Modeling Social Events

Bryan Ball and Asher Krim
Machine Learning and Computational Statistics
Professor David Sontag, NYU, Spring 2014
Project Proposal

Introduction

In this project, we would like to make predictions by modeling a large corpus of human social events data known as GDELT. We will focus on modeling things such as the relationship between social events and stock index performance, as well as social unrest with unemployment rates. Additionally, we will attempt to predict the intensity of future events.

Getting to Know Our Data

The Global Database of Events, Language, and Tone (GDELT) is a massive database which attempts to map every event or conflict involving individual people, organizations, or countries into one cohesive context. The dataset details the event, who is involved, and how the rest of the world feels about it. It is updated daily, and the back-catalog reaches back to 1979. At the heart of GDELT is the simplification of events into basic attributes, including: The date and location, the two actors, the tone (emotional response to the event), and a category the event falls under (such as armed struggle or demonstration).

Accurately predicting the behaviour of the financial markets is probably impossible. Just the same, we are interested in mapping the GDELT data to stock indexes and unemployment rates. Following the recent events involving Russia and Ukraine the Russian stock market plummeted, suggesting a link between financial markets and negative social events. The data that we will use is pulled from Yahoo and includes starting and closing prices, as well as daily high and low values. Our unemployment data is pulled from EuroStat.

Ingesting and Pre-Processing

Because our main data source is so well maintained, relatively little work will need to be done for cleaning the data. The GDELT website includes many documents which clarify the data attributes and supply the possible values that they can take. We will follow the usual procedures for transforming categorical values into feature vectors.

For those questions which we intend to ask about individual countries, we will include all rows of data in which that country is either Actor1 or Actor2 and prune the rest. For unemployment prediction we will focus on events internal to the country.

Questions We Will Try To Answer

Question 1: How much cooperation or conflict will there be in a particular country tomorrow and what are its causes?
Question 2: How much do different stock indexes change based on the world events of that day?
Question 3: What is the connection between the unemployment rate in different European countries and the events in that country this month?

Project Plan

The Goldstein scale measures world events on a continuous cooperation-conflict scale. Events of extreme conflict (e.g. war) will have extreme negative values while cooperative events (e.g. giving economic aid) will have positive values. For each country, we will aggregate these values over the course of a day to convey a measure of the total conflict or cooperation. This creates a univariate time series model where each day is an observation. Rather than a simple univariate ARIMA model, it would be better to create a number of different time series variables based on the countries involved, event code (code specifying the nature of the event), tone (the nature of people's feeling about the event), and quad category (separates events into material conflict, material cooperation, verbal conflict, and verbal cooperation). First, we will create a VARMA model. This predicts the next value in a time series vector (where each entry would correspond to the total of Goldstein scale values for events in a particular category and the length of the vector would be the number of categories). This is like an ARMA model, but the individual entries are vectors instead of scalars and the coefficients are matrices instead of scalars. The prediction for total Goldstein scale values on a particular day would be the sum of the entries of the vector. The parameters to be tuned for the number of previous days and shock to include in the model. An additional model we could use to predict would be ridge or lasso regression. The difficulty is that these models are not built for time series. There are two ways to approach this. One is to have a different create a different variable for each amount back in time. This would multiply the number of variables by the number of lags we wish to include. Another approach would be to reduce the weight of events on previous days the farther back they occur. For example, the sum of the Goldstein scale for a particular day one day ago would be multiplied by exp(-a*1), two days ago would be multiplied by exp(-a*2), and so on. This "a" parameter could be tuned in cross validation. For both models, to cross validate, we would take sliding windows, say a year, train a model, and predict the next day. Then slide the window forward and repeat. Then we would use a final period of time for testing.

The second problem we will investigate is predicting the performance of stock indexes based on the occurrence of events. More specifically, we will use penalized regression models such as ridge or lasso where the target value is the daily percent change in the stock index. We would repeat this problem on a number of different indexes. Much like the previous problem, each feature will be an aggregation of daily events of a particular type (e.g. one feature could be Nigerian material conflicts with other African nations). For this problem, we could experiment with a number of aggregation methods, the simplest being the number of events (e.g. if there were two Nigerian material conflicts with other African nations on a particular day, that feature would have a value of 2). Since markets react to news almost instantaneously, there would no point in including data from previous days. However, markets do include assumptions about the future in their current price. So for a more complex model, our features would need to be how many more events occurred than were expected to. In this case, we would need to build a time series model like the one described in the first problem that predicted the number of events in each category. The features in this model for predicting the change in market price would be the residuals in each category. We would try both models - the simpler model which uses actual event count and the more complex model that uses event count residuals. Our evaluation metric would be square error. Cross validation would involve sliding windows like were explained in the first problem. It would be dishonest for the time series model used in the more complex model to be trained on the full data set. This means that we would have to build a new time series model for each new window.

The modelling of the third problem will be very similar to that of the second with a few major differences. First, the events will be aggregated over the course of a month rather than a day to form the features. Second, the target variable will be the change in the following month's unemployment rather than the same day's index price change. Market prices within seconds or minutes while change in

unemployment can be a much slower process. Third, we will only look at events within the country we're investigating.

Project Timeline

4/1: Decide which of the above problems we will try to solve first
4/4: Decide on programming language/ framework, get all data, and load it in
4/11: Build Features
4/20: Build models, identify model inadequacies
4/27: Choose final form of models, cross-validate them, and test them
5/08: Write up paper
5/11: Create presentation and poster

Expectations From The Project

We realize that predicting the future is an uphill battle. In all likelihood, if we achieve accurate results it will mean that we have made a major mistake, or are predicting something trivial. Even producing poor results could be useful in showing that we were unable to find correlations in the data. We would expect to uncover interesting relationships which might shed light on the reasons things happen in the world.

The first problem would predict a daily country specific metric of cooperation and conflict. The one day ahead predictions would be telling of the levels of conflict to be expected the following day. The coefficients from the model would be more interesting. They would indicate which event types lead to prolonged cooperation or conflict and which event types are isolated. Additionally, they would show how the events in one country lead to events in another country. This would show how different countries are interconnected.

The actual values of the second problem's market predictions won't show anything interesting. They can't be used for trading strategies since we are using them to predict the return on the same day they occur rather than the next day. Again, the most interesting part of this model would be the coefficients it finds. They would show which event types most affect particular economies and show how strongly events in some countries affect markets in others.

The third problem's model will be interesting in two ways. First, we can predict unemployment. Second, the model's coefficients will tell us what types of events lead to changes in unemployment.

Links

GDELT - http://gdeltproject.org/
EU Public Opinion - http://ec.europa.eu/public_opinion/cf/index_en.cfm
EuroStat (Unemployment rates) -
http://epp.eurostat.ec.europa.eu/portal/page/portal/employment_unemployment_lfs/introduction