ANOMALY DETECTION

By : Anjana Yadav

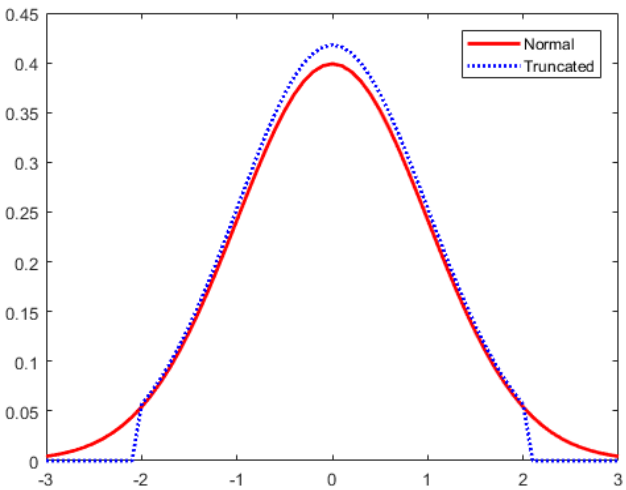
What is Anomaly Detection?

- Anomaly detection is the identification of items or events that do not conform to an expected pattern or to other items present. They indicate a variability in a measurement, experimental error or a novelty.
- It has many applications :
    - Intrusion detection : identifying strange patterns in network traffic that could be a signal hack
    - System health monitoring : spotting a malignant tumor in an MRI scan
    - Fraud detection in credit card transactions
- Anomalies can be broadly categorized as :
    - Point Anomalies are single instances of data if it's too far off from the rest. Eg. Detecting credit card fraud based on "amount spent."
    - Contextual Anomalies is context specific and is common in time-series data and Text analysis.
    - Collective Anomalies are a set of data instances that collectively helps in detecting anomalies. Eg. Someone trying to copy data from a remote machine to a local host unexpectedly could be flagged as a potential cyber attack.

What Algorithms can we use for Anomaly Detection?

STATISTICAL METHODS :

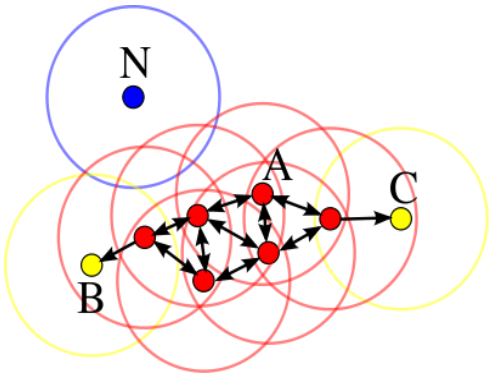
- Anomalous data point is the one that deviates by a certain standard deviation from the mean.
- Low Pass Filters :
    - Traversing mean over time-series data is not trivial.
    - We need a rolling window to compute the average across the data points.
    - Also called moving average, it's intended to smooth short-term fluctuations and highlight long-term ones.
- Z-Score or Extreme Value Analysis (EVA) :
    - The z-score is a metric that indicates how many standard deviations a data point is from the sample's mean, assuming a Gaussian distribution.
    - When computing the z-score for each sample on the data set a threshold must be specified.
    - By 'tagging' or removing the data points that lay beyond a given threshold we are classifying data into outliers and non outliers.



- Disadvantages:
    - The definition of threshold may frequently change, as malicious adversaries constantly adapt themselves.
    - Noise can be mistaken as anomaly.
    - Convenient for use in a low dimensional feature space and medium sized dataset.

MACHINE LEARNING

- Supervised
    - This method requires a labeled training set that contains both normal and anomalous samples for constructing the predictive model.
    - The most common supervised algorithms are Neural networks, support vector machine learning, k-nearest neighbors, Bayesian networks and decision trees.
    - Support Vector Machine :
        - SVM algorithm learns a soft boundary in order to cluster the normal data instances using the training set, and then, using the testing instance, it tunes itself to identify the abnormalities that fall outside the learned region.
    - Linear Models:
        - Projection methods that model the data into lower dimensions using linear correlations. Here data with large residual errors may be outliers.
- Unsupervised
    - These techniques do not require training data. The most common unsupervised algorithms are self-organizing maps (SOM), K-means, DBScan.
    - DBScan :
        - DBScan is a density based clustering algorithm, focused on finding neighbors by density (MinPts) on an ‘n-dimensional sphere’ with radius ε .
        - A cluster can be defined as the maximal set of ‘density connected points’ in the feature space.
        - It is sensitive to the MinPts parameter, tuning it will completely depend on the use case.
        - Core point : **A** is a core point if its neighborhood (defined by ε) contains at least the same number or more points than the parameter MinPts.
        - Border point : **C** is a border point that lies in a cluster and its neighborhood does not contain more points than MinPts.
        - Outlier : **N** is an outlier point that lies in no cluster and it is not ‘density reachable’ nor ‘density connected’ to any other point.
        - It is effective when the distribution of values in the feature space can not be assumed.
        - Selecting the optimal parameters eps, MinPts and metric can be difficult since it is very sensitive to any of the three params



What Datasets are available?

The following sites provide large variety of datasets based on the anomaly use case :

- Outlier Detection Datasets: ODDS provide a large variety of datasets based on the use case
  - http://odds.cs.stonybrook.edu/
- The UCSD annotated dataset is available
  - http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm
- University of Minnesota unusual crowd activity dataset :
  - http://mha.cs.umn.edu/Movies/Crowd-Activity-All.avi
- Signal Analysis for Machine Intelligence :
  - http://vision.eecs.yorku.ca/research/anomalous-behaviour-data/
- Kaggle Dataset :
  - The Numenta Anomaly Benchmark (NAB) is a novel benchmark for evaluating algorithms for anomaly detection in streaming and online applications.
  - It is comprised of over 50 labeled real-world and artificial Time series data files plus a novel scoring mechanism designed for real-time applications.
  - https://www.kaggle.com/boltzmannbrain/nab

REFERENCES :

https://towardsdatascience.com/a-brief-overview-of-outlier-detection-techniques-1e0b2c19e561
https://machinelearningmastery.com/how-to-identify-outliers-in-your-data/
https://www.datascience.com/blog/python-anomaly-detection
https://www.allerin.com/blog/machine-learning-for-anomaly-detection