Milestone #5 TF: Zeerak Ahmed November 28, 2016

Project:

Martingale Based Anomaly Detection

Group Members:

Melissa Curran, Alexander Dubitskiy, Deepika Sinha, Liangliang Zhang

Proposal of Future Work

In our project, we are developing a method of detecting anomalies in non-labeled time series data with adjustable confidence and sensitivity. We will be applying this to the adjusted closing price of Microsoft stock.

What We are Trying to Detect:

1. Unexpected short term (trading day) changes in the stock price. It is not necessarily large drops or gains. It could be the absence of those in a context of some other events (the company quarterly earnings released) that usually trigger a price change.
2. Unexpected long- or medium-term (weeks or months) high or low stock price contrary to evidence (company performance, broker expectation).

What We’ve Learnt So Far:

1. The stock price and also a few derivatives we analyzed (first and second differences, difference between stock price and 7, 30 and 120 days moving averages) are not normally distributed. We cannot use Z-score to detect outliers with defined sensitivity and confidence level.
2. We tried to use empirical outlier detection methods (the Turkey rule) on the above data and found it difficult to interpret and justify the results.
3. Based on the above we’ve decided the best way forward would be to use a multidimensional dataset that introduces some context (overall market and economy performance, the company performance, significant events like investments and acquisitions).

Challenges:

1. It looks like using martingales for testing exchangeability might be a good method of detecting medium- or long-term shifts but might not be appropriate for the short term (local) anomalies detection. The martingales introduce a delay in the outlier detection (the price for building up the confidence level) so they might not be able to pick a sharp but short lived price changes (especially swings).
2. We think the key part of the project is to select a good strangeness function. We believe we might need to use different strangeness functions to detect short- and long-term changes.
3. We find it challenging to obtain the historical consensus and estimates, but we think it is essential for us to have it.

Performance Metrics to Evaluate the Model: Measuring the performance of an unsupervised learning algorithm is very challenging. In an exemplar run of our program, we will use the news to validate detected anomalies. For example, if our program detects an anomaly on a particular date, we will check the news for that date and the previous few dates to determine if a particular event triggered the anomalous change in stock price (e.g., quarterly earnings report, merger, acquisition, or negative news report). This information from the news will help us determine whether the anomalies we have detected are normal responses to company activity or whether something else might be going on.

Future Work:

1. Get historical consensus and estimates for Microsoft stock and update our exiting dataset.
2. Some of our data sources are not automatically updated (actually most of them apart from the MSFT stock price from Yahoo Finance). We need to develop a method of keeping them up to date.
3. Evaluate One Class SVM as our strangeness function. The evaluation has to be done in the context of detecting both short- and long-term anomalies.
4. Evaluate Isolation Forest as our strangeness function. The evaluation has to be done in the context of detecting both short- and long-term anomalies.
5. Based on the evaluations, develop a strangeness function and plug it into our anomaly detection pipeline.
6. Develop the automated validation method. Given a past date, we want to get a probability of the significant company-related event happening.

Deliverables:

1. An up-to-date historical dataset about Microsoft stock performance and some context data, as described, in the form of CSV file.
2. A method of detecting anomalies in the above data in a form of python program.
3. A method of validation of the anomaly detection in a form of a python program.
4. A comprehensive investigation of the methods performance with different parameters in the form of an ipython notebook. This is essentially the model selection process.
5. A working prototype where a user can select confidence level and/or sensitivity and get a list of anomalies in through a command-line program.

Files Included in This Submission:

1. Proposal of Future Work (Proposal.docx)
2. iPython notebook containing our current code (cs109\_anomaly\_milestones\_4\_and\_5.ipynb)
3. PDF of the iPython notebook (cs1019\_anomaly\_milestones\_4\_and\_5.pdf)
4. Folder containing our datasets (datasets)
5. Folder containing helper code for our iPython notebook (cs109)
6. A review of existing anomaly detection methods for time series data (Introduction)