Application for new LRZ account

Nicholas Mitchell

March 8, 2016

Contents

1   User information

Name	Nicholas Mitchell
Phone	+49 15757 954 968
Email	nicholas.william.mitchell@student.uni-augsburg.de
Nationality	British

2   Project Description

Within the framework of his Master thesis the user performs statistical analysis of financial markets in combination with social media data (sentiment analysis) in an attempt to better explain uncertainty within the market and to ultimately increase predictive accuracy of current models, which do not include novel data such as sentiment analysis.

The sentiment analysis work has already been completed, meaning all the text-information has now been quantified, ready to combine with e.g. stock market returns. Several combinations of the data are selected to compare subsets containing differing information.

The work includes the usage of component-wise gradient boosting, allowing robust modelling of wide data sets ($p > n$) due to the inherent nature of variable selection that the method presents. Variants of gradient boosting are also implemented, including the usage of Bernoulli/Binomial distributions and stochastic gradient boosting. There are a relatively large number of hyper parameters that must be tuned, which means a large amount of computational time - cross-validation efforts especially contributing to the final run-time. Results are to be compared to standard modelling techniques such as AR(n) models.

3   Tool-set

Python was used for the collection of sentiment analysis. The algorithms that scored the text for its sentiment were written in a host of languages. From this point on, however, R will likely be the sole programming language. The algorithms that consume the most time include the gradient descent algorithm i.e. the minimisation of the relevant loss-function and then 25-fold bootstrap cross-validation of the results. Within R, the main packages used include:

- mboost
- caret
- gbm
- doParallel/doMC
- data.table