Reflection for Example Coursework, Week 0

Rachel R, Sept 2018

Our group of two were set the task of visualising "conn_sample.log" to produce a captioned figure.

We went about finding representations first starting with the notes (treated as unseen) and then finding examples on the internet. I worked primarily with R, and Peter P worked primarily with Python, working to combine our results into a single figure made jointly in R.

The project contains two clear challenges:

1. The dataset is very difficult to work with. It contains a large quantity of missing data, many "quasi-categorical" variables, as well as quantities with very strange distributions. As students starting in this field, we were not familiar with what the data "meant".
2. We were required to find interesting ways to visualise the data by doing a web search. Unfortunately the literature doesn't give clues for how to standardize the data in the first place.

We weren't sure exactly how to go about the literature search, since the space is so vast. We ended up looking for content dedicated to cyber security, and content for categorical variables. The cyber-security specific content is all python based, and the general data-science material didn't help us with the data processing. As a result, although we tried several standardisation techniques that were in the lectures, and some that we made up ourselves, we feel that better understanding of what the data meant in cyber security would have helped pre-process the data for visualisation. Visualisation might feel like a separate task to modelling, but it isn't. An example is the use of t-sne, which the literature is very consistently positive about and which is a massively cited paper[^1] (>5K citations). However, we were not able to make it work satisfactorily because we didn't understand how to scale our data for input.

More positively, we were able to make sense of the relationship between service and ports by both graph and other methods such as the balloonplot that were new to us. This produced a crisp and clean visualisation that made the associations very clear, so it was chosen to go in the Figure, along with the correlation matrix generated in Python. This made the time-correlations in when IP addresses were active very clear. Between these plots, it is possible to learn a lot about the data, even without knowing anything about cyber security. We capture 4 important dimensions: time, service, ports, and IP address. Since the time measurement encompasses data volume, the only main concept that we omitted was the packet size information. We did explore this but there wasn't enough signal to justify making it to the main plot.

R and Python are both well served in the visualisation tools available, and the

final results look professional. The main limitations for visualisation are the example code available on the internet. There is an almost infinite amount of general visualisation tools, so we focussed on those that looked most interesting for our categorical cyber data and were discussed in a data-science or machine learning context. Unfortunately, most cyber data does not seem to come with visualisation - it is mostly tables and summary statistics of the data. R in particular had few resources dedicated to cyber security.

Overall, I would like to return to this problem with a better understanding of cyber security data so that I knew how to perform the pre-processing.

[1]: Maaten L 2008, "Visualizing Data using t-SNE". Journal of Machine Learning Research.