Machine Learning for Predicting The Dow Jones Industrial Average

Ryan Madden
Northwestern University
EECS 349: Machine Learning
RyanMadden2017@u.northwestern.edu

Abstract

The stocks comprising the Dow Jones Industrial Average (DJIA) are some of the largest and most successful publicly traded stocks in the United States (e.g. Microsoft, Apple, Exxon Mobil). Being able to predict which of these stocks will be successful based on a previous week's data, even at margins slightly above 50%, could prove to be incredibly lucrative. I attempt to predict which stocks in the Dow Jones Industrial Average will have a net value gain over the following week given data for the current week.

1. Introduction

The Dow Jones Industrial Average is a collection of 30 publicly traded stocks. The value of the DJIA is a compilation of the value of each of these stocks. This value is used as a metric to measure the strength of the U.S. economy, among other things. The closing value of the DJIA is reported each day on major news networks across the U.S.

2. Project Goal

My goal is to predict which stocks in the DJIA will have a net value gain over the following week given data for the current week. Investing money in the stock market is a gamble because it is difficult to know whether a stock will rise or fall in value. By achieving my goal I can help investors make confident and safe investments.

3. Data

My dataset includes 750 instances of Dow Jones stock weekly metrics including sixteen features. The data represents the performance of every DJIA stock in the first and second financial quarters of 2011. The data was retrieved from the UCI Machine Learning Repository [1]. The features in the dataset include the financial quarter, the stock symbol, the date of the last business day of that week, the weekly opening, high, low, and close prices for the stock, the volume of the stock traded over the week, the percent change in price of the stock, the percent change in volume over the last week, the previous weeks volume, the next week's open and closing prices, the percent change in the next week's price, the days until the next dividend for the stock, and the percent return on the next dividend. I partitioned the data into two sets based on financial quarter. There are 360 examples in the first quarter and 390 in the second quarter.

4. Approach

I initially partitioned the dataset by quarter with the intention to use the first quarter's data as both a

development and training set, while the second quarter would serve as a test set. My goal was to evaluate the effectiveness of different classifiers using 10-fold cross-validation on the first quarter's data and then test the best classifiers on the second quarter's data, using the first quarter as a training set.

I converted the data from UCI into .arff format and used Weka to test different classifiers.

4.1 Features

I first converted the final feature, percent change in next weeks price, into a binary classifier. The feature would be 1 if a stock went up in the next week and 0 if the stock went down. I then removed the superfluous features quarter, stock symbol, and date. Finally, I removed the features that reported the next week's opening and closing price, as these would not be known in real life when attempting to predict stock behavior. After removing these five features there remained 11 final features:

- Opening price
- Closing price
- Weekly high price
- Weekly low price
- Volume
- Percent change in price
- Percent change in volume
- Previous week's volume
- Days to next dividend
- Percent return on next dividend
- Result (price goes up or down)

4.2 Classifiers

I tested several classifiers:

- ZeroR: ZeroR is the most naïve of the machine learning algorithms. It gives an initial result that can be used to compare the effectiveness of other classifiers.
- NNge: An algorithm similar to nearest neighbor. I chose this algorithm because I expected stocks with similar attributes to behave similarly.
- FT: Functional trees are similar to decision trees, except that each node contains a function calculated from many different attributes that determines how to classify an example.
- Logistic regression: This algorithm classifies examples with a multinomial logistic regression model

5. Results

I tested each of the above classifiers on the first quarter's data using 10-fold cross-validation as well as testing on the second quarter's data using the first quarter as a training set (Figure 1). Although I tested many other classifiers, few performed better than 52% on either the cross-validation or the testing on the second quarter's data.

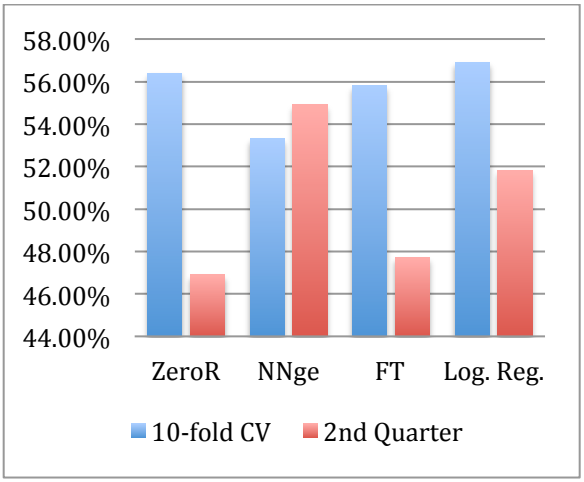

Figure 1. Testing Accuracy

6. Discussion

Very few classifiers provided any sort of reliable results. As demonstrated in Figure 1, nearest neighbor was the only algorithm that performed noticeably above 50% for both the cross-validation on the first quarter's data (53.3%) and the testing on the second quarter's data (54.9%). This makes some intuitive sense, since stocks with similar characteristics in a given week could be expected to behave similarly. Although ~55% correct classification on the test set may not seem like a success, it is important to remember the context of the stock market. Due to the general uncertainty that comes with stock trading, an algorithm with 55% accuracy could be used for great monetary gain.

7. Future Work

I believe further steps can be taken to improve the accuracy of the nearest neighbor algorithm. One such step would be to create several new features using a combination of the opening, closing, weekly high, and weekly low prices in order to assist the algorithm in determining which stocks are similar. Although two stocks may be trading at different values, the delta of their opening and closing or high and low prices may be similar. This indicates that the stocks are behaving similarly and should receive the same classification.

In addition, this classifier needs to be validated on further data. Evaluating the classifier on stock data from other years and other quarters would indicate whether this classifier truly applies well to this context.

8. References

[1] Brown, M. S., Pelosi, M. & Dirska, H. (2013). Dynamic-radius Species-conserving Genetic Algorithm for the Financial Forecasting of Dow Jones Index Stocks. Machine Learning and Data Mining in Pattern Recognition, 7988, 27-41. https://archive.ics.uci.edu/ml/datasets/Dow+Jones+Index34