Anomaly Detection

MATH 4422 Project

Project Topic: Anomaly Detection

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.

Project Objective: Data Accuracy

Anomaly detection is used in preprocessing to remove anomalous data from the dataset. This is done for a number of reasons. Statistics of data such as the mean and standard deviation are more accurate after the removal of anomalies, and the visualization of data can also be improved. In supervised learning, removing the anomalous data from the dataset often results in a statistically significant increase in accuracy. Anomalies are also often the most important observations in the data to be found such as in intrusion detection or detecting abnormalities in highly sensitive datapoints.

Theoretical Background: Definition

Many attempts have been made in the statistical and computer science communities to define an anomaly. The most prevalent ones include:

1. An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism.
2. Anomalies are instances or collections of data that occur very rarely in the data set and whose features differ significantly from most of the data.
3. An outlier is an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data.
4. An anomaly is a point or collection of points that is relatively distant from other points in multi-dimensional space of features.
5.Anomalies are patterns in data that do not conform to a well defined notion of normal behavior.
6. Let T be observations from a univariate Gaussian distribution and O a point from T. Then the z-score for O is greater than a pre-selected threshold if and only O is an outlier.

Theoretical Background: Why detect anomalies?

Data is a very important asset in our current world. The more data one has the more one can curve out possibilities that may otherwise seem unavailable. Data gathering is happening every moment. The data of our mobile, laptop, internet, the data in our electric meter, gas meter, financial transaction tools, data collected about weather, traffic, celestial events. Data all around us is being collected endlessly. But in this collection procedure we often collect data which are unwanted or arouses suspicion. These are generally referred to as anomalies. Some anomalies harm our data collection model distorting output or destroying desired data points. Some anomalies enhance our understanding of certain behavioral models. So anomalies can be considered good or bad but the importance of detecting them is much more visible.

Data can control major events in the world, it can create or destroy many things for instance one starts a business and in order to upgrade their business they need solid data evidence of their production, supply, consumer status, financial aspect etc. Now if they were to collect anomalies in their data they may face serious financial backlash which may inadvertently destroy their entire business. So it is essential we detect anomalies from Data and then we can judge and take a decision as what to do with the anomalies.