A_Abstract

Nicholas Mitchell

March 31, 2016

Contents

1 Abstract

The emergence and ensuing explosion in popularity of social media platforms since the beginning of the twenty-first century, facilitated and catalysed by technological advancements in global connectivity, has created an abundance of data in completely new dimensions to traditionally available data. This has lead to rise in popularity of terms such as *big data*, *sentiment analysis* and *machine learning*, to name but just a few. The task of the data scientist is to extract information from masses of information in a way that allows new insights to be made. This study utilises a combination of these developments with the aim of enhancing data sets commonly used to analyse and predict the movements of stock markets; with particularly interest given to the Dow Jones Industrial Average (DJIA). Social media data from the Twitter platform has been used to add greater breadth to our data set; the results of multi-model sentiment analysis performed on individual tweets being incorporated as additional covariates into a forecasting model.

Component-wise gradient boosting was selected as the methodology for its inherent features, such as variable selection and scalability to high-dimensional data sets. These are major advantages, considering both the large number of covariates produced by sentiment analysis and the scarcity of prior knowledge regarding their individual influence levels.

The ultimate aim of was to determine the extent to which the addition of social media data, in the form of sentiment analysis, to traditional financial market data enhances the predictive accuracy of a set of forecasting models. Over numerous modelling conditions, we were able to show that the predictive accuracy was indeed increased through the inclusion of social media data. Gaussian models were shown to cope well with the inclusion of social media data and so wide data sets ($p \gg n$), returning strong results with errors equal or smaller to those measured in traditional data sets. Binomial modelling produced higher predictive accuracies still, however tended to be less robust to larger data sets, with performance being diminished in tests utilising increased-size data sets. Stochastic gradient boosting was shown not to perform as well as component-wise gradient boosting, returning noisier results - some models failing to outperform a naive model.