Anomaly Detection in Streaming Data

By Harsh Kr.

1. What is an Anomaly: An anomaly, or outlier, is an observation in a dataset that significantly deviates from the norm.

2. Data Stream Generation: Initially, we create a random data stream to simulate various types of data. This provides a base dataset that resembles real-world data without specific patterns. A seasonal pattern is incorporated into the data to add periodic changes. This is achieved by generating a sinusoidal function, which models regular, repeating fluctuations. To simulate real-world data variability, random noise is added. This noise introduces randomness and ensures the data isn't too predictable or smooth. Occasional anomalies are introduced to mimic rare but significant deviations from the normal data. These are the points that we need to identify using our model.

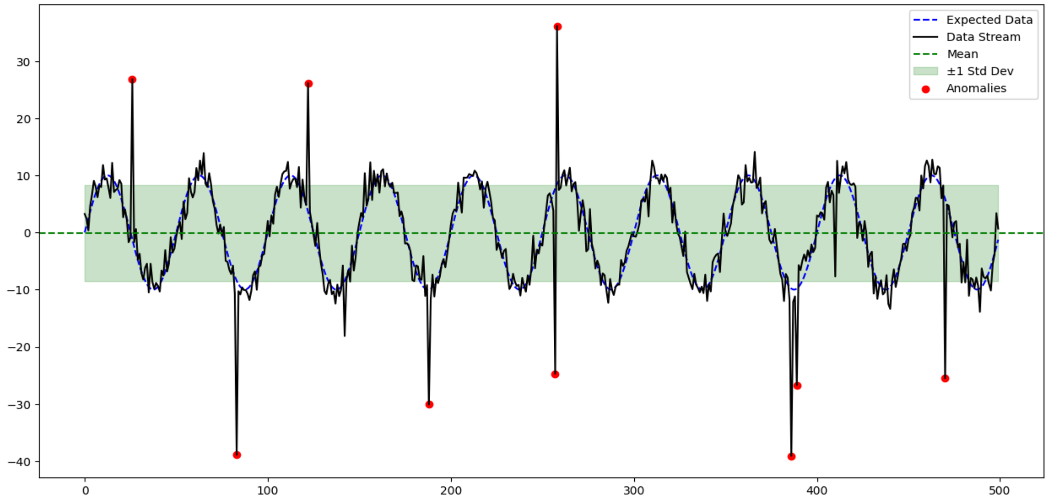
3. Methods for Anomaly Detection: The following are some ways from which we can find out the anomalies from a dataset.

- From Standard Deviation of previous data.
- Z-score Based Detection.
- Exponentially Weighted Moving Average Based Detection.
- Using Machine Learning Models like SGD classifier..

3.1. Using Standard Deviation

The core idea is to identify data points that deviate significantly from the norm. This technique starts by calculating the mean and standard deviation of all previous data points up to the current point in the data stream. The mean provides a central value around which the data points are expected to cluster, while the standard deviation measures the average distance of each data point from this mean, indicating the data's variability. As new data points arrive, we compute the residual, which is the absolute difference between the current data point and the mean of the previous data. This residual is then compared against a threshold value, which is a multiple of the standard deviation. If the residual exceeds this threshold, the data point is flagged as an anomaly.

The threshold factor determines the sensitivity of the anomaly detection process. A higher threshold factor means the method will flag fewer anomalies, as it requires a larger deviation from the mean to be considered abnormal.

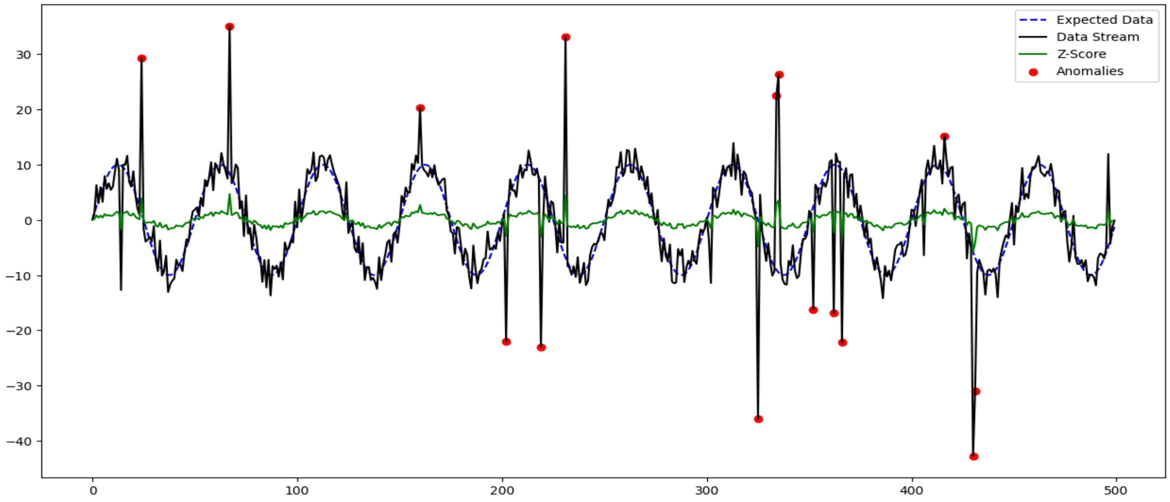


The standard deviation-based anomaly detection method has limitations due to its sensitivity to changes in data distribution and assumptions of normality. As data evolves, the mean and standard deviation may not accurately reflect current patterns, leading to false positives or negatives. This method also struggles with non-normally distributed data or when anomalies are not extreme deviations from the mean. The Z-score method improves on this by normalizing deviations, making it more robust to varying data patterns and providing a more reliable measure of anomaly significance.

3.2. Z-Score Based

In the Z-score-based anomaly detection method, we calculate the Z-score for each data point by measuring how many standard deviations it deviates from the mean of the entire dataset. This is done by subtracting the mean from the data point and dividing by the standard deviation. Anomalies are identified when the absolute value of the Z-score exceeds a predefined threshold.

This approach helps normalize the data, making it easier to identify outliers regardless of the underlying data distribution. By using Z-scores, we can effectively pinpoint significant deviations, offering a more consistent and reliable anomaly detection compared to the standard deviation method, especially in datasets with varying or non-normal distributions.



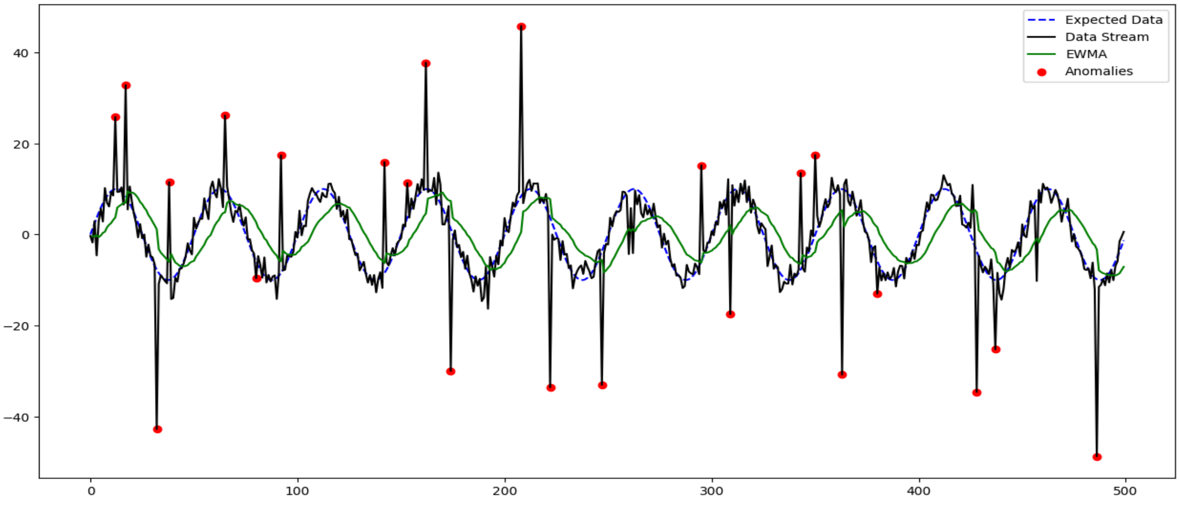
This approach works well for normally distributed data but falls short when data patterns change over time or when the data is not normally distributed. Its static nature makes it less effective in adapting to new trends or shifts in data.

The Exponentially Weighted Moving Average (EWMA) method addresses these limitations by giving more weight to recent data points. This adaptive approach allows EWMA to respond more effectively to changes in data patterns and provides a more dynamic measure of what constitutes normal behavior, making it better suited for real-time anomaly detection in evolving datasets.

3.3. EWMA Based

The Exponentially Weighted Moving Average (EWMA) method is an advanced technique used for anomaly detection based upon the principle of smoothing recent observations more heavily than older ones. This method gives greater weight to more recent values, which allows it to adapt to recent trends and fluctuations more effectively. By adjusting the weighting factor, EWMA can be tuned to detect anomalies that significantly deviate from the most recent patterns, making it particularly useful for datasets with trends and seasonal variations. This approach helps in setting a dynamic threshold that evolves with the data, improving the detection of outliers that deviate from the current data behavior.

EWMA works by applying a decay factor, often referred to as the smoothing parameter or alpha (α), to control the rate at which older observations are discounted. This parameter determines the weight of the most recent data relative to previous observations. A higher alpha value means that recent observations are weighted more heavily, making the model more responsive to recent changes, but potentially more susceptible to short-term noise. Conversely, a lower alpha value results in a smoother trend that reacts more slowly to recent changes, which can be beneficial for identifying more stable patterns but may overlook sudden anomalies. This balance is crucial for tuning the EWMA to effectively capture the nuances in the data stream while minimizing false positives or missed detections.



Despite its advantages, EWMA has some limitations. One key drawback is its sensitivity to the choice of the smoothing parameter, which can influence the method's ability to detect anomalies. A poorly chosen parameter might either overreact to minor fluctuations or fail to detect significant anomalies. Moreover, EWMA may not perform well in the presence of abrupt changes or anomalies that deviate drastically from both recent and historical data, as it primarily focuses on recent trends.

To address these limitations, machine learning-based anomaly detection methods, such as SGD classifier or isolation forest can offer a more robust alternative. Although these are computation-heavy methods, they ensure a precise prediction irrespective of the data.