Application of K-Means Clustering and Neural Network to Stock Return Prediction

Chong Sun

BAYLOR
UNIVERSITY

April 25, 2017

Outline:

Types of Data

- *Macroeconomic Factors* refer to overall conditions within a defined market that affect all properties within that market,
  - Inflation Rate, Interest Rate, Unemployment Rate, Political Crisis, Debt Crisis and Exchange Rate etc;
- *Fundamental Factors* include the firm-specific information that contributes to the economic well-being and the subsequent financial valuation of a company, security or currency,
  - Cash Flows, Profit Margins, Price-to-Earning Ratio, Price-to-Book Ratio and Size of the company etc;
- *Technical Factors* refer to stock market trading patterns and past price moving pattern,
  - Moving Average, MACD, and Momentum etc.

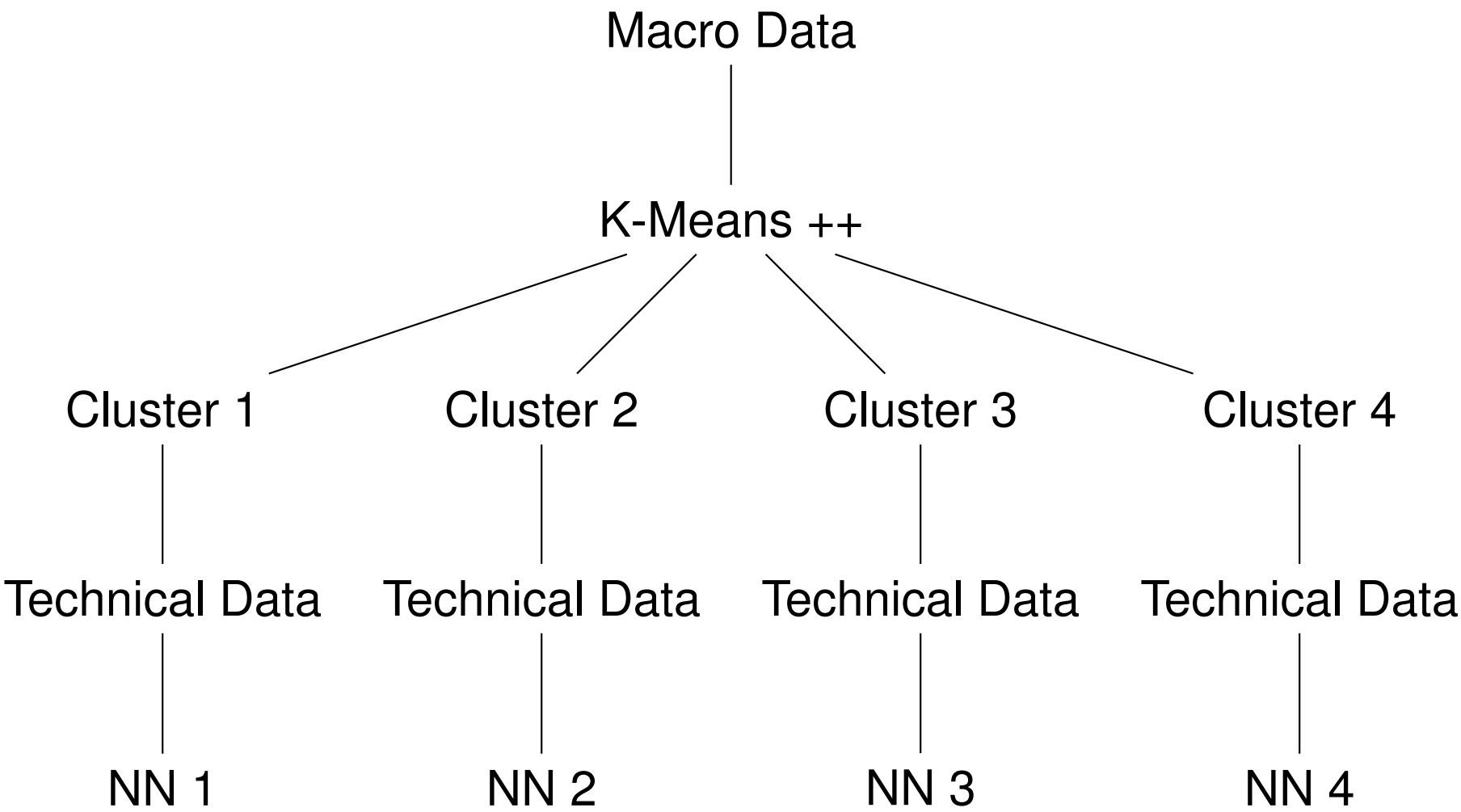
Challenges

- Too many types of data, high chance of overfitting;
- Small number of data of each type for different companies;
- Data snooping tendency;
- Stock market data tends to be highly noisy and nonlinear.

Traditional Approaches

- Hidden Markov model;
- Neural net work based on macroeconomic, fundamental and technical factors combined;
- Some researchers also employed clustering or other unsupervised learning methods based mainly on firm-specific fundamental factors.

My Approach


Macro Data
K-Means ++
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Technical Data
Technical Data
Technical Data
Technical Data
NN 1
NN 2
NN 3
NN 4

My Approach

- We employ 4-layer neural network with 4 nodes in the 1st hidden and 3 nodes in the 2nd hidden layer;
- The goal is to predict price moving direction of each stock in the next month.

Ideas behind the Approach

- Business cycle theory: there are four business cycles- Peak, Recession, Trough, Expansion, thus we have 4 clusters.

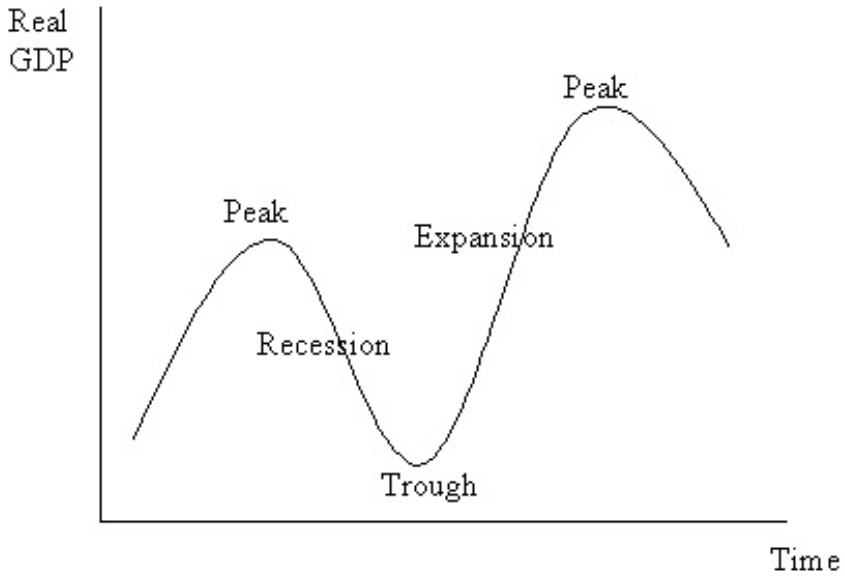


Figure 11.2

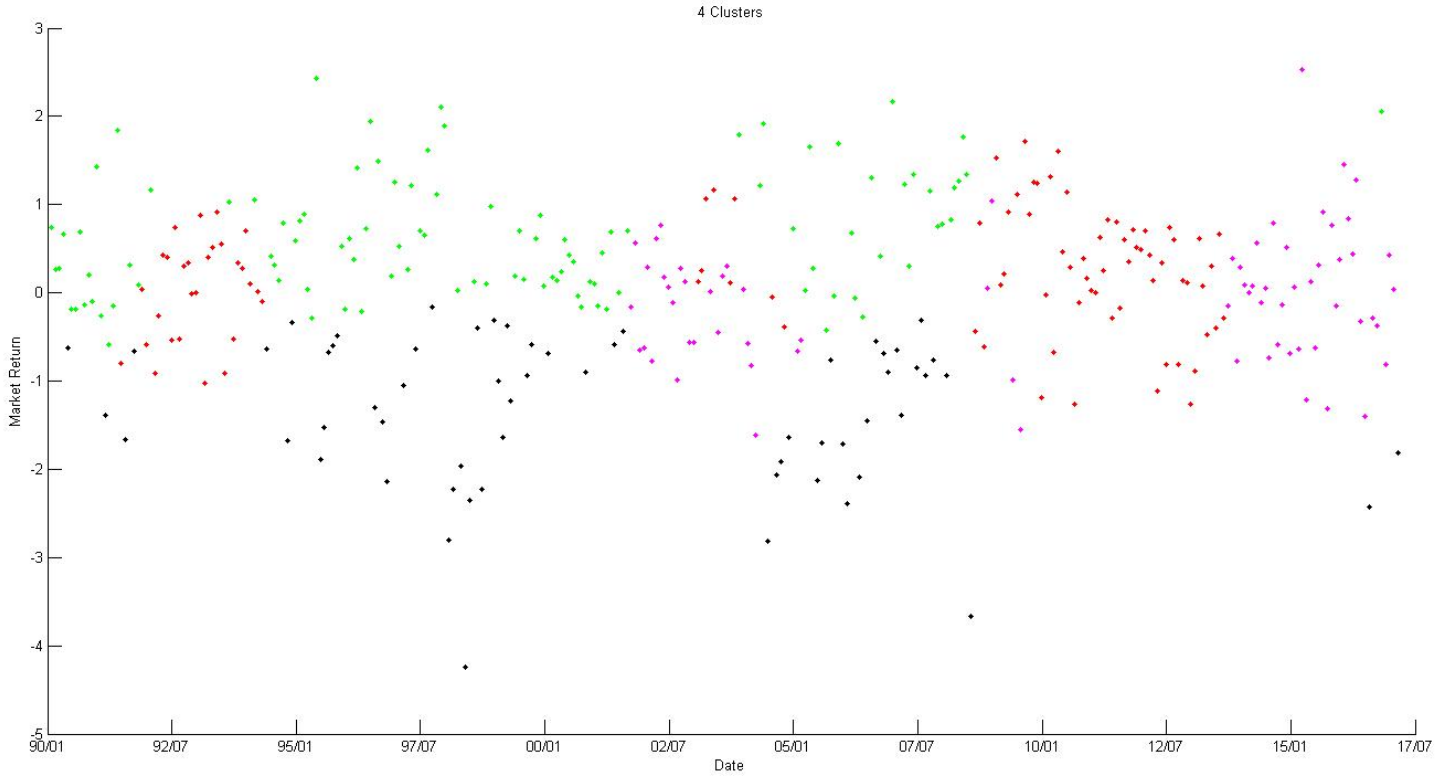
Ideas behind the Approach Cont'd

- Stock returns tend to show quite different patterns during different periods of time;
- Stock market data are highly nonlinear, thus neural network;
- Short term stock market data are quite noisy, thus we try to predict monthly return instead of daily.

Data Collection: Clustering

- Macroeconomic factors for clustering:
  - Inflation rate;
  - Fed effective interest rate;
  - Unemployment rate;
  - S & P 500 monthly return.
- Frequency: monthly data;
- Time period: Jan. 1990 - Feb. 2017.

Clustering



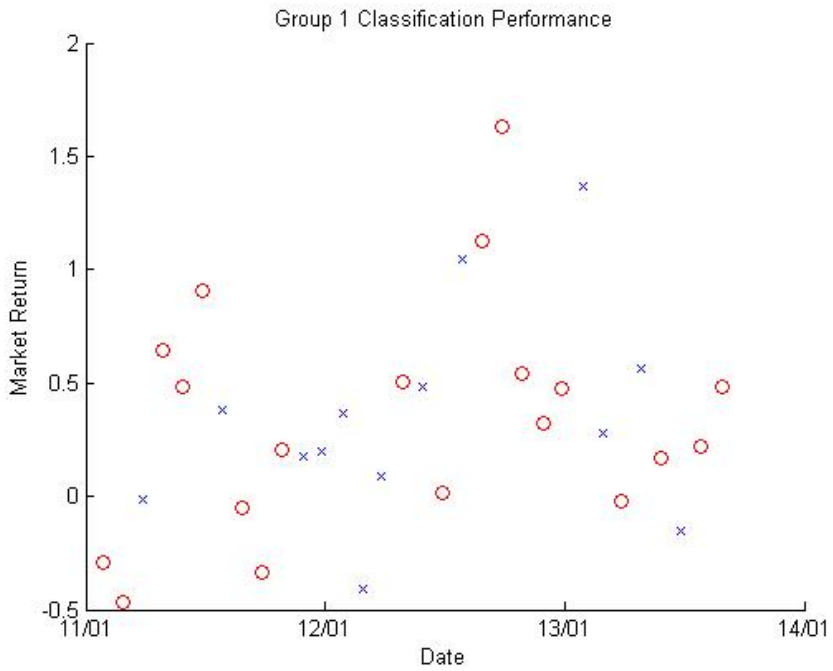
- Here we let y- axis denote market return, since it is most obvious indicator of macroeconomic conditions.

Data Collection: Neural Network

- Technical and fundamental factors for predicting:
  - return of current month;
  - 2 month moving average of stock price;
  - 3 month moving average of stock price;
  - size of the company;
- Frequency: monthly data;
- Company: Apple.

Results: Group 1

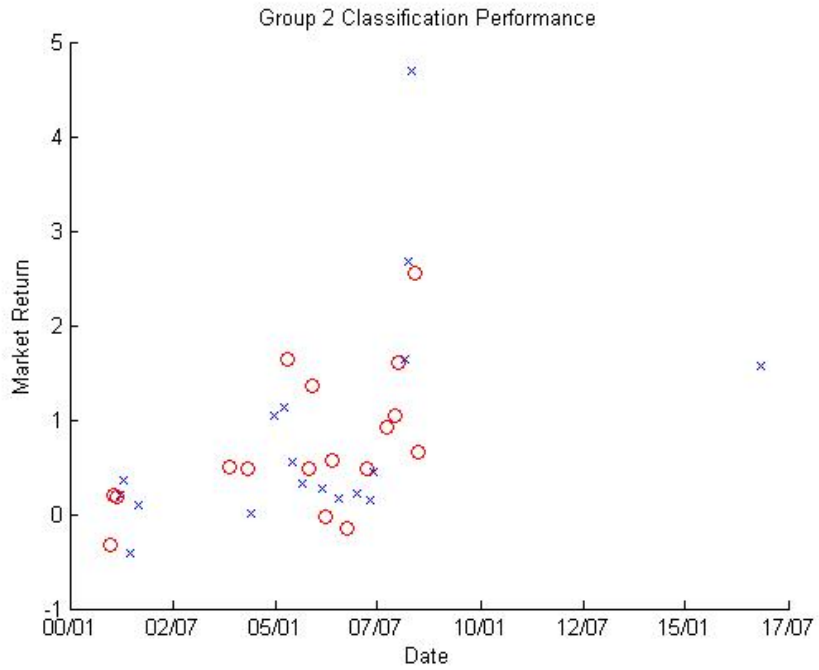
We compare the classification on the validation set of group 1.



Size: 93 points;
In sample Classification Error: 0;
Out sample Classification Error: 40%.

Results: Group 2

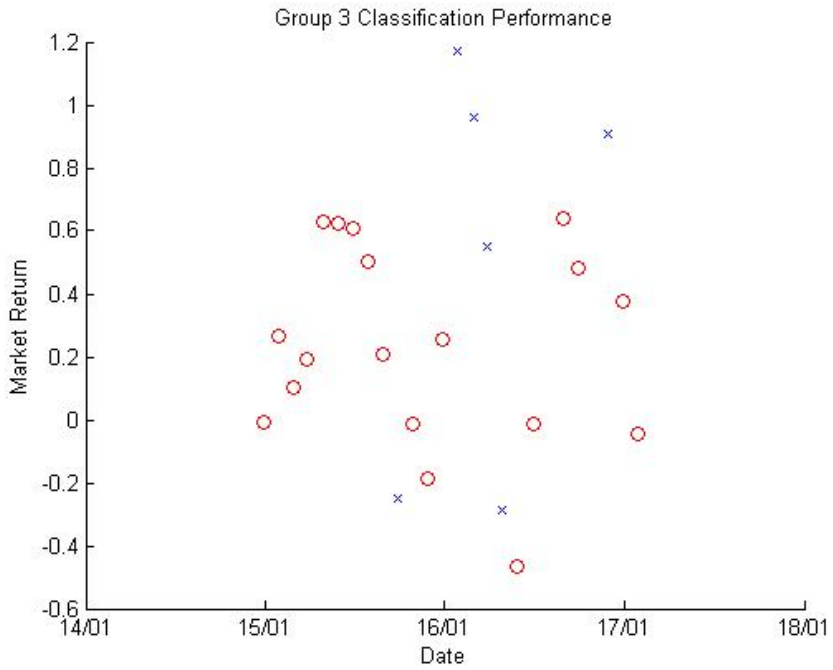
Classifications on the validation set of group 2.



Size: 104 points;
In sample Error: 1.43%;
Out sample Error: 51.43%.

Results: Group 3

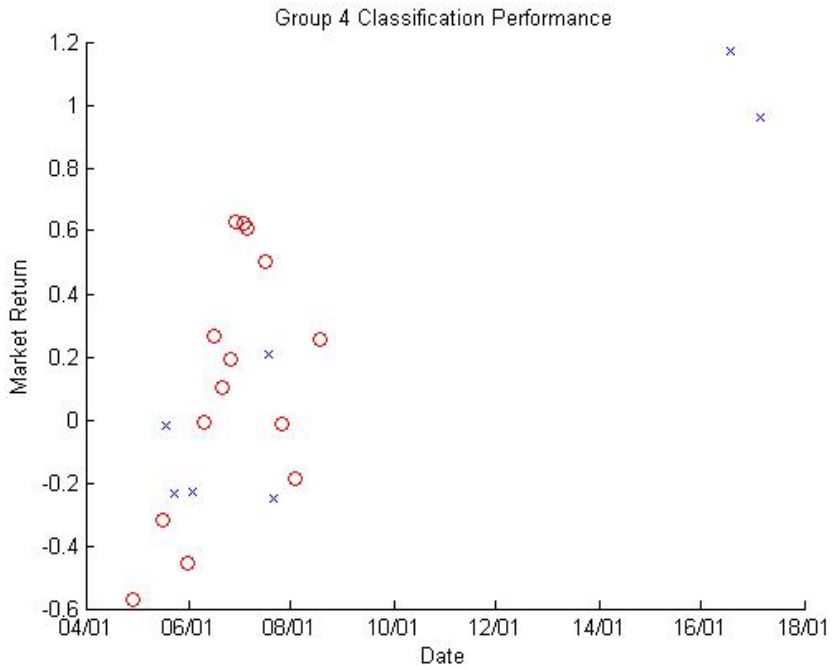
Classifications on the validation set of group 3.



Size: 69 points;
In sample Error: 2.17%;
Out sample Error: 25%.

Results: Group 4

Classifications on the validation set of group 4.



Size: 60 points;
In sample Error: 0;
Out sample Error: 33%.

Current Conditions

- Current macroeconomic factor
  - Unemployment rate: 4.5%;
  - Inflation rate: 2.4%;
  - Fed effective rate: 0.82%;
  - Market return: 0.48%.
- Apple data
  - Two-month price moving average: 95.09;
  - Three-month price moving average: 94.79;
  - Current return: -0.014%;
  - Volume: 25,273,400 dollars.

Prediction

- The current macroeconomic state belongs to group 1.

Prediction

- The current macroeconomic state belongs to group 1.
- Based on the neural network of group 1, Apple stock will generate positive return next month.

Question for You

Will You Bet 1 Million Bucks on Apple Now?!

Conclusion and Future Work

- Conclusion
  - Stock market return is highly noisy and hard to find a perfect model;
  - Even within same type of model, different initialization and choosing different factors will yield quite different results;
  - Better to put your money in index funds than in active funds.
- Future Work
  - Add more fundamental factors to neural network input set;
  - PCA and regularization to avoid overfitting problems;
  - Cross Validation to fight the problem of small number of data points;
  - Convolutional Neural Network.