Data driven approach for predicting stock index movement and prices using S&P 500 Data (Category: Finance Commerce)

Saksham Gakhar (06190199), sakshamg@stanford.edu
Lei Fang (06037833), lfang2@stanford.edu
Joseph Ballouz (06037006), jballouz@stanford.edu

September 27, 2020

Contents

1 Motivation

What problem are you tackling? Is this an application or a theoretical result?
This is an application project. We intend to devise algorithms that are based on years of training information that can be used to assist investors and financial analysts make predictions and gain insights into the behavior of stock prices of top 500 companies by predicting the direction of stock price index movement and future stock prices. We also intend to gain insight about the main clusters/classes in which companies may be categorized (from the 500 companies in our database) to see how coherently some perform than the others. This might be useful in understanding relations between these different players.

2 Method

What machine learning techniques are you planning to apply or improve upon?

2.1 Input features

We have data for each day for each of the 500 companies in the "S&P 500" index. Some of the **raw features** in the data are the stock closing price, the opening price, highest and lowest price that day and the volume of stocks. The **derived attributes** that we are planning to use include percent change in price, percent change in volume, percent return on next dividend, moving average and few more as we experiment with the data.
The daily data is available from 'S&P500' from February 2013 to February 2018. We intend to divide our data into training, validation and test data sets (illustrated in Fig.1) and temporally separate the validation and test sets so that the training data is immune from information from future. There are inherent assumptions that need to be addressed such as constancy of political conditions, general economic conditions and traders' expectations.

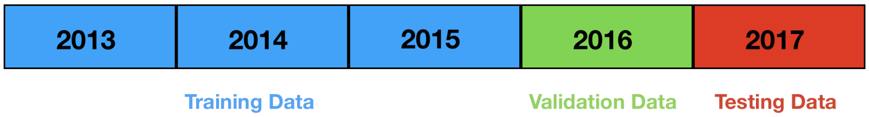

Figure 1: Data division illustration

2.2 Outputs

1. **For classification problem**, our label will symbolize the movement of the close price the day after: **0** if the price goes down (that means *sell*

indication for the user) and **1** if it goes up (*buy* indication for the user).

2. **For the regression problem**, the aim would be to output the stock price given a particular set of input features.
3. **For the clustering and unsupervised learning problem**, we aim to cluster the companies in a high dimensional feature space. We plan to implement types of clusters – **first**, clustering companies based on mean features (time averaged features); *second*, clustering based on short time averaged features. We do second kind of clustering over time. Then we should be able to see how the companies' relationship changes over time.

2.3 Techniques

1. Classification using SVM
2. Random Forest
3. Neural Networks
4. Clustering using k-means algorithm
5. Unsupervised learning for EM (Expectation-Maximization) algorithm

3 Intended experiments

What experiments are you planning to run? How do you plan to evaluate your machine learning algorithm?

3.1 Experiments

We intend to work on the classification problems using 'quick-&-dirty' implementations of logistic regression first (both with and without regularization) to understand the bias and variance in data so that we can appropriately select Kernel functions for SVM. We also intend to do forward feature search so that from the set of available features, we can select the few that are the most relevant for the problem.
For unsupervised learning, multiple experiments will be run based on the desired outcomes. As of now, our understanding of unsupervised learning implementation is limited. So we will describe these experiments in more detail in our future submissions.

3.2 Evaluation of the algorithm (performance metrics)

1. **For classification problem**: F1 score (for binary classification), Area under the Receiver Operating Characteristic (ROC)
2. **For regression problem**: Accuracy (w.r.t. the original test data)

3. **For clustering and unsupervised learning problem**: Methods of evaluation will include (1) ‘intra-cluster distance’ (a measure of how close each member of the cluster is to every other member); (2) the‘inter-cluster’ distance (how close each cluster of companies is to other clusters)