**Simple Anomaly Detection algorithms for Streaming Data — Machine Learning**

[[Leonardo Mauro P. Moraes](https://medium.com/@leomaurodesenv?source=post_page-----92cfaeb6f43b--------------------------------)](https://medium.com/@leomaurodesenv?source=post_page-----92cfaeb6f43b--------------------------------)

[[Sinch Blog](https://medium.com/wearesinch?source=post_page-----92cfaeb6f43b--------------------------------)](https://medium.com/wearesinch?source=post_page-----92cfaeb6f43b--------------------------------)

[Leonardo Mauro P. Moraes](https://medium.com/@leomaurodesenv?source=post_page-----92cfaeb6f43b--------------------------------)

·

[Follow](https://medium.com/m/signin?actionUrl=https%3A%2F%2Fmedium.com%2F_%2Fsubscribe%2Fuser%2Ff4d6e632704c&operation=register&redirect=https%3A%2F%2Fmedium.com%2Fwearesinch%2Fsimple-anomaly-detection-algorithms-for-streaming-data-machine-learning-92cfaeb6f43b&user=Leonardo+Mauro+P.+Moraes&userId=f4d6e632704c&source=post_page-f4d6e632704c----92cfaeb6f43b---------------------post_header-----------)

Published in

[Sinch Blog](https://medium.com/wearesinch?source=post_page-----92cfaeb6f43b--------------------------------)

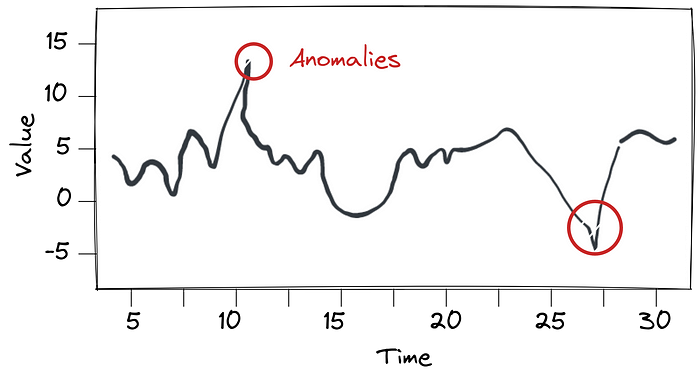
·

4 min read

·

Jan 13, 2022

81



by [Leonardo Mauro P. Moraes](https://www.linkedin.com/in/leomaurodesenv/)

**Abstract**

Have you ever heard about Anomaly Detection for Streaming Data? Basically, Streaming Data is data that is continuously generated over time. For example, the number of songs downloaded per time, internet usage per time, travel time on a road along the day. As you can see, this data changes over time. Meanwhile, Anomaly Detection refers to Machine Learning techniques to detect suddenly peaks or drop-offs in the data.

See some practical examples for this kind of machine learning technique: (1) detect high time travel on roads to offer another alternative route; (2) detect trend topics in Twitter volume data; (3) detect peaks in CPU usage to automatically provide more cloud computers. In general, you can detect anomalies/outliers in any kind of Streaming Data.

In this article, we are going to learn two simple techniques to detect anomalies in near real-time streaming data: *Moving Average* and *Exponential Moving Average*. To demonstrate the generality of the models, we are going to explore four different types of Streaming Data.

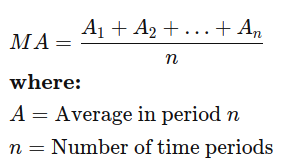
*The****code****is available at the end of the article.*

**What are the algorithms requirements?**

* Predictions must be made online; they cannot look ahead
* Algorithms must run automatically and unsupervised
* Algorithms must learn continuously and adapt to dynamic environments

**Moving Average**

Moving Average is a common type of average used in Streaming Data problems. Moving Average is a calculation used to analyze/smooth data by creating a new series of averages from subsets of the data points. In other words, this technique simply averages the *n* latest data points.



Moving Average formula

*See this one-minute video for better understanding:*

How to Calculate the Moving Average

In this way, a new averaged data results from the average of the previous *n* data points. This calculation happens for every data point in your dataset. Thus, the algorithm needs a parameter *n* to indicate how many periods it will use to compute the average. Here, we used the “averaged data” as “expected value”, thereby we can compare the current value with the expectation.

Meanwhile, how can we predict whether a value is an anomaly or not? To achieve this, we define a dynamic range from the “expected value” ***x\_value***. The range is computed from the standard deviation ***std*** from the *n* lastest points. Thus, the range of a non-anomaly is from (***x\_value*** - **std**) to (***x\_value***+ ***std***). So, if the value is in the range, it is not an anomaly; but if the value is out of the range, it is an anomaly/outlier.

**Note**. At this moment, we have created a simple and powerful algorithm to detect anomalies in Streaming Data. However, the simpleness of this algorithm is only able to treat simple Streaming Data where the data increases and decreases uniformly or with small changes over time. While, we have the *Exponential Moving Average* algorithm that is able to fit to more changes in the data, such as seasonal data.

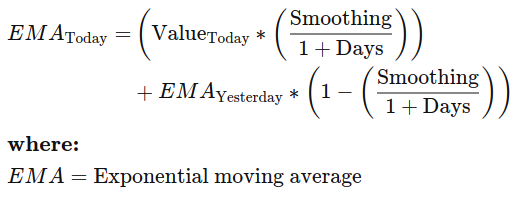


Types of Streaming Data

**Exponential Moving Average**

Exponential Moving Average is a weighted average algorithm that focuses on the most recent data by assigning more weight and significance to the most recent data; roughly, it is a Moving Average with weights. “An exponentially weighted moving average reacts more significantly to recent data changes than a simple moving average, which applies an equal weight to all observations in the period.” (Infopedia)

Also, the Exponential Moving Average needs a parameter called *alpha* (or smoothing factor) which determines the importance of the last record, and it’s decaying for the next records. For example, if *alpha=0.5*: record-1 has 50% importance, record-2 has 30% importance, and so on. Thus, the algorithm allows smoothing the “expected value”, prioritizing the most recent data.



Exponential Moving Average formula

Finally, the anomaly detection happens in the same way as the Moving Average. So, we just have to compute the standard deviation and define a range for non-anomaly data points as seen before. Experiments with real data showed that the Exponential Moving Average performs better than Moving Average in data with high variance (e.g., many peaks); also, it learns faster due to its smoothing factor.

**Discussion**

We learned two simple algorithms to automatically detect anomalies in Streaming Data. If you want to see the robustness of these algorithms, consider accessing Jupyter Notebook on Kaggle. The notebook presents the implementation of both algorithms, as well as experiments in real data, such as detect outliers in CPU utilization and network usage.

***Curiosity.****In finance, moving average techniques are stock indicators commonly used in technical analysis. The reason for calculating the moving average of a stock is to help smooth out the price data by creating a constantly updated average price. (*[*Investopedia*](https://www.investopedia.com/terms/m/movingaverage.asp)*)*

**Kaggle — Jupyter Notebooks**

We have the first notebook describing and detailing the techniques; the second is just an evaluation in another dataset; and the last one is an extension with new algorithm developed by [Mani Sarkar](https://www.kaggle.com/neomatrix369).

* [leomauro/anomaly-detection-streaming-data/](https://www.kaggle.com/leomauro/anomaly-detection-streaming-data/)
* [leomauro/anomaly-detection-streaming-data-evaluation/](https://www.kaggle.com/leomauro/anomaly-detection-streaming-data-evaluation)
* [neomatrix369/anomaly-detection-streaming-data-extended/](https://www.kaggle.com/neomatrix369/anomaly-detection-streaming-data-extended/)

**References**

* [Moving averages (Rob J Hyndman et al.)](https://otexts.com/fpp2/moving-averages.html)
* [Moving Average (MA) (investopedia)](https://www.investopedia.com/terms/m/movingaverage.asp)
* [Exponential Moving Average (EMA) (investopedia)](https://www.investopedia.com/terms/e/ema.asp)
* [Moving average and exponential smoothing models (Robert Nau)](https://people.duke.edu/~rnau/411avg.htm)