Anomaly Detection

Patrick Rubin-Delanchy
patrick.rubin-delanchy@bristol.ac.uk

School of Mathematics, University of Bristol

November 27, 2018

Why anomaly detection

A lot of cyber-security problems can be phrased as anomaly detection. To give an example, the MS-ISAC guide to DDOS attacks (2017) suggests to:

- "Look for a large number of SYN packets, from multiple sources, over a short duration"
- "Look for a large number of inbound UDP packets over irregular network ports coming from a large number of source IP addresses"
- "look for a significant amount of inbound ICMP traffic from a large number of sources"
- "look for a large number of inbound traffic from a significant number of source IP addresses with a destination port of 80"

More generally: I think of anomaly detection as the science of finding interesting features in data.

Intended learning outcomes

ILO1 To recognise and apply a range of models for dynamic network data, and their estimation

ILO2 To understand core anomaly detection concepts and tools, including mastering theory and interpretation of hypothesis tests, controlling false positive rates and performing meta-analysis

ILO3 To apply these anomaly detection tools to analyse real large-scale data and report the results

Intended learning outcomes

ILO1 To recognise and apply a range of models for dynamic network data, and their estimation

- We'll focus mostly on tractable and scalable Bayesian techniques, very much inspired by the paper: Heard, Nicholas A., et al. "Bayesian anomaly detection methods for social networks." *The Annals of Applied Statistics 4.2* (2010): 645-662.

ILO2 To understand core anomaly detection concepts and tools, including mastering theory and interpretation of hypothesis tests, controlling false positive rates and performing meta-analysis

ILO3 To apply these anomaly detection tools to analyse real large-scale data and report the results

Intended learning outcomes

ILO1 To recognise and apply a range of models for dynamic network data, and their estimation

ILO2 To understand core anomaly detection concepts and tools, including mastering theory and interpretation of hypothesis tests, controlling false positive rates and performing meta-analysis

- We'll learn about the fundamentals of hypothesis testing and meta-analysis, but also practical issues (such as discreteness) and connections to machine-learning.

ILO3 To apply these anomaly detection tools to analyse real large-scale data and report the results

Intended learning outcomes

ILO1 To recognise and apply a range of models for dynamic network data, and their estimation

ILO2 To understand core anomaly detection concepts and tools, including mastering theory and interpretation of hypothesis tests, controlling false positive rates and performing meta-analysis

ILO3 To apply these anomaly detection tools to analyse real large-scale data and report the results

- In the assessed coursework and problem classes, we will use the techniques to (try to) find red-team activity in computer network data, specifically the Los Alamos National Laboratory authentication dataset.

Aims

1. To give you a firm mathematical footing in anomaly detection, in a world where everyone is making simple and easily avoided mistakes
2. To make you fully appreciate the challenges of pushing such theory into deployment

Aims

1. To give you a firm mathematical footing in anomaly detection, in a world where everyone is making simple and easily avoided mistakes
2. To make you fully appreciate the challenges of pushing such theory into deployment
    - Remember that anomaly detection finds model discrepancies, not necessarily features of interest. And big datasets are notoriously hard to model.

Course Outline

Week 1 *Fundamentals of hypothesis testing*, including the null and alternative hypotheses, the test statistic and the p-value, power and the Neyman-Pearson lemma

Week 2 *Combining p-values*, including Fisher's, Stouffer's, Simes', higher criticism and how you choose between them

Week 3 *Multiple testing*, including controlling false positives, the familywise error rate, and the false discovery rate

Week 4 *Miscellanea*, including handling discrete test statistics, Bayesian p-values, Monte Carlo p-values, and connections to machine-learning

Weeks 4/5/6 *Modelling dynamic network data*: tractable and scalable Bayesian techniques. This will include three problem classes, and assessed coursework handed out in Week 4 and **due by Thursday 20th December**

Software

For the problem classes you will need:

1. a laptop with access to internet
2. R, RStudio or other editor installed and working, and similarly python.

Further points and information

- My office hour is on Tuesdays 12:00-13:00, Office 4.15 (School of Maths)
- The lecture slides and other relevant material will usually be made available at the end of every week
- Towards the end of the unit, we'll run through a mock exam question