Sample UCL-styled A0 scientific poster LaTeX

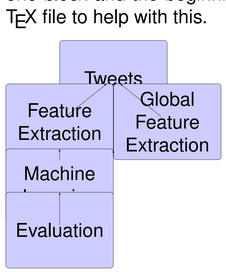
Shaun Dowling, Alessandro Ialongo, Andrey Levushkin, Matthieu Louis
University College London

Introductory segment

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Vestibulum justo. Praesent leo. Sed consectetuer. Aenean pretium, diam quis mattis porttitor, elit velit scelerisque sapien, sed convallis ipsum lectus non neque. Morbi in mi eu neque luctus scelerisque. Curabitur odio. Mauris a mi. Aenean iaculis erat vel sapien. Curabitur nulla velit, feugiat quis, imperdiet sit amet, vestibulum ac, diam. Suspendisse a metus. Pellentesque vulputate venenatis eros. In auctor, eros nec sodales faucibus, erat nisl facilisis nisl, ut rutrum nisl nunc eu est. Quisque eget ante at nunc varius ultrices.

First Piece of Content

It is worth fiddling around a bit with positioning of the text blocks to get the spacing even. I like a vertical gap of about 0.4 “block units” between the horizontal bar at the end of one block and the beginning of the next. There are contruction lines at the end of the TEX file to help with this.



Figures (and labels using the LaTeX picture environment) work as you would expect.

Feature Extraction

The key objective of this project is to use information contained within Twitter posts to predict the mood of the markets towards certain stocks. Tweets contain a great deal of information, including the text itself, linking between users, linking to entities (either explicitly view a tag or plain text) and network effect via re-tweets.

In their raw form, Tweets are not convenient for inference. As such we will focus a great deal of our time distilling the information contained within the Tweets into a form that we can use easily, ensuring all important aspects of the Tweets are preserved. This approach gives a nice layer of abstraction between the Information Retrieval and the Machine Learning steps in the pipeline. We will be gathering these features using two independent MapReduce jobs.

Global Statistics

Firstly, we will run a global job to gather statistics that depend on the entire dataset (for example tf/idf). The statistics available and the mannerr in which they are gathered will be a significant area of experimentation.

Some statistics that we plan to start with are:

- user average sentiment - to help us determine if someone’s good opinion of an entity is only because they are generally positive. This could be global or regarding a specific company.
- inverse document frequency - to be used with term frequency within tweets to potentially highlight particularly significant phrases

Machine Learning

The feature extraction process provides the raw material on which to construct models to explain the data and to formulate predictions about stock prices given new Tweets. In our project we will consider three main statistical models to uncover the patterns in the data:

- Linear Regression
- Support Vector Machine (SVM)
- Gaussian Process

Each of these models will make use of a portion of the available Twitter data (between 70% and 90% of the data, ordered chronologically) in combination with stock price data to extract the optimal parameters (according to a loss function). The parameterised models will then be used to predict the more recent performance of the relevant stocks given the remaining portion of the Twitter data. These predictions will then be evaluated against the real-world performance.

Linear Regression We will use regularised linear regression with the MSE (mean squared error) loss function (ridge regression) to give us a simple baseline on our predictive performance. With this simple model, we hope to find a linear relation (i.e. linear coefficients) between the features previously extracted and the stock prices. The (Tikhonov) regularisation will be parameterised by a lambda value which will determine the extent to which more complex (larger) coefficients will be penalised in our MSE loss function.

SVM When predicting simple increase or decreases in stock prices (see evaluation), the problem becomes a classification rather than a regression one. Thus we can use a support vector machine (experimenting with possible kernels and respective parameters) to fit a hyperplane in the feature space between the time splits in the Twitter data that corresponded to increases in the relevant stock prices from those that corresponded to decreases.

Gaussian Process For the regression task we will also attempt to describe the stock market as a Gaussian interaction of multiple samples defined by a suitable (kernel) covariance matrix. Defining this matrix will allows us to model periodicities, and the decay of correlation between adjacent samples. Out of the three, this model is the most elaborate as it allows for the greatest flexibility on the relation between the features and the stock prices. In fact the kernel covariance can encode very complex nonlinear relations between features and data-points. This requires also a higher degree of cross-validation to select the most effective kernel and its hyperparameters.

Performance Evaluation

ML algorithm performance

To demonstrate the effectiveness of our learning algorithms we proceed by splitting the data into testing and training sets. Than we predict the future stock price using the training set and compute the mean squared error of the prediction using the test set. An example of discrepancy between the predicted and actual price that we hope to minimise is shown below.

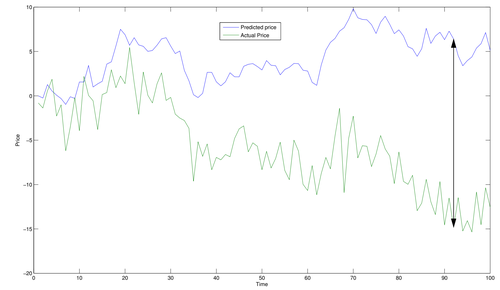


Figure 1: The arrow demonstrates the differences between the predicted price and the actual price in the test set that we hope to minimise.

By carrying out the above procedure with different sets of test and training data we will be able to establish the predictive power of our model as well as determine how far into the future we will be able to make effective predictions. It is expected that the performance of the prediction will deteriorate for predictions further into the future as shown in the figure above.

If the system will be used for training than the complex regression problem can be reduced to a simpler classification problem. Instead of predicting the exact price we will instead classify whether stock is going to go up or down after a specific amount of time. When evaluating this approach we intend to use the number of misclassification as a metric for determining the predictive power of our algorithm.

Back-testing

While mean square error in prediction is useful in evaluating algorithm effectiveness, low mean square error does not directly translate into trading performance.

The simple price model will need be further extended to incorporate stock liquidity and ensure that gains can be realised. Further optimisation is possible by incorporating trading fees as well as liquidity rebates to ensure that the system not only maximises prediction power but also profitability.