Anomaly Detection

State of the art algorithm breakdown

Peter Tillotson

Acumed Training

September 24, 2016

Table of Contents

Definition: Anomaly

Anomaly from the Oxford Dictionary:

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

What is normal?

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

data:yahoo

Methods for anomaly detection

- Curve fitting: Builing a model directly from signal stats
- Classification
- Clustering

Definition: Anomaly

- **Seasonality:**
- **Underlying trend**

Curve fitting

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.

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

Twitter's Seasonal Hybrid ESD

- Time series decomposition
- Generalised ESD (Extreme Studentized Deviate)

Generalized ESD (Extreme Studentized Deviate) I

The Generalized ESD**ref:rosner** 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