**Overview**

Electricity theft detection using a data driven approach, although have a very promising outcome, is prone to produce false results due to erroneous and noisy data consumption patterns spitted out by the smart meters due to either failure of smart meters, unscheduled maintenance or unreliable transmission of data. This erroneous data must be dealt with before applying any data analytics or machine learning approaches to detect anomalies.

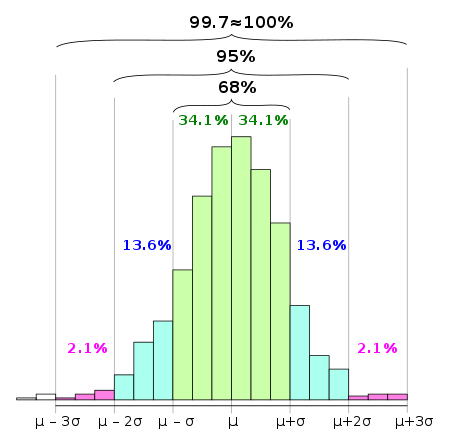
More often than not there are erroneous values (particularly outliers) that may result in the data pattern due the above mentioned reasons. These outliers are treated in [1], with the “Three-sigma rule of thumb”.

**Three-Sigma Rule of Thumb**

The three sigma rule of thumb states that for values of a random variable in a normal distribution an **event is considered to be practically impossible if it lies outside three standard deviations of the mean**. Even for many reasonably symmetric unimodal distributions**, almost all of the population lies within three standard deviations of the mean**. This is **supported by Chebyshev's inequality which guarantees that, for a wide class of**[**probability distributions**](https://en.wikipedia.org/wiki/Probability_distributions)**, no more than a certain fraction of values can be more than a certain distance from the**[**mean**](https://en.wikipedia.org/wiki/Expected_value)**.**

If X is a normally distributed random variable then mathematically almost 99.7% values lies within 3 standard deviations from the mean as

Graphically;



In [1], the three-sigma rule of has been slightly modified to suit the particular dataset. **Any value outside the 2 standard deviation limit is treated as an outlier**. Therefore even if our dataset doesn’t follow a Gaussian distribution, taking 2std would compensate for most of the values. Hence if denotes the daily energy consumption for a consumer for one day and is vector consisting of multiple values then an outlier is corrected by the function;

Suppose if is vector containing probability distribution of a random variable then the python code for getting the range of 3 sigma rule is,

References

[1]Z. Zheng, Y. Yang, X. Niu, H. Dai and Y. Zhau, "Wide and Deep Convolutional Neural Networks for Electricity-Theft Detection to Secure Smart Grids - IEEE Journals & Magazine", *Ieeexplore.ieee.org*, 2017. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8233155. [Accessed: 13- Dec- 2019].