Efficient Data Stream Anomaly Detection

Project Overview

This project implements a real-time anomaly detection system using the Z-Score method. The system continuously simulates a data stream that represents various metrics, such as financial transactions or system metrics. The primary objective is to identify unusual patterns in the data that may indicate anomalies.

Code Structure

Function Descriptions

- **`simulate_data_stream`**
  - **Description:** This function generates a continuous stream of data points that simulate real-time data. The data points are derived from a sine wave function combined with random noise to create a realistic data flow.
  - **Usage:**

    ```
    stream = simulate_data_stream()
    ```

- **`update_mean_std`**
  - **Description:** This function updates the rolling mean and standard deviation for a specified window size as new data points are added. It ensures the calculations remain accurate by adjusting the mean and standard deviation based on the latest data.
  - **Parameters:**
    - `prev_mean`: The previous mean of the data window.
    - `prev_std`: The previous standard deviation of the data window.
    - `new_data`: The new data point being added.
    - `old_data`: The old data point being removed from the window.
    - `window_size`: The size of the rolling window.
  - **Returns:** A tuple containing the updated mean and standard deviation.
- **`detect_and_visualize_anomalies`**
  - **Description:** This function detects anomalies in real-time data using the Z-Score method. It visualizes the data stream and highlights detected anomalies on a plot.
  - **Parameters:**
    - `data_stream`: A generator that yields real-time data points.
    - `window_size`: The size of the rolling window for calculating the mean and standard deviation (default is 30).
    - `threshold`: The Z-Score threshold for detecting anomalies (default is 3).

- **Returns:** None. The function continuously updates the plot as new data is processed.

Z-Score Method Explanation

The Z-Score method is a statistical technique used to determine how far a data point is from the mean in terms of standard deviations. The Z-Score is calculated using the formula:

$Z=(X−\mu)\sigma Z = \frac{(X - \mu)}{\sigma}Z=\sigma(X−\mu)$

Where:

- XXX: The current data point.
- $\mu\mu$: The mean of the data points in the rolling window.
- $\sigma\sigma$: The standard deviation of the data points in the rolling window.

An anomaly is flagged if the absolute value of the Z-Score exceeds a predefined threshold (e.g., 3), indicating that the data point is significantly different from the expected range.

Running the Project

To run the project, ensure you have Python 3.x installed along with the required libraries. You can install the necessary libraries using the following command:

```
pip install numpy matplotlib
```

Then, run.