Forecasting Economic Growth Using Business News Data

Reinier Maat

August 30, 2016

Abstract

Business news sentiment can be conjectured to both reflect how the economy is doing, as well as influencing the spending of consumers and investors who read it. It is likely, therefore, there is predictive value in the sentiment that is contained in business news. In this project, we will collect business news over a certain time span, perform sentiment analysis on it, and forecast economic growth (GDP). These forecasts are relevant for e.g. investors making investment decisions, or for policy makers who wish to meet a certain economic policy objective.

Milestones

Problem Statement

Predict GDP (economic growth) from the sentiment that is contained in New York Times business news, perhaps supplemented with traditional economic indicators.

Literature Survey

The impact of news headlines has previously been studied in the context of exchange rates:

- Zhang, D., Simoff, S.J. and Debenham, J., 2005, December. Exchange rate modelling using news articles and economic data. In Australasian Joint Conference on Artificial Intelligence (pp. 467-476). Springer Berlin Heidelberg.

A comprehensive review of sentiment analysis can be found here:

- Pang, B. and Lee, L., 2008. Opinion mining and sentiment analysis. Foundations and trends in information retrieval, 2(1-2), pp.1-135.

Some theoretical background reading of how economic time series are predicted in econometrics is located here: (advanced, not required)

- Stock, J.H., 2001. Forecasting economic time series. A Companion to Theoretical Econometrics, Blackwell Publishers, pp.562-84.

Data collection and exploration

Students will collect data by requesting lead paragraphs from the New York Times Article Search API. It is important they only select business news. Once collected for a sizable time frame, they will need to preprocess the data to produce sentiment scores by month or week. These sentiment scores can be generated by applying scores from a pre-existing dictionary, like SentiWordNet, or they can choose to perform unsupervised learning to cluster of words in two sentiment groups. This data set is then combined with traditional economic indicators from sources elsewhere (e.g. the FED or the World Bank) to form a complete data set for GDP prediction. Data exploration can be done by visualizing all variables throughout time, and identifying correlations in the data.

Naive Solution

Seasonal AR model with only the leading indicators (simple linear regression with a 12 month lag).

Proposal

Challenges student will face in this project include:

- Collecting data from a public REST API
- Preprocessing data to form useful sentiment features
- Combining different data sources into one
- Evaluating different predictive methods, from linear regression to anything that sklearn offers the student to perform regression on the data (like support vector regression and the like).
- Cross-validating to choose the most suitable hyperparameters for regularization and so forth
- Report on the used methodology and make an out of sample prediction