Udacity Machine Learning Nanodegree
Capstone Project Proposal
Jibo Wen

Stock Segmentation

Domain Background

Stocks are typically categorized by sectors, such as energy, financials, technology, healthcare etc, or by their market capitalization, such as S&P 500, S&P MidCap 400 or S&P SmallCap 600. Different segmentation methods will present very different stock properties and behaviors, as from time to time, different sectors have different growth rate (or decline rate) and volatility. The purpose of this project is to apply unsupervised learning to the stocks trading on NYSE and NASDAQ, and find an underlying clustering mechanisms based on their financial conditions and growth rate. Such clustering mechanisms can present the stock market to investors from a different perspective, in which stocks exhibit similar financial conditions in each cluster, therefore, are likely to show similar performance. This new clustering mechanisms will also be compared to an existing mechanism such as grouping by sectors to test which methods yield better results based on the homogeneity of the stocks in each group.

Problem Statement

The core of this project is to find more fundamental ways to group stocks, so that stocks in each group will have similar financial standings and behavior. In order to find a stock segmentation scheme, there are two problems need to be solve:

1. Identify relevant features to use for clustering
2. Find a clustering method based on the characteristics of the features

The first problem can be tackled with PCA analysis to reduce the dimensionality of the dataset, or finding underlying dimensions that better capture the variability of the stocks. The features that will be examined in this problem are financial statistics of each stocks including price ratios and growth rate, details of which will be provided in the Dataset and Input section.

The second problem can be solved with unsupervised clustering algorithms such as K-Means or Gaussian Mixture Model. Such model can be further utilized to build a stock recommendation system by providing a list of stocks similar to a given stock. However, a stock recommendation system will not be the focus of this project and its performance will not be tested.

Datasets and Inputs

The dataset for this project is scraped from Yahoo Finance by using a python library called yahoo-finance 1.4.0. First, a list of NYSE and NASDAQ symbols can be downloaded from NASDAQ website, and then be used for extracting their financial statistics (or features) including[1]:

1. Price Earning Ratio (PE)

[1] Details regarding how the dataset is obtained can be found in Stock_Data_Retrieval.ipynb, and the cleaned up dataset that will be used for this project can be found in all_stock_data.csv

2. Price Sales (PS)
3. Price Book (PB)
4. Short Ratio (SR)
5. Divident Yield (DY)
6. Price Earning Growth Ratio (PEGR)
7. Percent Change from Year High (PCYH)
8. Percent Change from Year Low (PCYL)

Feature 1-6 will be based on the most recent quarter as of 06/02/17. Feature 7-8 will be based on the year to date data as of 06/02/17. Out of over 4000 stocks listed on NYSE and NASDAQ, only 2869 stocks whose financial data is complete are chosen for this study (More details in Stock_Data_Retrieval.ipynb notebook). Below is a sample dataset:

Symbol	DY	PB	PCYH	PCYL	PE	PEG	PS	SR	SECTOR
PIH	0.00	0.951	-0.171156	0.320283	33.1538	0.00	1.3706	0.52	Finance
FLWS	0.00	2.460	-0.105300	0.297700	27.3500	1.13	0.5700	6.33	Consumer Services
FCCY	1.16	1.300	-0.172700	0.474400	15.7000	0.86	3.1800	0.64	Finance
SRCE	1.63	1.760	-0.087200	0.522200	20.0600	1.86	4.7100	3.36	Finance
JOBS	0.00	3.410	-0.028300	0.555700	27.4800	7.42	7.5400	3.50	Technology

The reason these features are chosen is they are all ratios rather than absolute values, as absolute values may not represent a stock accurately. For example, a stock with a lower price is not necessarily cheaper than a stock with a higher price. The value of a stock is more accurately represented by their PE or PB ratios. A summary of each feature is provided below:

	DY	PB	PCYH	PCYL	PE	PEG	PS	SR
count	2869.000000	2869.000000	2869.000000	2869.000000	2869.000000	2869.000000	2869.000000	2869.000000
mean	1.923789	5.262124	-0.137261	1.036934	50.919360	4.647964	3.663936	5.036375
std	2.590263	32.518166	0.131984	33.707553	219.507561	140.381934	14.516095	5.342447
min	0.000000	0.070700	-0.942500	0.000000	0.075000	-4098.680000	0.010000	0.000000
25%	0.000000	1.360000	-0.188500	0.173100	15.500000	0.000000	1.010000	1.830000
50%	1.250000	2.130000	-0.103400	0.315600	21.670000	1.300000	2.250000	3.470000
75%	2.720000	3.760000	-0.041500	0.518400	34.030000	2.290000	4.192800	6.510000
max	35.740000	1191.530000	0.295000	1805.749800	6385.000000	4299.500000	529.420000	58.350000

Udacity Machine Learning Nanodegree
Capstone Project Proposal
Jibo Wen

Solution Statement

As several features might correlate with each other, PCA is necessary to compress the features so that they can be visualized and segmented at a lower dimension. Then a classification method such as K-Means or Gaussian Mixture will be applied to the dataset based on the distribution of the features (whether they are close to uniform or normal), and find each group's center points.

A potential solution to the problem is a classification scheme that divide stocks into groups. The quality of such classification scheme can be evaluated based on its silhouette score. Such score will be compared to that of the benchmark model to determine which classification scheme is superior. The solution should be reproducible and should always be the same.

Benchmark Model

The benchmark classification model is to categorize stocks by their corresponding sector labels such as energy, technology, etc. This is a basic categorization method used by most stock brokers (if not all) and data providers to study the stock market segment by segment. The benchmark model and the solution model can be compared by calculating their silhouette scores. The one with high score indicates a more cohesive grouping scheme.

Evaluation Metrics

The evaluation metrics is the silhouette score (as mentioned above), which is calculated using the mean intra-cluster distance and the mean nearest-cluster distance for each sample (sk-learn documentation). A higher score will indicate a better model.

Project Design

A theoretical workflow will comprise the following:

1. Obtain and clean up the dataset
2. Feature normalization
3. Study the correlation between each pair of features
4. Study the distribution of each features and remove potential outliers
5. Feature transformation (PCA) and examine the transformed features (how much variation they explain).
6. Visualized the transformed dataset
7. Create clusters by using K-Means or Gaussian Mixture, determine the number of clusters best suitable for the dataset by evaluating their silhouette score.
8. Recover the data by applying inverse transformation.
9. Compare the model clustering to the bench model clustering.

Future study of the clustering model can be building a recommendation system by finding the nearest stock in the same group.