Definition (approx. 1-2 pages)

Project Overview

Historically hedge funds have been using market knowledge and intuition to gather signals that are helpful in the prediction of the returns of stocks. Lately an onslaught of evidence has been gathering that machine learning could be a valuable tool in profitable analysis. This brings rise of the machine learning quantitative analyst and this project takes a look into a hypothetical problem within profession.

In this scenario, I am a machine learning engineer working as a quantitative analyst for a hedge fund. Furthermore this hedge fund trades in the S&P 500 index but can use data from other indexes for analysis in these trades. The company provides the machine learning engineer data to use for profitable analysis. More specifically this data contains 10 columns, each one representing a open close prices for a stock ordered according to dates. Each column name is named S1 through S10. More specifically, S1 trades in the united states as part of the S&P 500 Index while stocks S2, S3 through S10 trade in Japan in the Nikkei Index. S1 is missing the latter portion of the dataset in comparison with stocks S2 through S10. More specifically, S1 contains 50 data points, while S2 through S10 contains 100 data points each.

The goal for the company in providing the machine learning engineer this dataset is for the engineer to build a model that forecasts the latter portion of S1, as a function of S2 through S10. The model should predict the latter 50 rows for S1. In addition, fund researchers point out not all of the columns S2 through S10 are important in predicting S1.

Problem Statement

Cross validation will be implemented on the dataset to make sure the testing set is not a statistical anomaly that is generally not representative of the real environment. Furthermore this will assure that we have a more averaged training error that has more potential to represent actual environment testing. For each cross validated training, the dataset will go through dimensionality reduction using principal component analysis. After we have a new dataset which was transformed by principal component analysis, I will apply an ensemble technique. An ensemble is a collection of various machine learning algorithms each of which could be weighted and summed into a final classification or regression solution. The ensemble will contain linear regression or general linear model, logistical regression, and k nearest neighbors. This ensemble will return the model which should predict the actual stock open to close changes for the latter portion of S1. This is the solution to the problem: Predict the 50 latter data points of the stock S1.

Metrics

To check for accuracy or testing error I will be using mean squared error. Basically this metric takes the difference from actual values and predicted values, squares it, and then returns a sum of all these values for each training point. The squaring is done so that negative differences don't cancel positive differences during the summing process. The reason I'm fond of this particular metric is because mean squared error is the second moment of the error. Thus it includes both variance of the estimator and its bias.

Analysis (approx. 2-4 pages)

Data Exploration

A data set is present for this problem. The features are other stock prices. Basically we have 9 columns representing a feature in the input space, and we have one column representing the output. The output column is S1 and the input space contains 'S2 through S10'. For the first fifty points, we have the values for S1, and this subset of the dataset will be used for training. The latter fifty data points, S1 is missing its values. Furthermore I have provided a few rows of the dataset so that the reader may understand better on what we're working with.

	date	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
0	5/30/2014 0:00	0.446531	-0.828505	1.268874	1.468834	1.207883	0.622798	1.821350	-0.699674	1.890508	0.523177
1	6/2/2014 0:00	0.242901	-0.245839	-0.558069	-0.990142	-0.497502	-0.137380	0.158878	0.081083	-1.311177	0.175995
2	6/3/2014 0:00	-0.327468	-0.745564	0.254992	0.322715	0.832613	0.091484	-0.405513	-0.323873	0.920877	-0.066132
3	6/4/2014 0:00	0.098084	-0.636639	-0.840802	0.643013	0.167639	0.017852	0.293241	0.871516	0.201038	-0.166004
4	6/5/2014 0:00	0.127334	-0.710122	1.396653	0.135177	1.888105	0.583958	0.949694	0.428746	1.050444	1.196328

Using the pandas dataframe, I was able to call the describe function which is pasted below.

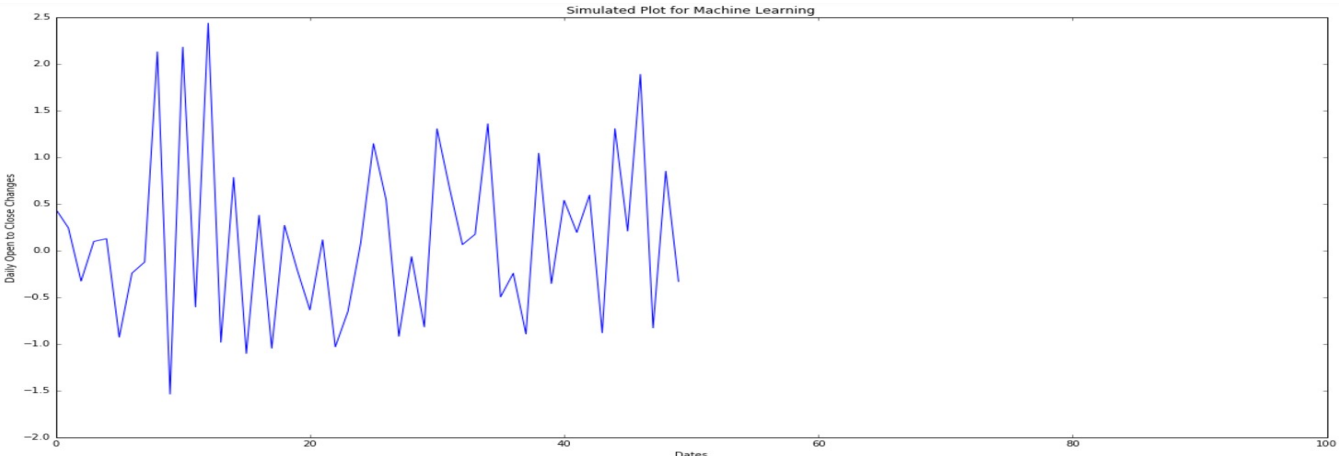
	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
count	50.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000
mean	0.118352	-0.177025	0.239531	0.245163	0.350003	0.119914	0.209739	-0.082404	0.294292	0.014811
std	0.930020	1.105111	2.178191	1.207634	3.198760	1.020103	2.110917	0.767700	2.212564	0.721940
min	-1.537622	-3.318346	-4.817725	-2.048516	-8.426485	-2.506426	-5.439535	-2.365902	-5.180856	-2.108822
25%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
50%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
75%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
max	2.435610	1.979053	6.623951	3.662557	9.696483	2.954939	5.747429	2.114004	5.844546	1.872833

The mean shows the average for each stock in terms of open to close changes per day. All of these are positive except for S2 and S8. S5 seems to be the most profitable stock with an average open to close change of .35 and a maximum daily change of 9.69. In particular S1 has a positive mean with value 0.118352. This mean is most similar to S6. Furthermore it's standard deviation was 0.930020 which is

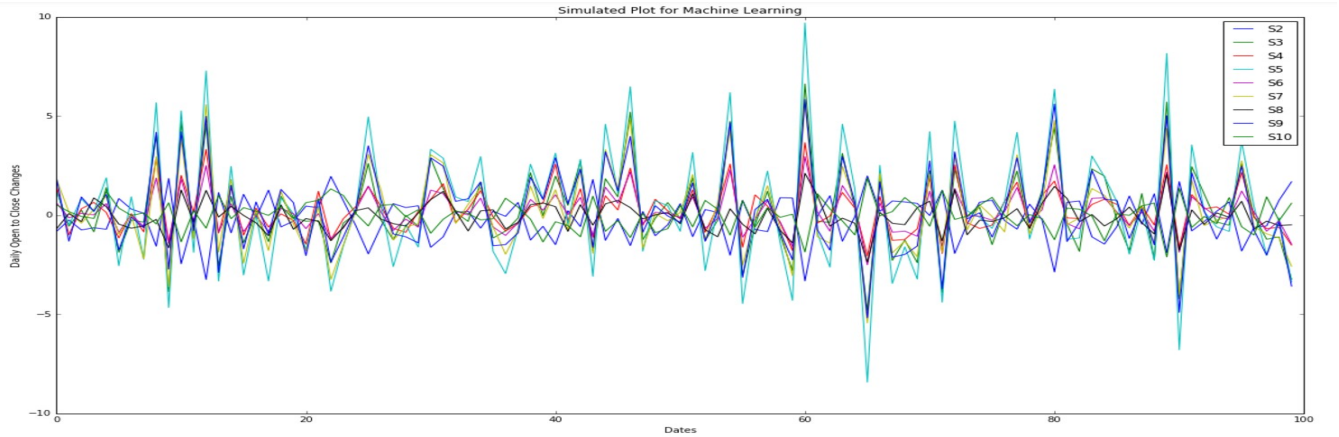
most similar to S6. Simply reading the data, a reader realizes S5 is the most different. It is sporadic in terms of the other stock price changes with a standard deviation of 3.198, while still being the most profitable.

Exploratory Visualization

The S1 data is plotted below



Next the input space is plotted below, albeit with heavy overlap.



Comparing both graphs, it is important to note, that the y axis is scaled differently. Also some stocks shown in the second plot are more sporadic while others are less, in comparison with S1. These graphs show the input space on one end and an output space with what is to be predicted. It gives the data scientist a birds eye view of the data, in order to create models. By analyzing new predictions in

comparison with the input space, one might be able to understand right away that the model did something right or wrong visually.

Algorithms and Techniques

Once more the algorithms I will test in an ensemble are linear regression or general linear model, logistical regression, and k nearest neighbors. The dataset will first be transformed principal component analysis for each cross validated training and testing sets. Principal component analysis is a powerful mathematical tool which reduces dimensionality, so that one knows what new features, from the transformed dataset, brings the most variation. Furthermore, it rearranges the data on new axises, which in fact represents the new features. For example if one of the stock brings little variation to the overall stock price, it could be dropped in a mathematically efficient manner saving both time and energy for the computation while conserving an effective dataset to build the model. The ensemble will be applied with the listed machine learning algorithms using exhaustive parameter tuning with the help of gridsearchcv, an sklearn function. To start I apply the ensemble to 4 of the most valuable pca components. Next pca_inverse transform is used to bring the model dataset back to the original space. Similarly the general linear model, logistical regression and reinforcement learning was applied. A weighted ensemble will be used as the final model.

Benchmark

Unfortunately there's no benchmark to compare the model to an actual result. This capstone project is given by a quantitative analyst firm looking for a machine learning engineer. Although the date for the submission (to the firm) has passed, I find this project

valuable since a real firm uses this to seek a valuable skill set.

Methodology (approx. 3-5 pages)

Data Processing

In this section, all of your preprocessing steps will need to be clearly documented, if any were necessary. From the previous section, any of the abnormalities or characteristics that you identified about the dataset will be addressed and corrected here. Questions to ask yourself when writing this section:

- If the algorithms chosen require preprocessing steps like feature selection or feature transformations, have they been properly documented?
- Based on the **Data Exploration** section, if there were abnormalities or characteristics that needed to be addressed, have they been properly corrected?
- If no preprocessing is needed, has it been made clear why?

Implementation

Refinement

Results (approx. 2-3 pages)

Model Evaluation and Validation

Justification

Conclusion (approx. 1-2 pages)

Free-form Visualization

Reflection

Improvement

Work Cited:

1. Mean squared error *http://www.vernier.com/til/1014/*
2. Understanding PCA *http://setosa.io/ev/principal-component-analysis/*
3.