Predicting DJIA movements from the fluctuations of a subset of stocks?

Veronique Moore

CS229 Course Project Report - December 12, 2014

1. Introduction

It is the belief that the Dow Jones Industrial Average (DJIA) is the pulse of the U. S. Economy, that is: DJIA leads the movements of the stock market. True, or false? Regardless, the price behavior of DJIA is closely monitored daily by numerous investors around the world. So, what truly influences DJIA's movements up or down, one might ask?

By definition, invented in 1896 DJIA is a price-weighted average of 30 significant stocks (from 30 large publicly owned companies based in the U.S.) traded on the New York Stock Exchange and the Nasdaq. Consequently, DJIA is a strong authority in its own right and many factors contribute to its stability. The majority of the underlying factors contributing to DJIA movements have been commonly attributed to key economic indicators such as GDP, M2, CPI PPI, to name a few, as well as other factors including the Asian and European market performances.

This paper attempts to explain DJIA movements from a non-traditional approach, exploiting some techniques and algorithms of Machine Learning. We wish to answer the following two questions (later referred to as "our two questions of interest"): based on the fluctuations of a subset of possibly influential stock indices used as training data for a chosen model, can we accurately predict DJIA movements?

In addition, if we partition such selected subset of stock indices into 3 clusters, is one cluster more representative than the remaining data, and therefore a better predictor of DJIA movements?

2. Data Collection, Cleaning and Preprocessing

Source of Data: www.finance.yahoo.com

We screen stock indices, which had an average trading volume of above 7 Million between January 1, 2006 and December 31, 2009, covering 1007 trading days.

For each such index, and for ^DJI (the DJIA index), we extract 1007 daily historical prices. We then clean the data obtained, discarding all records that are found to be incomplete. We find that a total of 92 indices qualify to work with, serving as the 92 features of our model; we refer to them as the Full Set (see Appendix for the specific indices.)

We preprocess the cleaned data by converting, for each index, the difference between the current day and previous day prices into daily fluctuations of 0 for a fluctuation down (negative difference) or 1 for a fluctuation up (positive difference.)

3. Experimental design

We attempt to answer our two questions of interest in four scenarios: using the Full Set and using 3 subsets of the Full Set. With the K Means algorithm where K=3, we partition the Full Set into Cluster 1 with 21 features, Cluster 2 with 59 features and Cluster 3 with 12 features (see Appendix for the specific indices.)

In scenarios 1, 2, 3 and 4 we use features from the Full Set, Cluster 1, Cluster 2 and Cluster 3 respectively, as the input to our model.
For each scenario, out of the 1007 trading days, we reserve the first 807 days to train, and the remaining 200 days to test our Machine Learning model, making the ratio of training set to test set roughly 4:1.

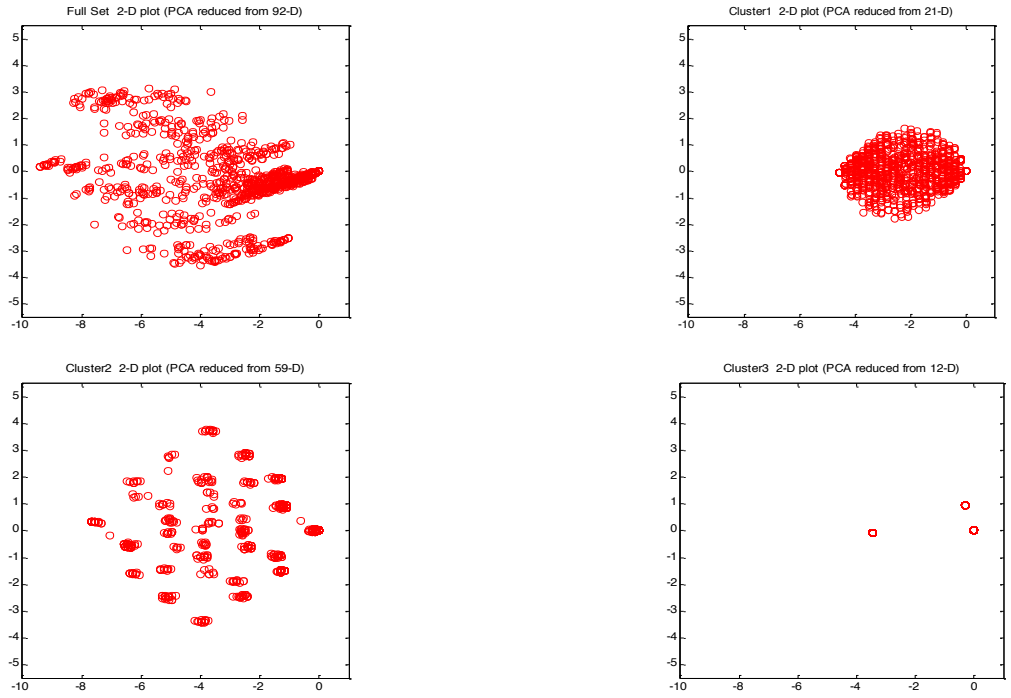
4. Models

Our problem is a classification one, where the daily prediction is either 1: ^DJI will move (fluctuate) up, or 0 : ^DJI will move (fluctuate) down. We consider five such models: the linear regression, the support vector machine (SVM), the support vector machine with RBF kernel, the naïve Bayes and the naïve Bayes with ksDensity kernel classification models.
Using these models, not only do we wish to answer our two questions of interest, but also we want to compare the models' performances on our data by way of two related measures: the accuracy and the F1 score (see Appendix for formulas.)

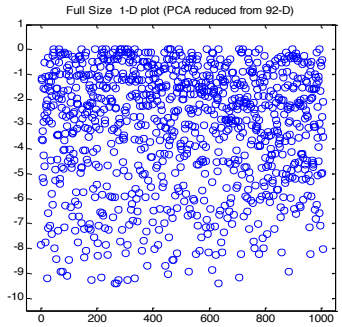
5. Dimensionality Reduction and Data Visualization

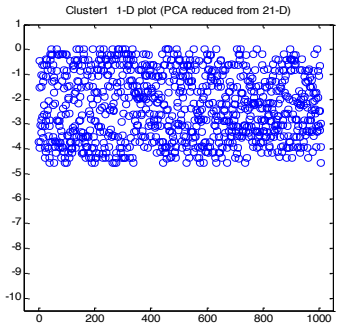
On each of the four sets of data, we perform the Principal Component Analysis algorithm to compress and reduce the dimension of the features from 92 to 2, and from 92 to 1. We are then able to visualize our data on a 2-D, and 1-D plots for the Full Set, Cluster 1, Cluster 2 and Cluster 3.

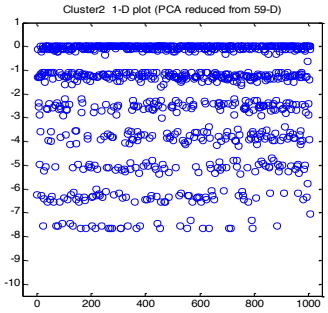
6. Cluster Results and Analysis

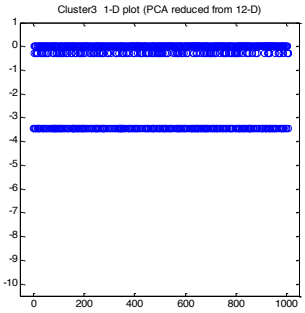


The 2-D plots reveal that there is a strong association between the first and second principal components of Cluster 1; the data is indeed very tight together compared to the results obtained on the Full Set, Cluster 2 and Cluster 3. We conclude that the data in Cluster 1 is fairly homogeneous.









The 1-D plots show no linearity in the first principal components, a result that is expected given the binary nature of the data. However data in Cluster 1 is confirmed to be more tight and less scattered than that of the Full Set, Cluster 2 and Cluster 3, making it more homogeneous as previously found.

7. Model Results and Analysis

Model		Linear Regression	SVM	SVM RBF	Naïve Bayes	Naïve Bayes ksDensity
Accuracy	**Full Set** (92 indices)	78.00%	78.50%	58.00%	57.00%	57.50%
	Cluster 1 (21 indices)	79.50%	78.50%	57.50%	57.00%	57.50%
	Cluster 2 (59 indices)	56.00%	53.50%	58.00%	57.00%	57.50%
	Cluster 3 (12 indices)	52.00%	52.00%	52.00%	42.50%	57.50%
F1 Score	**Full Set** (92 indices)	0.8070	0.8106	0.7273	0.7261	0.7302
	Cluster 1 (21 indices)	0.8285	0.8106	0.7213	0.7261	0.7302
	Cluster 2 (59 indices)	0.5600	0.4364	0.6250	0.7261	0.7302
	Cluster 3 (12 indices)	0.4947	0.4947	0.4947	NaN	0.7302

These results show that set wise, Cluster 1 and the Full Set tend to provide comparable, better and consistent results with the highest accuracy and highest F1 Score, across all five classification

models; actually, the linear regression model applied to Cluster 1 performs better than on the Full Set. In general, Cluster 2 only performs slightly better than Cluster 3, but can easily be removed entirely from our learning data, without negative impact to our prediction results.
Cluster 3 does the worst at predicting values, can be considered noise and be eliminated entirely from the entire set, in an additional step of cleaning our data. Indeed, Cluster 3 has the lowest overall accuracy of 42.50% and an F1 Score of NaN, using the Naïve Bayes model; the detailed results show that out of 200 test values, the number of true positive, false positive, false negative, and true negative predicted values is equal to 0, 0, 85 and 115, respectively.
Model wise, the results show that the linear regression model is our model of choice, followed by SVM with comparable results. The SVM RBF, Naïve Bayes and Naïve Bayes ksDensity have comparable poor results and should not be given any preference, without further tweaking.

8. Conclusion

In this paper we wanted to know if, using Machine learning techniques and algorithms, given the fluctuations of a subset of possibly influential stock indices as training data, we could accurately predict DJIA movements, and if a specific cluster of those indices was more relevant and a better predictor of those movements than the rest of the data.
From the linear regression, the support vector machine, the support vector machine with RBF kernel, the naïve Bayes, and the naïve Bayes with ksDensity kernel classification models used, our results indicate that we are indeed able to make the best predictions of DJIA price movements with 79.50% accuracy and an F1 Score of 0.8285, working with the linear regression model and using only Cluster 1. However, while certainly remaining relatively satisfactory, this prediction accuracy is not viewed by us as being very robust; in our opinion, it would be prudent to supplement our predictions in a real-time fashion with other, sometimes subjective, factors such as the current news events in and outside the U.S., or what is often referred to as the market belief or sentiment.

9. Future work

Additional analysis
Identify which indices best predicted the 2008 stock market crash, during which DJIA plummeted 3,600 points between September 19, 2008 and October 10, 2008.

Improving the current predictions
Further study of why cluster 3 did so poorly; perhaps come up with some additional "rectifying features" such as those that account for index volatility, use them on Cluster 1 to improve accuracy.

Construct a dedicated Kernel, which effectively captures the true geometric Brownian motion nature of the stock indices fluctuations in order to improve the current accuracy.

References

CS229 Course Notes
www.coursera.org/course/ml
http://openclassroom.stanford.edu/MainFolder/CoursePage.php?course=MachineLearning
www.finance.yahoo.com
www.wikipedia.org

www.investopedia.com

Appendix

1. Full Set indices: AA, AAL, AAPL, ABX, AIG, AMAT, ATVI, AVP, BAC, BBD, BBRY, BSX, BTU, C, CHK, CLF, CMCSA, CSCO, CSX, CX, DIA, DOW, EBAY, EEM, EFA, EMC, EWJ, EWT, EWZ, F, FCX, FOXA, FXI, GILD, GLW, HAL, HBAN, HK, HPQ, HST, INTC, ITUB, IWM, IYR, JBLU, JPM, KEY, KO, LUV, MDR, MDT, MGM, MRK, MS, MSFT, MU, NE, OIH, ONNN, ORCL, PBR, PBR/A, PFE, PG, QQQ, RFMD, RIG, SCHW, SDRL, SLB, SUNE, T, TLM, TLT, TSM, VWO, VZ, WFC, WFT, WMB, X, XLB, XLE, XLF, XLI, XLK, XLP, XLU, XLV, XOM, XRX, YHOO

2. Cluster 1 indices: AA, AAL, AAPL, ABX, AIG, AMAT, ATVI, AVP, BAC, BBD, BBRY, BTU, CHK, CX, DIA, DOW, EBAY, EEM, EFA, EMC, EWT

3. Cluster 2 indices: C, CLF, CMCSA, CSCO, EWZ, F, FCX, FOXA, FXI, GLW, HAL, HBAN, HK, HPQ, HST, ITUB, IWM, IYR, JBLU, JPM, KEY, LUV, MDR, MDT, MGM, MRK, MS, MU, NE, OIH, ONNN, ORCL, PBR/A, PFE, PG, QQQ, RFMD, RIG, SDRL, SLB, SUNE, T, TLM, TLT, VWO, VZ, WFC, WFT, WMB, X, XLE, XLF, XLI, XLK, XLP, XLU, XOM, XRX, YHOO

4. Cluster 3 indices: BSX, CSX, EWJ, GILD, INTC, KO, MSFT, PBR, SCHW, TSM, XLB, XLV

5. Model Accuracy and F1 Score formulas:

   Let tp, fp, fn and tn be the number of true positive, false positive, false negative and true negative model predicted values, then:

   by definition, $p = \text{precision} = tp / (tp + fp)$ and $r = \text{recall} = tp / (tp + fn)$

   $\text{Accuracy} = (tp + tn) / (tp + fp + fn + tn)$

   $F1\ \text{Score} = 2 * (p * r) / (p + r)$