Milestone 5: Proposal of Future Work

Kevin Shain, Mali Akmanalp, José Ramón Morales Arilla

11/28/16

Feature selection

Unigrams and Bigrams

In dealing with natural language, extracting the most meaningful features is essential for building a model. It is not clear how any given word relates to any other word. Naively, we could view all the words as independent features with the individual word count per document or term-frequency-inverse-document-frequency as the quantitative value of that feature for a given document. You could go beyond this and treat neighboring words as bigrams because the relative placement of words might have some information. A simple example is that the "strong" in "strong coffee" means something quite different than in "strong steel". There are many such words and phrases like that in English so including bigrams in our feature "vocabulary" makes sense.
We have experimented with creating a unigram and bigram vocabulary from the news data. The unigrams are chosen based on a reasonable maximum occurance frequency, minimum occurance frequency, and eliminating stopwords. The bigrams are a little bit more difficult to extract. The reason is that there are many more bigrams possible and their occurances are especially sparse. For this reason, we use the Dunning Log Likelihood Ratio to evaluate how meaningful the bigrams are. Essentially, the log likelihood ratio determines how likely the constituent words in the bigram are to occur together divided by their likelihood to occur separately. Then, we threshold the log likelihood ratios to keep only the bigrams with words that occur together quite frequently compared to occuring separately. As an example, the bigram with the greatest log likelihood ratio in our dataset is "New York". This makes a lot of sense as especially "York" almost never occurs without being preceeded by "New".
The current state of this feature selection algorithm is that none of the features are particularly correlated with the consumer confidence index (CCI). This is not entirely surprising as one would expect that each unigram or bigram's occurances have little effect on CCI. Some future work in this direction is to set better thresholds on the unigrams and bigrams to get just the most meaningful vocabulary and the least noise. Also, the model choice can do some "averaging" over the features that might reduce noise and enhance correlation with CCI so experimenting with high-dimensional models may be good to try.

Latent Dirichlet Allocation (LDA) Topic Modeling

A promising approach to dimensional reduction of our vocabulary is topic modeling. We can choose the number of features we want to end up with as the number of topics in our LDA model. This algorithm assigns words in the vocabulary to topics and the documents to topics in a way that most clearly separates the documents into the desired number of topics. In the end, we get a probability of each document coming from each of the topics. If we chose to have 10 topics, this compresses our high dimensional feature space into just 10 features. We can make sense of these topics by looking at the words assigned to them and see if they are related and what that relation is. Though one could imagine that the these topics are not correlated to CCI, they seem to represent a powerful dimensional reduction that is not too closely correlated to the average sentiment as calculated from a sentiment dictionary. For this reason, we might expect that average sentiment might be the most important feature, but the topics could convey different and somewhat important information.

More sophistication with sentiments

The SentiWordNet dictionary has a little more going on than just a positive or negative sentiment score for each word. In fact, the first difficulty is that the sentiment score depends on how the word is being used as there are often multiple definitions of a word and it can sometimes have different meanings depending on the part of speech. Figuring out the correct definition each time a word is used seems like it would be a very difficult problem, but part of speech tagging is something that has been done many times and could be transferable to our dataset. Also, the scores are not a single number. There is a positivity score, negativity score, and objectivity score. One can imagine that these scores are likely to be highly correlated, but there correlation is not defined in this dictionary. It will likely benefit our model if we can transform the three separate scores into one so that we don't have problems with multicollinearity.

Modeling

Since we are dealing with data points that are correlated in time, we have to think about how to incorporate that temporal information in the model. One thing to try might be a vector autoregression model. In that case, we say that any lag less than p determined by the AR(p) model that we choose has a coefficient for predicting the next value in the series. This allows us to use all of the features to combine and predict the next value of all of the predictors. This then allows us to predict all of our features one time step ahead. With that, we can fit a model without time dependence to determine how the future feature prediction determines the future CCI prediction.

A possibly more flexible approach is to just include certain lags or differences in time as new features. This is a potentially better because it takes the time information and formulates it such that we can treat all observations as independent and use the more complex models that we learned in class. Furthermore, it seems that the topic information is highly seasonal so incorporating lags of 12 months or differencing compared to 12 months prior could lead to a parsimonious model that incorporates seasonality. In other words, certain topics come up at certain times of year so the value of the topic is mostly a measure of time, but the difference compared to the same time last year could hold the most important information. The information about important lags and the amount of differencing needed can be extracted by fitting an ARIMA model using statsmodels. This gives a more systematic way of determining the important lags than including them all as features in a linear regression model.

Additionally, it is likely that we should choose a model that incorporates interactions with the average sentiment score. This is because the topics mainly tell whether the articles are talking about technology, banking, stocks, bonds, advertising, housing etc., but they don't say whether that topic occured in a positive of negative context. Therefore, the interaction of sentiment and topic could be very useful in the model.

Lastly, it is likely that including some more economic leading indicators may be important to incorporate into the model. These can naturally be included with lags alongside the CCI and news-based features to incorporate the temporal data.