Novelty / Outlier detection

Anomaly Detection

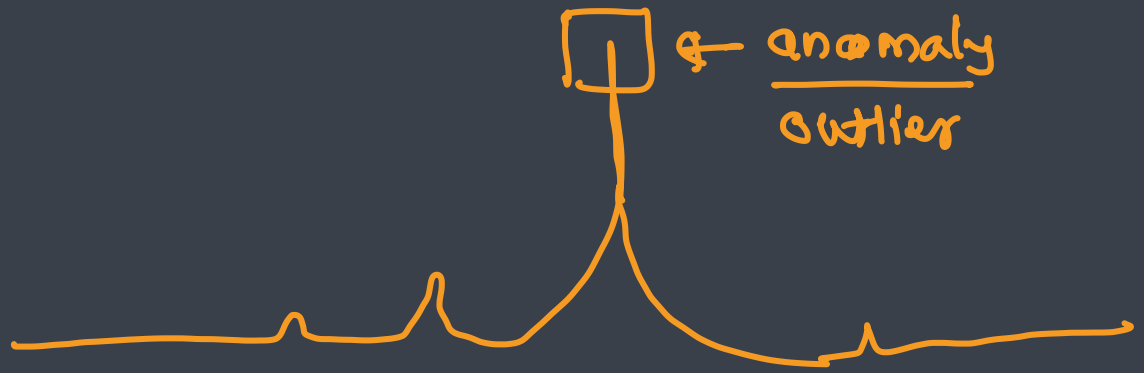
- ML
  - Supervised
  - unsupervised
    - clustering
- DL

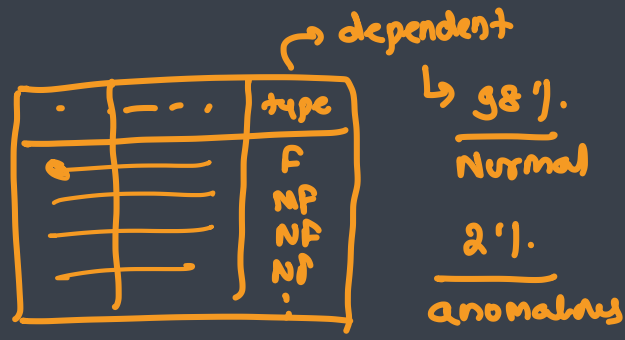
↓

rare pattern

Introduction

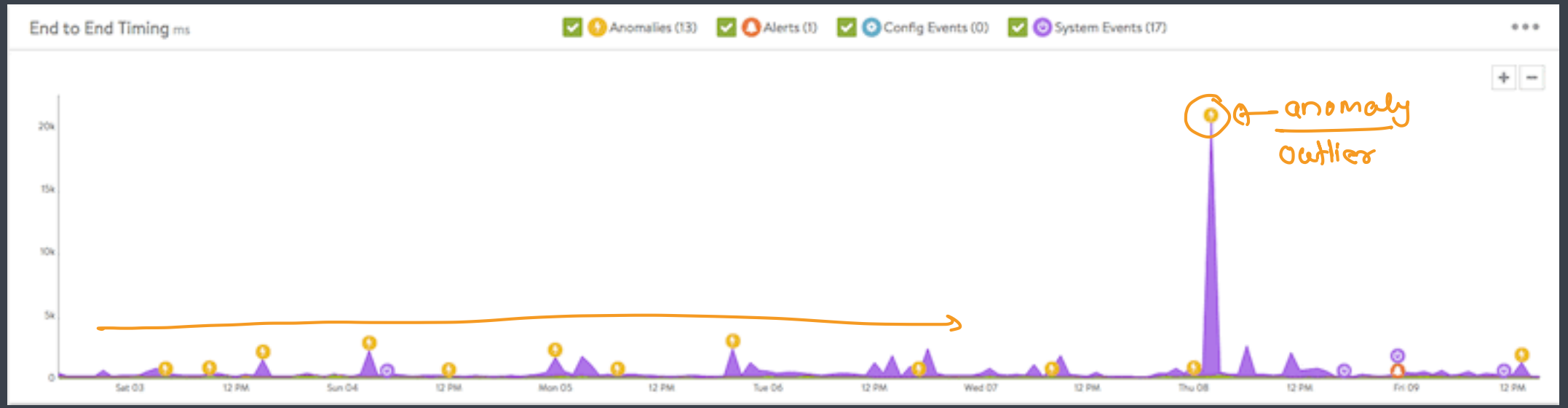
- Anomaly detection, also called outlier detection
- It is the identification of unexpected events, observations, or items that differ significantly from the normal
- Often applied to unlabelled data by data scientists in a process called unsupervised anomaly detection
- Any type of anomaly detection rests upon two basic assumptions:
  - Anomalies in data occur only very rarely → imbalanced
  - The features of data anomalies are significantly different from those of normal instances





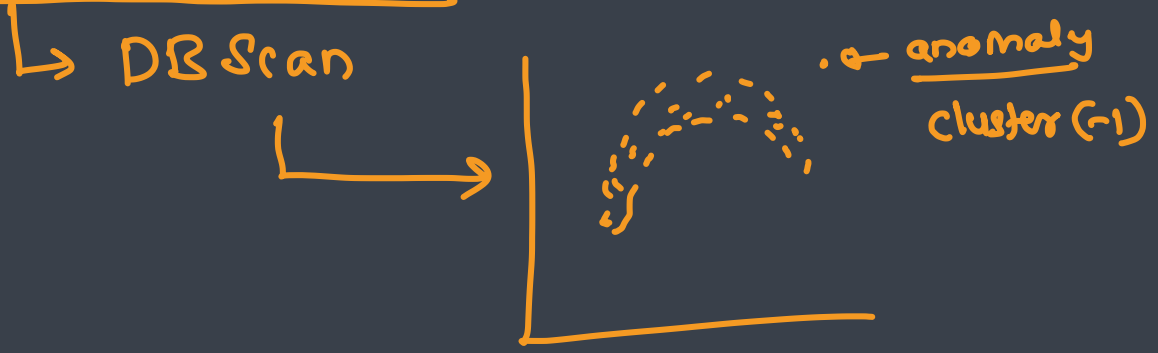
Introduction

- Typically, anomalous data is linked to some sort of problem or rare event such as hacking, bank fraud, malfunctioning equipment, structural defects / infrastructure failures, or textual errors
- For this reason, identifying actual anomalies rather than false positives or data noise is essential from a business perspective



What is it ?

- Anomaly detection is the identification of rare events, items, or observations which are suspicious because they differ significantly from standard behaviors or patterns
- Anomalies in data are also called standard deviations, outliers, noise, novelties, and exceptions
- In the network anomaly detection/ network intrusion and abuse detection context, interesting events are often not rare—just unusual
- For example, unexpected jumps in activity are typically notable, although such a spurt in activity may fall outside many traditional statistical anomaly detection techniques
- Many outlier detection methods, especially unsupervised techniques, do not detect this kind of sudden jump in activity as an outlier or rare object. However, these types of micro clusters can often be identified more readily by a cluster analysis algorithm.



Methods

types

- supervised → 100% complete labelled data
  - semi-supervised → partial labelled data
- unsupervised → 100% unlabelled
- reinforcement

semi-supervised → create a model with available data
→ use model to impute values in missing positions of dependent column
→ now the dependent column has 100% data

Anomaly detection methods

- There are three main classes of anomaly detection techniques:
  - Unsupervised
  - Semi-supervised → dependent variable is partially complete
  - Supervised → 100 % complete dependent variable
- Essentially, the correct anomaly detection method depends on the available labels in the dataset

Supervised method → classifier → 100% complete dependent variable

- Supervised anomaly detection techniques demand a data set with a complete set of "normal" and "abnormal" labels for a classification algorithm to work with
- This kind of technique also involves training the classifier
- This is similar to traditional pattern recognition, except that with outlier detection there is a naturally strong imbalance between the classes
- Not all statistical classification algorithms are well-suited for the inherently unbalanced nature of anomaly detection
- The advantage of supervised models is that they may offer a higher rate of detection than unsupervised techniques. This is because they can return a confidence score with model output, incorporate both data and prior knowledge, and encode interdependencies between variables.

↳ model → further prediction

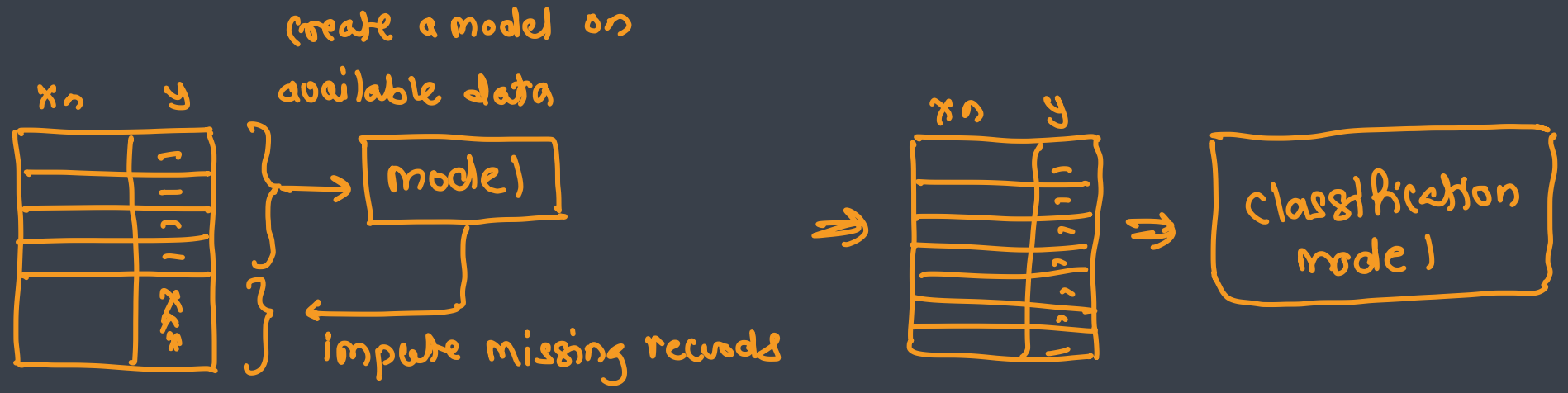
Nagios
Datadog

→ alert
↳ sms
↳ email

Semi-supervised method

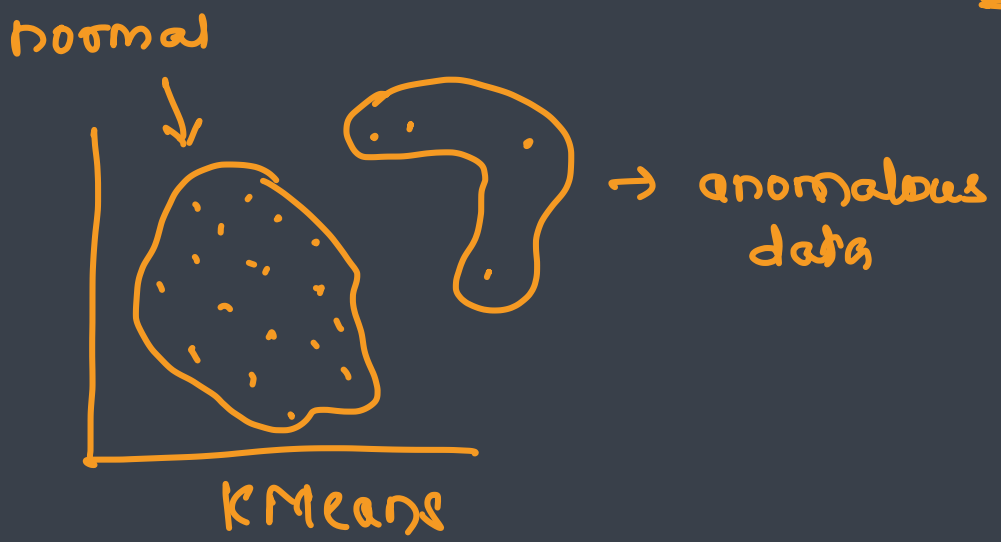
partially complete dependent variable

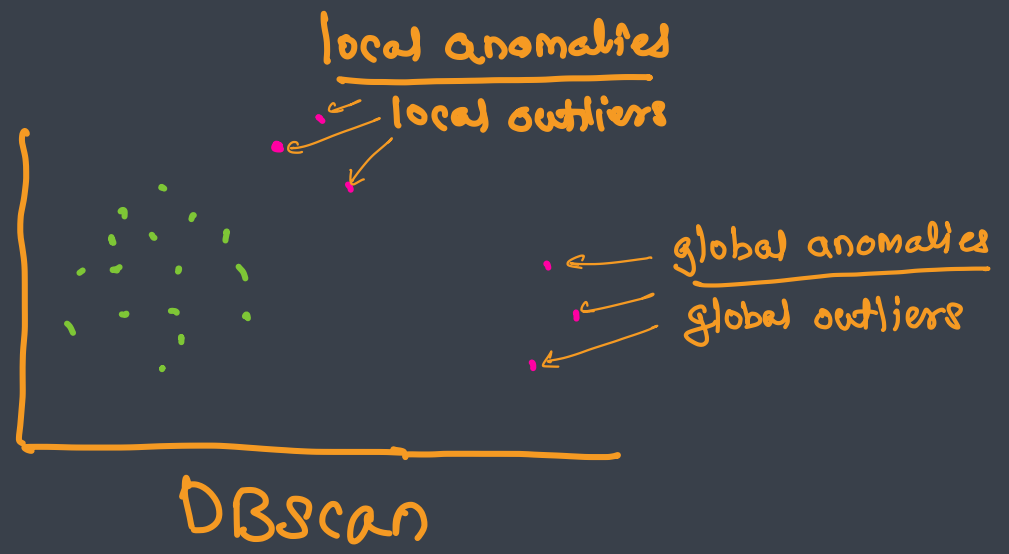
- Semi-supervised anomaly detection techniques use a normal, labelled training data set to construct a model representing normal behavior
- They then use that model to detect anomalies by testing how likely the model is to generate any one instance encountered
- A semi-supervised anomaly detection algorithm might also work with a data set that is partially flagged
- It will then build a classification algorithm on just that flagged subset of data, and use that model to predict the status of the remaining data



Unsupervised method → clustering → DBScan

- Unsupervised methods of anomaly detection detect anomalies in an unlabelled test set of data based solely on the intrinsic properties of that data
- The working assumption is that, as in most cases, the large majority of the instances in the data set will be normal 98%. → normal , 2%. → anomalous
- The anomaly detection algorithm will then detect instances that appear to fit with the rest of the data set least congruently
- The most popular unsupervised anomaly detection algorithms include Autoencoders, K-means, hypothesis tests-based analysis, and PCAs → DBScan





Types

Types of Anomalies

- **Network anomalies**
  - Anomalies in network behavior deviate from what is normal, standard, or expected
  - To detect network anomalies, network owners must have a concept of expected or normal behavior
  - Detection of anomalies in network behavior demands the continuous monitoring of a network for unexpected trends or events
- **Application performance anomalies**
  - These are simply anomalies detected by end-to-end application performance monitoring
  - These systems observe application function, collecting data on all problems, including supporting infrastructure and app dependencies
  - When anomalies are detected, rate limiting is triggered and admins are notified about the source of the issue with the problematic data
- **Web application security anomalies**
  - These include any other anomalous or suspicious web application behavior that might impact security such as CSS attacks or DDOS attacks.

Importance

Why is it important ?

- It is critical for network admins to be able to identify and react to changing operational conditions
- Any nuances in the operational conditions of data centers or cloud applications can signal unacceptable levels of business risk
- On the other hand, some divergences may point to positive growth
- Therefore, anomaly detection is central to extracting essential business insights and maintaining core operations

Importance

- Consider these patterns - all of which demand the ability to discern between normal and abnormal behavior precisely and correctly:
  - An online retail business must predict which discounts, events, or new products may trigger boosts in sales which will increase demand on their web servers
  - An IT security team must prevent hacking and needs to detect abnormal login patterns and user behaviors
  - A cloud provider has to allot traffic and services and has to assess changes to infrastructure in light of existing patterns in traffic and past resource failures
- A evidence-based, well-constructed behavioral model can not only represent data behavior, but also help users identify outliers and engage in meaningful predictive analysis
- Static alerts and thresholds are not enough, because of the overwhelming scale of the operational parameters, and because it’s too easy to miss anomalies in false positives or negatives
- To address these kinds of operational constraints, newer systems use smart algorithms for identifying outliers in seasonal time series data and accurately forecasting periodic data patterns

Techniques

Anomaly detection techniques

- In searching data for anomalies that are relatively rare, it is inevitable that the user will encounter relatively high levels of noise that could be similar to abnormal behavior
- This is because the line between abnormal and normal behavior is typically imprecise, and may change often as malicious attackers adapt their strategies
- Furthermore, because many data patterns are based on time and seasonality, there is additional baked-in complexity to anomaly detection techniques
- The need to break down multiple trends over time, for example, demands more sophisticated methods to identify actual changes in seasonality versus noise or anomalous data
- For all of these reasons, there are various anomaly detection techniques
- Depending on the circumstances, one might be better than others for a particular user or data set
- A generative approach creates a model based solely on examples of normal data from training and then evaluates each test case to see how well it fits the model

↳ classification

Clustering-Based Anomaly Detection → unsupervised

- Clustering-based anomaly detection remains popular in unsupervised learning
- It rests upon the assumption that similar data points tend to cluster together in groups, as determined by their proximity to local centroids
- K-means, a commonly-used clustering algorithm, creates ‘k’ similar clusters of data points. Users can then set systems to mark data instances that fall outside of these groups as data anomalies. As an unsupervised technique, clustering does not require any data labelling.
- Clustering algorithms might be deployed to capture an anomalous class of data. The algorithm has already created many data clusters on the training set in order to calculate the threshold for an anomalous event. It can then use this rule to create new clusters, presumably capturing new anomalous data.
- However, clustering does not always work for time series data. This is because the data depicts evolution over time, yet the technique produces a fixed set of clusters.

Density-Based Anomaly Detection

- Density-based anomaly detection techniques demand labelled data
- These anomaly detection methods rest upon the assumption that normal data points tend to occur in a dense neighbourhood, while anomalies pop up far away and sparsely
- There are two types of algorithms for this type of data anomaly evaluation:
  - K-nearest neighbor (k-NN) is a basic, non-parametric, supervised machine learning technique that can be used to either regress or classify data based on distance metrics such as Euclidean, Hamming, Manhattan, or Minkowski distance.
  - Local outlier factor (LOF), also called the relative density of data, is based on reachability distance

Support Vector Machine-Based Anomaly Detection → supervised

- A support vector machine (SVM) is typically used in supervised settings, but SVM extensions can also be used to identify anomalies for some unlabelled data
- A SVM is a neural network that is well-suited for classifying linearly separable binary patterns—obviously the better the separation is, the clearer the results.
- Such anomaly detection algorithms may learn a softer boundary depending on the goals to cluster the data instances and identify the abnormalities properly
- Depending on the situation, an anomaly detector like this might output numeric scalar values for various uses

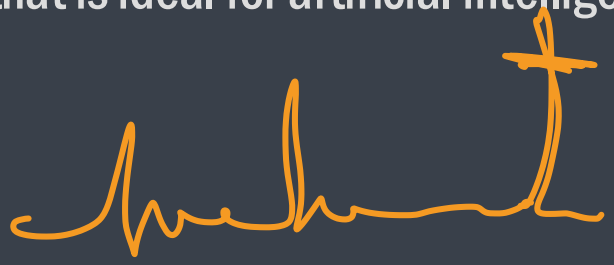
Use Cases

Anomaly Detection Use Cases

- Anomaly based intrusion detection
  - An anomaly based intrusion detection system (IDS) is any system designed to identify and prevent malicious activity in a computer network
  - A single computer may have its own IDS, called a Host Intrusion Detection System (HIDS), and such a system can also be scaled up to cover large networks. At that scale it is called Network Intrusion Detection (NIDS)
  - This is also sometimes called network behavior anomaly detection, and this is the kind of ongoing monitoring network behavior anomaly detection tools are designed to provide
  - Most IDS depend on signature-based or anomaly-based detection methods, but since signature-based IDS are ill-equipped to detect unique attacks, anomaly-based detection techniques remain more popular
- Fraud detection
  - Fraud in banking (credit card transactions, tax return claims, etc.), insurance claims (automobile, health, etc.), telecommunications, and other areas is a significant issue for both private business and governments
  - Fraud detection demands adaptation, detection, and prevention, all with data in real-time

Anomaly Detection Use Cases

- Data loss prevention (DLP)
  - Data loss prevention (DLP) is similar to prevention of fraud, but focuses exclusively on loss of sensitive information at an early stage
  - In practice, this means logging and analyzing accesses to file servers, databases, and other sources of information in near-real-time to detect uncommon access patterns
- Anomaly based malware detection
  - Malware detection is another important area, typically divided into feature extraction and clustering/classification stages
  - Sheer scale of data is a tremendous challenge here, along with the adaptive nature of the malicious behavior
- Medical anomaly detection
  - Detecting anomalies in medical images and records enables experts to diagnose and treat patients more effectively
  - Massive amounts of imbalanced data means reduced ability to detect and interpret patterns without these techniques
  - This is an area that is ideal for artificial intelligence given the tremendous amount of data processing involved



Anomaly Detection Use Cases

- Anomaly detection on social platforms
  - Detecting anomalies in a social network enables administrators to identify fake users, online fraudsters, predators, rumor-mongers, and spammers that can have serious business and social impact
- Log anomaly detection
  - Log anomaly detection enables businesses to determine why systems fail by reconstructing faults from patterns and past experiences
- Internet of things (IoT) big data system anomaly detection
  - Monitoring data generated in the field of the Internet of things (IoT) ensures that data generated by IT infrastructure components, radio-frequency identification (RFID) tags, weather stations, and other sensors are accurate and identifies faulty and fraudulent behavior before disaster strikes
  - The same is true of monitoring industrial systems such as high-temperature energy systems, power plants, wind turbines, and storage devices that are exposed to massive daily stress