# INTRODUCTION

In data mining, anomaly detection (also outlier detection) is the identification of rare items, events or observations which raise suspicions by differing significantly from the majority of the data. Typically, the anomalous items will translate to some kind of problem such as bank fraud, a structural defect, medical problems or errors in a text. Anomalies are also referred to as outliers, novelties, noise, deviations and exceptions.

## WHAT ARE ANOMALIES?

Before getting started, it is important to establish some boundaries on the definition of an anomaly. Anomalies can be broadly categorized as:

* **Point anomalies:** A single instance of data is anomalous if it's too far off from the rest. Business use case: Detecting credit card fraud based on "amount spent."
* **Contextual anomalies:** The abnormality is context specific. This type of anomaly is common in time-series data. Business use case: Spending $100 on food every day during the holiday season is normal, but may be odd otherwise.
* **Collective anomalies:** A set of data instances collectively helps in detecting anomalies. Business use case: Someone is trying to copy data form a remote machine to a local host unexpectedly, an anomaly that would be flagged as a potential cyber-attack.

## MULTIVARIATE ANOMALY DETECTION

As we have noted above, for identifying anomalies when dealing with one or two variables, data visualization can often be a good starting point. However, when scaling this up to high-dimensional data (which is often the case in practical applications), this approach becomes increasingly difficult. This is fortunately where multivariate statistics comes to help.

When dealing with a collection of data points, they will typically have a certain distribution (e.g. a Gaussian distribution). To detect anomalies in a more quantitative way, we first calculate the probability distribution p(x) from the data points. Then when a new example, x, comes in, we compare p(x) with a threshold r. If p(x)<r, it is considered as an anomaly. This is because normal examples tend to have a large p(x) while anomalous examples tend to have a small p(x).

# CATEGORIES OF ANOMALY DETECTION

Three broad categories of anomaly detection techniques exist.

* Unsupervised anomaly detection techniques.
* Supervised anomaly detection techniques.
* Semi-supervised anomaly detection techniques.

## UNSUPERVISED ANOMALY DETECTION

Unsupervised anomaly detection techniques detect anomalies in an unlabeled test data set under the assumption that the majority of the instances in the data set are normal by looking for instances that seem to fit least to the remainder of the data set.

## SUPERVISED ANOMALY DETECTION

Supervised anomaly detection techniques require a data set that has been labeled as "normal" and "abnormal" and involves training a classifier (the key difference to many other statistical classification problems is the inherent unbalanced nature of outlier detection).

## SEMI-SUPERVISED ANOMALY DETECTION

Semi-supervised anomaly detection techniques construct a model representing normal behavior from a given normal training data set, and then test the likelihood of a test instance to be generated by the learnt model.

## APPLICATIONS

Anomaly detection is applicable in a variety of domains, such as

* Intrusion detection.
* Fraud detection.
* Fault detection.
* System health monitoring.
* Event detection in sensor networks.
* Detecting ecosystem disturbances.

It is often used in preprocessing to remove anomalous data from the dataset. In supervised learning, removing the anomalous data from the dataset often results in a statistically significant increase in accuracy.

## POPULAR TECHNIQUES

Several anomaly detection techniques have been proposed in literature. Some of the popular techniques are:

1. Density-based techniques (k-nearest neighbor, local outlier factor, isolation forests, and many more).
2. One-class support vector machines.
3. Bayesian Networks.
4. Hidden Markov models (HMMs).
5. Cluster analysis-based outlier detection.
6. Fuzzy logic-based outlier detection.
7. Ensemble techniques, using feature bagging, score normalization and different sources of diversity.

The performance of different methods depends a lot on the data set and parameters, and methods have little systematic advantages over another when compared across many data sets and parameters.