – m – 9 – 3*m*n as described in the benchmark model section.

For each such trial, models described in the *solution statement* are trained and tuned using a certain number of *validation trials* using $(Xtrain_i^{(Q)}, Ytrain_i^{(Q)})$.

Fig 1. Proposed regression model for a single stock for single future day adjusted close price prediction



A general R^{th} validation trial consists of :

train points : {

$(Xtrain_{i,\,0}^{(Q)}, Ytrain_{i,\,0}^{(Q)}),$

$(Xtrain_{i,\,1}^{(Q)}, Ytrain_{i,\,1}^{(Q)}),$

...

$(Xtrain_{i,\,NtrainQ-R-2}^{(Q)}, Ytrain_{i,\,NtrainQ-2}^{(Q)})$

}

compactly denoted as $Dtrain^{(Q,R)}$

and *single* validation point: $(Xtrain_{i,\,NtrainQ-R-1}^{(Q)}, Ytrain_{i,\,NtrainQ-1}^{(Q)})$ denoted compactly as $Dvalid^{(Q,R)}$

The number of validation trials are limited by $Ntrain_{Q.}$ By letting the minimum number of training points required in a validation trial to 3*m*n, the maximum value of R becomes R_{max} = NtrainQ – 1 – 3*m*n which in-turn puts an upper bound to Q, given by Q_{max} = S – m – 11 – 3*m*n.

The model has two main parameters to be tuned a) learning rates and b) number of nearest neighbors using validation trials. For a given setting of parameters, the model predicts 5*N real values for a single validation point after training on $Dtrain^{(Q,R)}$ and this is repeated for all validation trials.

Validation score per each parameter setting of the model is given by :

Validation score = $\frac{1}{5N}\sum_{i=0}^{N-1}\sum_{j=0}^{4} EvalMetric_R(\hat{Y}_{i,j,R},\ Y_{i,j,R})$

where R is the validation trial index, $\hat{Y}_{i,j,R}$ and $Y_{i,j,R}$ are the model predicted outputs for validation point R and actual target of the validation point R respectively. i and j are the stock and future day indices. Model parameters are chosen based on this validation score. Then the tuned model is trained on whole of $(Xtrain_i^{(Q)}, Ytrain_i^{(Q)})$ and evaluated on $(Xtest_i^{(Q)}, Ytest_i^{(Q)})$. This whole process is then repeated for every test trial Q and Test score is calculated using :

Test score = $\frac{1}{5N}\sum_{i=0}^{N-1}\sum_{j=0}^{4} EvalMetric_Q(Ypred_{i,j,Q},\ Ytest_{i,j,Q})$

References

[1] Ou, Phichhang, and Hengshan Wang. "Prediction of stock market index movement by ten data mining techniques." Modern Applied Science 3.12 (2009): 28.

[2] Patel, Jigar, et al. "Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques." Expert Systems with Applications 42.1 (2015): 259-268.

[3] http://www.investopedia.com/terms/a/adjusted_closing_price.asp

[4] https://en.wikipedia.org/wiki/Bollinger_Bands

[5] Machine Learning for Trading Course by Udacity - https://www.udacity.com/course/machine-learning-for-trading--ud501

[6] https://en.wikipedia.org/wiki/List_of_S%26P_500_companies

[7] http://scikit-learn.org/stable/modules/model_evaluation.html#r2-score