Jie Yuan

From Data to Solutions, Week 4

Oded Netzer described applications of big data techniques in market research, specifically the use of network analysis to measure the online perception of many car brands, as well as the use of text analysis to predict the likelihood of loan repayments. The first part of his talk focused on how businesses are taking advantage of social media to better respond to consumers, including favoring those with larger social media presences. An interesting anecdote involved a reporter successfully turning on the air conditioning on a flight by tweeting at the airline. Overall however, I believe the talk was geared more towards an audience with little experience in big data, not a group of PhD students. Too much time was spent on general claims about the revolutionary nature of big data, and on a very high-level description of what text mining is. At least in the way the material was presented, I did not see anything particularly impressive or novel from a modeling standpoint. The entire discussion on loan repayment prediction involved using Naive Bayes classifiers on single words or n-grams to predict a binary variable, which is done regularly now even in business settings.

The main substance of Dr. Netzer's talk involved discovering co-occurrences of products in forum posts, specifically for car models. These are discovered by identifying the occurrence of terms referring to each car in a forum post, and identifying pairs of models which appear together with high frequency. This is done using a calculation called lift, which models the joint probability of the two car models appearing against the prior probabilities of the two appearing independently. The car pairs are then inserted into a graph with force-directed edges, so that the more connected cars appear closer together. In this plot, the nodes arrange so that clusters emerge of frequently-compared cars. When k-means clustering is applied, clusters appear roughly indicating the type or expensiveness of the car model. The goal of this application is to see which cars compete with one another, and to gauge public perception of a given car brand based on which cluster it falls into. A similar graph analysis was performed by identifying pairs where owners of one car switched to another model when buying a new car. The structure and location of the individual models matched with high accuracy the graph obtained through text mining, which was quite impressive.

One thing that was not described is how the class memberships indicated by the dotted lines in the graph images were determined. My suspicion is that these were drawn arbitrarily, since they are just a little bit too perfect – in the first "Perceptual Map of Brands," the dotted lines of two classes just barely encompass and separate the Cadillac and Lincoln, both outliers in their respective classes. If the intention was to show the k-means classes, it would have been more proper to simply color the points, without suggesting the existence of some distribution. Even if the dotted lines represent true deviations from a Gaussian distribution trained on the k-means clusters, this is not exactly proper either because the Gaussian distributions are applied post-hoc, and my understanding of k-means is that it is assumed during training that the clusters

are spherical – the covariances of the clusters are not taken into account. A mor rigorous solution could probably be obtained using a Gaussian mixture model to fit fully-described distributions to the resulting clusters – this is essentially the k-means approach but taking the class covariances into account.

One future direction I'd like to see this research go in is combining the co-occurrence networks with some of the models that have been developed for sentiment analysis. In the associated paper, the authors already have combined car and word association networks, but there is no directionality in the network - the representation only tells you which items are notable for which car, not whether each of these items is especially praiseworthy or deficient in the car. By obtaining labels of positive or negative attitude towards a particular component of the car or the car model itself, a more descriptive set of relationships can likely be obtained showing overall which car of a co-occurrence pair is preferred, and why.

I checked out the supplement to the paper to learn more about the text mining process involved in extracting car names and important terms. It appears to use a combination of conditional random fields and hand-crafted rules, as well as some manual corrections in a couple stages by both a car expert and users on Mechanical Turk. Involving humans in this analysis seems kind of like cheating, and making use of custom rules which don't appear to be listed clearly in the paper make the model much less useful for other applications and difficult to reproduce. Since the final task of identifying car co-occurrences is actually quite easy once the car occurrences themselves are extracted, I think it's worthwhile to really make the entire text mining process automated.

Additionally, I feel like going down the human labor route may not even be necessary – in their analysis of users who post a large volume of text, the authors performed their network analysis using only subsets of the data, and found the network to be reproducible even in extreme cases where a large fraction of the data was missing. If this was the case, then it would have been far easier to just skip over those problematic abbreviations entirely. Then the model would be fully automated and still present good results, which to me seems like a much more compelling case for more widespread adoption. If a large number of abbreviated car names are being lost, then additional tools can be used to perform more refined searches.

While looking into this problem I came across a Google autocomplete API which works pretty well in this regard. For example, if you wanted to search for "Honda Sivik," type the following into a terminal:

```
> curl 'http://suggestqueries.google.com/complete/search?client=firefox&q=Honda%20Sivik'

["Honda Sivik",["honda civic","honda civic 2016","honda civic si","honda civic 2015","honda civic type r","honda civic commercial song","honda civic for sale","honda civic lease","honda civic hatchback","honda civic coupe"]]
```