Anomaly Detection

Day 1: What is an anomaly? How do we spot them?
Day 2: Machine learning, clustering and classification
Day 3: Periodicity, and an entity centric view

Peter Tillotson

Acumed Consulting

June 5, 2016

Day I

What is an anomaly? How do we spot them?

Table of Contents

What we'll learn

By the end of today, we will have:

- a good understanding of anomaly detection
- discussed in detail a number of techniques commonly used to detect anomalies
  - linear regression
  - classification
  - clustering
- introduced Python Panda's and
- implemented a linear regression model in with Panda's

Prerequisites

Programming: basic understanding of programming concepts including: • flow control • functions • data structures • types (int, float)	Good
Python: be familiar with the Python	Some
Math: basic statistics (mean, variance)	Some

Definition: Anomaly

Anomaly from the Oxford Dictionary:

> *Something that deviates from what is standard, normal, or expected*

But that includes 5% of all data so what we're really interested in are events that are particularly odd or inherently risky. Ideally something that is actionable and that probably needs human intervention.

We want to avoid false positives, though whilst potentially rare these create noise and reduce an analysts trust of an automated solution.

We also want to avoid false negatives, real risks that we failed to spot.

Some code

```
import json
def my_function():
m_str = json.dumps({'name': 'peter', 'msg':'dont it look
    nice'})
print
my_function()
```

[1]

More often than not, time series data are ‘non-stationary’; that is, the values of the time series do not fluctuate around a constant mean or with a constant variance.

Day II

Machine learning, clustering and classification

Day III

Periodicity, and an entity centric view

What is Periodicity

Anything that repeats with a fixed interval can be said to periodic. If you can be guruanteed delivery of your signal, and have a trusted source for timestamps then it is possible to run Descrete Fourier Transforms and get a frequency domain representation of your signal.

Often this is not possible and the following approaches can be used to to handle trends and seasonality in data.

- Curve fitting - STL / Decomposition
- Twitter's Seasonal Hybrid ESD [2]
  - Time series decomposition
  - Generalised ESD
- Numenta's Hierarchical Temporal Memory [3]

Generalized ESD (extreme Studentized deviate) I

The Generalized ESD[4] test is defined for the two hypothesis:

H_0 There are no outliers in the data set

H_r There are up to r outliers in the data set

Compute:

$$R_i = \frac{\max_i |x_i - \mu|}{\sigma} \tag{1}$$

with μ and σ denoting the mean and standard deviation, respectively.

Remove the observation that maximizes $|x_i - \mu|$ and then recompute the above statistic with $n - 1$ observations. Repeat this process until r observations have been removed. This results in the r test statistics $R_1, R_2, \ldots, R_r$.

Generalized ESD (extreme Studentized deviate) II

Corresponding to the r test statistics, compute the following r critical values:

$$\lambda_i = \frac{(n-1)t_{p,n-i-1}}{\sqrt{(n-i-1+t^2_{p,n-i-1})(n-i+1)}} \tag{2}$$

for $i = 1, 2, \ldots, r$

where $t_{p,v}$ is the $100p$ percentage point from the t distribution with v degrees of freedom and

$$p = 1 - \frac{\alpha}{2(n-i+1)}$$

The number of outliers is the largest i such that $R_i > {}_i$.

References

[1] Y. B. Nikolay Laptev Saeed Amizadeh. (Mar. 2015), A benchmark dataset for time series anomaly detection, Yahoo, [Online]. Available: http://yahoolabs.tumblr.com/post/114590420346/a-benchmark-dataset-for-time-series-anomaly.

[2] A. Kejariwal. (Jan. 2015), Introducing practical and robust anomaly detection in a time series, [Online]. Available: https://blog.twitter.com/2015/introducing-practical-and-robust-anomaly-detection-in-a-time-series.

[3] J. Hawkins. (Sep. 2011), Hierarchical temporal memory, [Online]. Available: http://numenta.com/assets/pdf/whitepapers/hierarchical-temporal-memory-cortical-learning-algorithm-0.2.1-en.pdf.

[4] B. Rosner, "Percentage points for a generalized esd," *Technometrics*, vol. 25, no. 2, pp. 165–172, May 1983.