7_Conclusions

Nicholas Mitchell

March 31, 2016

Contents

1 Conclusions

1.1 Linear regression

This method was used to produce some results using a the simpler linear regression model, where no form of boosting is performed.

The same data set was used as with the comparison in Section `stoch-comparison`, namely with a frame-size of 60 days and correlation threshold, $\kappa = 80$ %.

No offsets were used for the coefficients, thereby not assuming any *a priori* knowlegdge of the model.

1. Autoregression (of matching order)
2. Fuzzy neural networks
3. Principal component analysis

1.2 Conclusions

The premise of this study was to include social media data as predictors - in the form of sentiment analysis performed on Twitter data - into traditional models, which would commonly make use only of financial market data, and investigate changes in the predictive capabilities of the resulting models. Comparisons were made between (1) traditional subsets, (2) subsets containing purely sentiment analysis data, and (3) subsets combining traditional market data variables with sentiment data variables. The use of five such subsets provided the primary source for comparison. The principal modelling tool has been component-wise functional gradient boosting, with the additional variation of stochastic gradient boosting being utilised on a small subset of the data for comparison.

It has been demonstrated that the inclusion of social media data in forecasting models does indeed enhance the performance of the forecasting models compared, in the vast majority cases. Results from the Gaussian family showed the subsets containing social media data outperforming the traditional subsets across the board, with comparable levels of mean-square error. In the paramter combinations of the correlation threshold $\kappa = 70$ % and 80 %, the errors produced by the subset containing both financial market data and social media data gave the highest predictive accuracy with the lowest errors. The binomial modelling produced overall better results to those from the Gaussian modelling, and was additionally able to amplify the performance advantage offered by the two larger data sets containing social media data in the paramter combinations using a larger frame-size. There were differences of up to 4 % in predictive accuracy, between traditional subsets and the combined subset, with improvements of at least 2 % in predictive accuracy in every individual value of lag used (with frame-size 0 60 days).

The utilisation of component-wise gradient boosting was instrumental in reaching the outcomes presented in Chapter `chapter-empirical-studies`. It provided a transparent method of handling large numbers of predictors efficiently by providing a statistically systematic method of variable selection while minimising the specified loss function. Stochastic gradient boosting was compared, introducing a factor of randomisation to the gradient descent in an order to reduce variance within the model. This, however, failed to outperform the component-wise boosting algorithm, with several models failing to produce predictive accuracies above 50 %.

1.3 Further comparisons

In order to explore how other methods compare to the component-wise gradient boosting, several different methods were used as benchmarks. In order to allow for more comparisons to be made, the number of subsets used was reduced[1]. For this section only two subsets were forward from Sections `results-gauss` and `results-bin` - namely the best and the worst performing subsets. (within a certain lag??) The comparison models chosen are the following:

1.4 Further work

Add this to the further work section at the end

A convenient side-effect of this methodology's variable selection ability is that it is no longer necessary to separate variable selection and model fitting, such as is often the case with wide data sets. An

[1] This is to save computational time as well as to make the comparison of results more straightforward.

interesting topic might be the effectiveness of feature selection, in comparison to techniques of *feature reduction* such as principal component analysis (PCA), where condensing many predictors into a handful that are frequently able to explain large amounts of the variance in a data set. The results from such models, however, are not easily interpreted directly, and are not necessarily able to be linked retrospectively to the input. This means it can be impossible to say how what impact each individual predictor had on the final model[2].

[2] An R package exists, named FactoMinoR [**?**], which allows some level of further analysis and interpretation of PCA results with respect to the input variables.