A Project Report on

**Anomaly Detection in ML**

Submitted by :

**Abhishek Darekar**

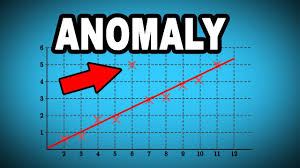
Chapter 1 : Introduction

### 1.1 What is an Anomaly?

Anomaly is something Abnormal/Outlier/Uncertainty.

Anomaly Detection to find out differences in the pattern to detect fraud in the system.

**In data mining, anomaly detection is referred to the identification of items or events that do not conform to an expected pattern or to other items present in a dataset. Typically, these anomalous items have the potential of getting translated into some kind of problems such as structural defects, errors or frauds. Using machine learning for anomaly detection helps in enhancing the speed of detection.**

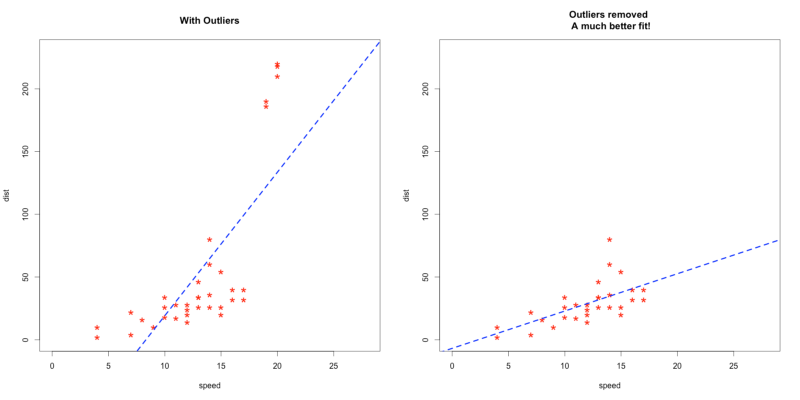


1. Hawkins defined anomaly as “an abbreviation which deviates sp much from the other observations as to arouse suspicious that it was generated by a different mechanism .
2. Anomaly detection has received considerable attention in the field of data mining due to the valuable insight that the detection of unusual events can provide in a variety of application
3. WE can use this to detect faulty sensor.
4. In anomaly detection Domain Knowledge is very important .
5. They are the Datapoints that are inconsistent with the distribution of majority of data points.
6. An example is the breed of dogs if a new species is seen it is novrl class and if any other breed comes in it is outlier .

### Why do we need to Detect Outliers?

### Outliers can impact the results of our analysis and statistical modeling in a drastic way.

### Check out the below image to visualize what happens to a model when outliers are present versus when they have been dealt with:



### 1.3 Novelty Detection vs Anomaly Detection

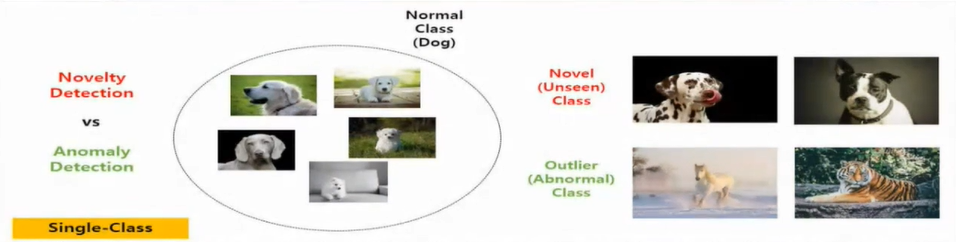
1.Novel(Unseen) Class :

- New unseen data which is similar to our data (Some Differences).

2.Outlier(Abnormal ) Class :

- New unseen data which is very dissimilar than our data.

Example :-



### 1.4 Types of Anomaly Detection : (There are many, we will look at some of the most common)

1. **Time-series anomaly** is like attack on a system .
2. **Video-Level** **Detection** : In Banks, ATM and other important places in cctv recording we can set certain limits and category as to if someone does any unwanted behaviour it will set an alarm .

-Video is mostly very expensive to store so, where it is not possible we use image analysis as alternative

1. **Image-level detection** : Can be used in cases where human cannot check the similarity and 2 two images we can find the percentage of similarity in 2 pictures .There are 3 types of categories .
   1. Anomaly Classification target (ex : Classify Aadhar card using photo on it.).
   2. Out-of-Distribution Detection (ex : Blur image, so machine doesn’t know what value to take in blurred portion).
   3. Anomaly Segmentation Target.

# 1.5 Anomaly Detection — Business Benefits

## Intrusion detection :

* Any nefarious activity that can damage an information system can be broadly classified as an intrusion.
* Anomaly detection can be effective in both detecting and solving intrusions of any kind.
* Common data-centric intrusions include cyberattacks, data breaches, or even data defects.

1. Mobile sensor data :

* For instance, a particular industry case study is that of the IBM Data Science Experience that developed a tool for anomaly detection using Jupyter Notebook for capturing sensor data from mobile phones and connected IoT devices.

### 1.6 Anomaly detection can be done using the concepts of Machine Learning. It can be done in the following ways :

1. **Supervised Anomaly Detection**:

* This method requires a labeled dataset containing both normal and anomalous samples to construct a predictive model to classify future data points.
* The most commonly used algorithms for this purpose are supervised Neural Networks, Support Vector Machine learning, K-Nearest Neighbors Classifier, etc.

1. **Unsupervised Anomaly Detection**:

* This method does require any training data and instead assumes two things about the data ie Only a small percentage of data is anomalous and Any anomaly is statistically different from the normal samples.
* Based on the above assumptions, the data is then clustered using a similarity measure and the data points which are far off from the cluster are considered to be anomalies.

### 1.7 PyOD (Python Outlier Detection Package) :

1. PyOD is a comprehensive and scalable python **toolkit for detecting outlying** objects in multivariate data.
2. It was developed back in 2017 and has been used in many academic research and commercial products

#### PyOD Uses **:**

1. It is featured for Unified API’s , detailed documentation and interactive examples across Various Algorithms.
2. In Advanced models, including Neural Networks and outlier Ensembles
3. Optimized performance with **JIT and parallelization** when possible, using **namba** and **joblib**.

### 1.8 Benchmark of Various outlier detection models :

1. **Linear Models** for Outlier Detection: When one increases or decreases with respect to other it is linear.
   1. Principal Component Analysis (**PCA**): Based on the contribution can we remove any of the feature and choose the most important ones is PCA. We remove least Imp features.
   2. Minimum Covariance Determinant (**MCD**): Covariance is the difference between std deviation and other variance values. So, using a limit from the midpoint we can detect outliers after a range.
   3. One-Class Support Vector Machine (**OCSVM**): We can take all the in-liner and remove the uncertain and out of the line problems.
2. **Proximity Based** Outlier Detection Models: Using the proximity to detect the outliers
   1. Local Outlier Factor (**LOF**)**:** Proximity Based Outlier ex: Car alarm rings when car gets close to some other car behind it.
   2. Clustering Based LOF (**CBLOF**)

* It classifies the data into small clusters and large clusters. The anomaly score is then calculated based on the size of the cluster the point belongs to, as well as the distance to the nearest large cluster
  1. **KNN** (uses distance to K-th nearest neighbour as Outlier)
* For any data point, the distance to its kth nearest neighbor could be viewed as the outlying score
* PyOD supports three kNN detectors:
* Largest: Uses the distance of the kth neighbor as the outlier score
* Mean: Uses the average of all k neighbors as the outlier score
* Median: Uses the median of the distance to k neighbors as the outlier score
  1. Histogram Based Outlier Score (**HBOS**)
* It is an efficient unsupervised method which assumes the feature independence and calculates the outlier score by building histograms
* It is much faster than multivariate approaches, but at the cost of less precision

1. **Probability Model** for outlier Detection:
   1. Angle-Based Outlier Detection (**ABOD**)

* It considers the relationship between each point and its neighbour(s). It does not consider the relationships among these neighbours. The variance of its weighted cosine scores to all neighbours could be viewed as the outlying score
* ABOD performs well on multi-dimensional data
* PyOD provides two different versions of ABOD:
* Fast ABOD: Uses k-nearest neighbours to approximate
* Original ABOD: Considers all training points with high-time complexity

1. **Ensemble** and **combination Framework**
   1. Isolation Forest

* It uses the scikit-learn library internally. In this method, data partitioning is done using a set of trees. Isolation Forest provides an anomaly score looking at how isolated the point is in the structure. The anomaly score is then used to identify outliers from normal observations
* Isolation Forest performs well on multi-dimensional data
  1. Feature bagging
* A feature bagging detector fits a number of base detectors on various sub-samples of the dataset. It uses averaging or other combination methods to improve the prediction accuracy
* By default, Local Outlier Factor (LOF) is used as the base estimator. However, any estimator could be used as the base estimator, such as kNN and ABOD
* Feature bagging first constructs n sub-samples by randomly selecting a subset of features. This brings out the diversity of base estimators. Finally, the prediction score is generated by averaging or taking the maximum of all base detectors