CMPT 318 – CYBER SECURITY

DR. UWE GLASSER, FALL 2021

ASSIGNMENT 2.

1. Answers:

Feature scaling is an essential preprocessing technique of input data for machine learning algorithms and beyond. Two widely used feature scaling techniques are normalization and standardization. For each of the following questions explain clearly and concisely:

(I) What is feature scaling?

- Feature scaling is a type of modeling that helps raw data that a human can understand be transformed into an input set of data that can be understood by a machine.
- Feature scaling also helps us converge data easier because of proper scaling.

(II) Why scaling features of a dataset is necessary?

- Gives us an insight of how the data behaves relative to a number, this not just makes it easier for a person to see certain patterns, normalization or standardization of data makes it easier for a machine to also detect anomalies in the system.
- Clustering algorithms (Machine Learning) will fail as they will be skewed by different (not standardized nor normalized) scales as they rely on distance metrics.

(III) What does normalization and standardization do to the data and the noise? Your answers to all three questions should not exceed 2 pages in total but provide technical descriptions including the use of mathematical notation.

Assignment 2

Tuesday, October 26, 2021 8:56 PM

(III)

What does normalization and standardization do to the data and the noise? Your answers to all three questions should not exceed 2 pages in total but provide technical descriptions including the use of mathematical notation.

Normalization – set the value from a min, max : 0 to 1

Standardization – set the value σ from the mean
(usually you are -3 to 3 sd from mean)

let σ stand and deviation of data.

Normalization steps:

$$X_{normal} = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

let X be a feature
where $X = \{1, \ldots, n\}$

This just normalizes the data to be in the set of 0 to 1

ie let $X = [1, 3, 5, 7]$ apply formula and

$X_{normal} = [0, 1/3, 2/3, 1]$ ← notice here that
maxima = 7 is 1
minima = 1 is 0

So range of $X_{normal} = [0, 1]$ exclusively.

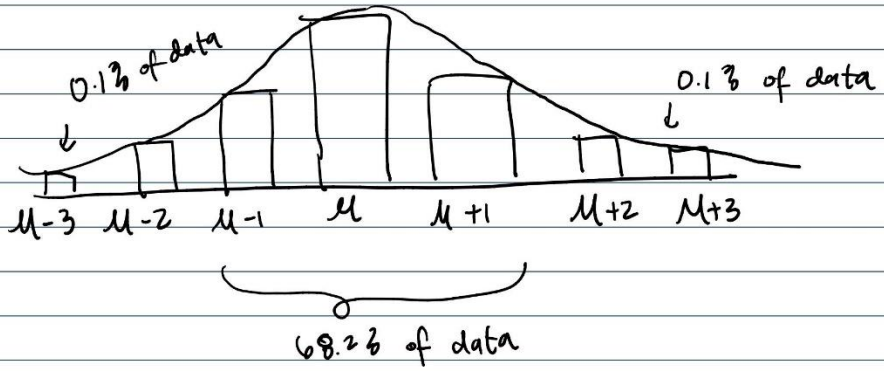
Standardization steps

$$X_{standard} = \frac{X - M}{\sigma}$$

let M is mean of dataset
let σ is stdeviation of data

This sets our data into a universe where you must be a certain standard deviation from the mean.

eg



The dataset will fall in range of $[-3, 3]$ this value means how far away you are from the mean.

Note: data is not exclusively to $[-3, 3]$ but anything farther away is very slim that it is improbable.

2):

MODEL DATA SET.

This set of data seems to be consistent with the flow of the output for the global intensity.

Some model numbers:

Week 1 to Week 10 (excluding week 8)

Mean range: [5.5,6.5]

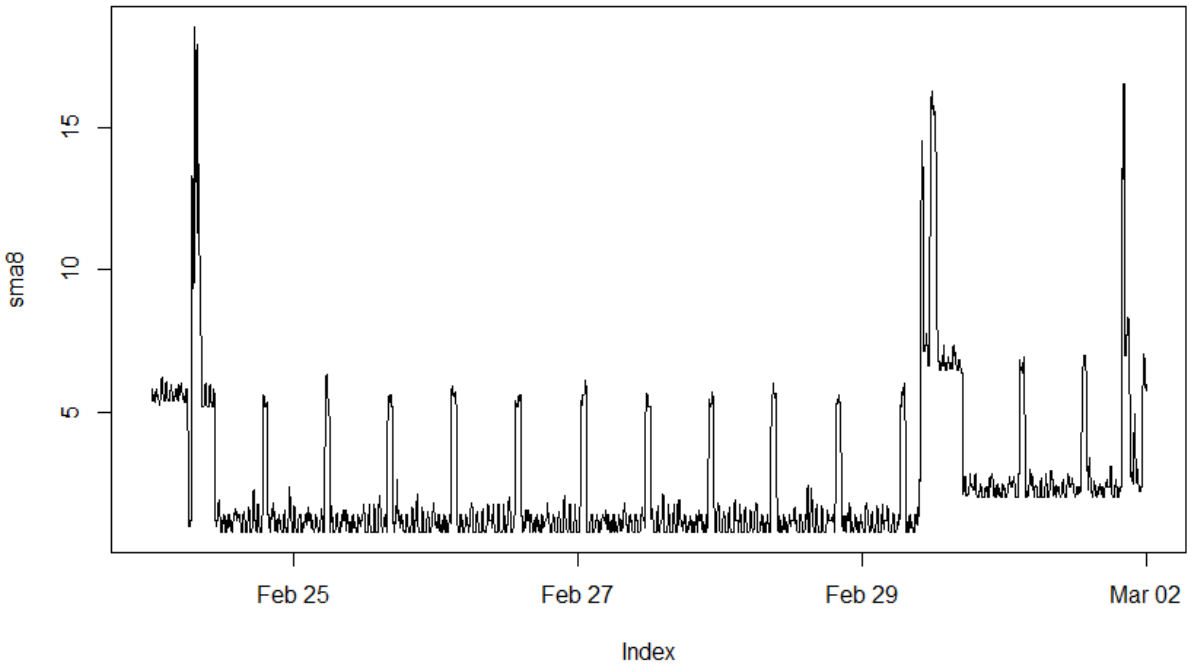
Median range [5.5,5.7]

Sd range: [4.3,4.9]

From the dataset the most anomalous weeks are the following:

The following are SMA (10 minutes):

WEEK 8:

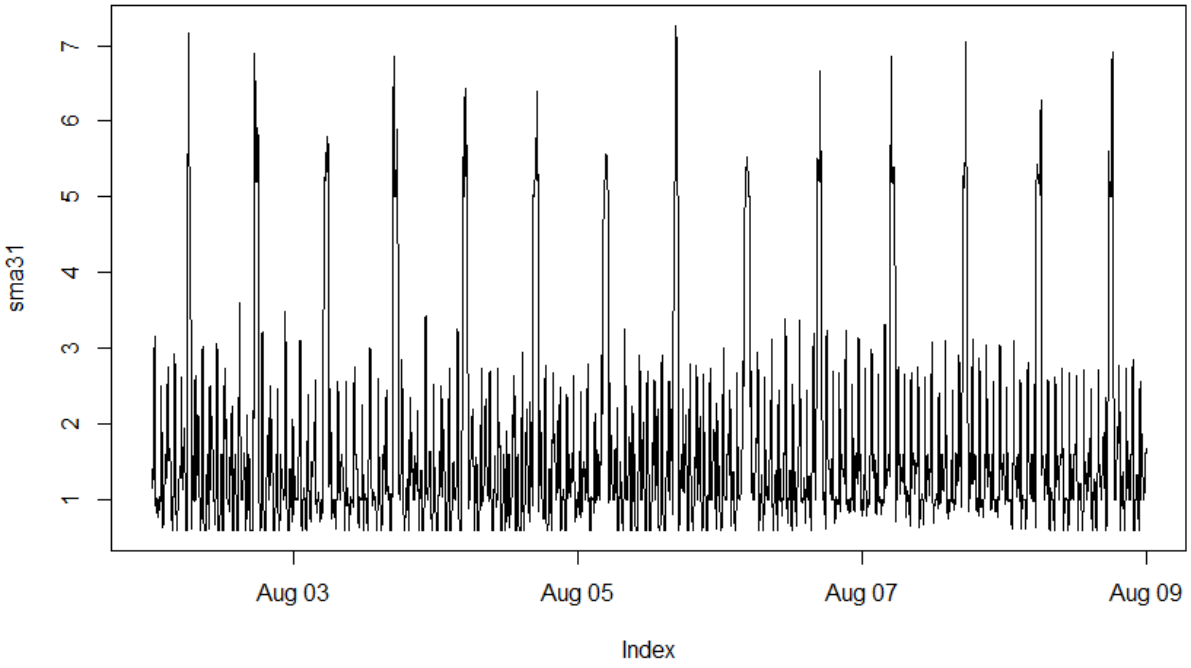


Mean: 2.39

Median: 1.38

Sd: 2.462335

WEEK 31: WORST ONE

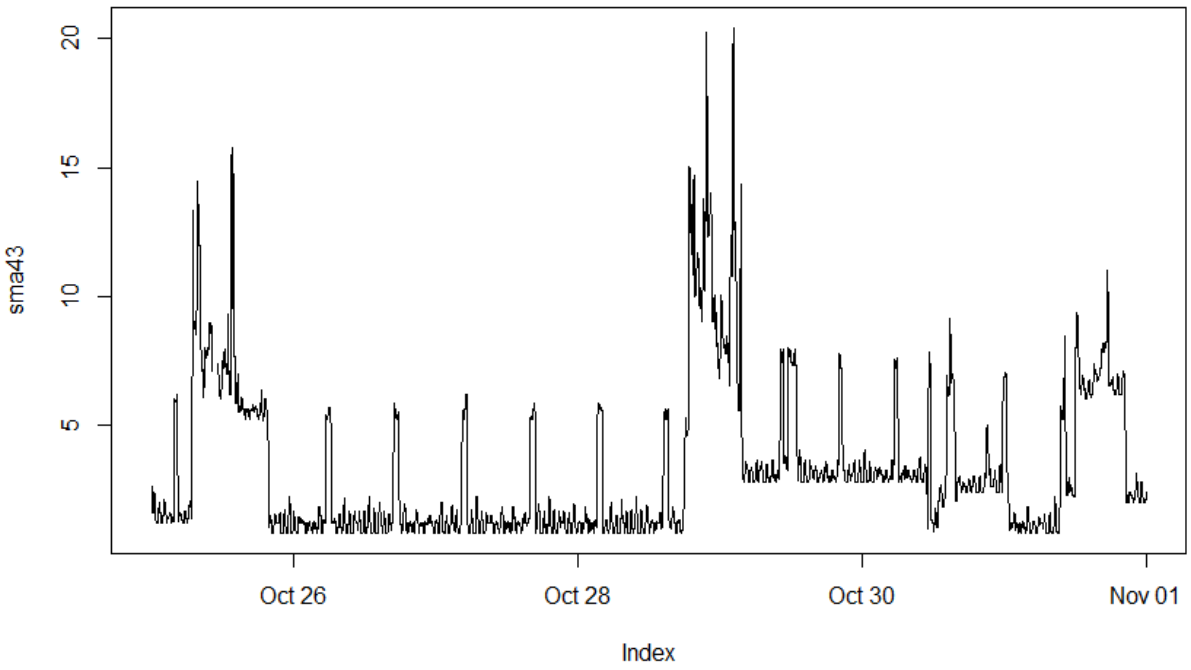


MEAN: 1.631661

MEDIAN: 1.26

SD: 1.185286

WEEK 43:

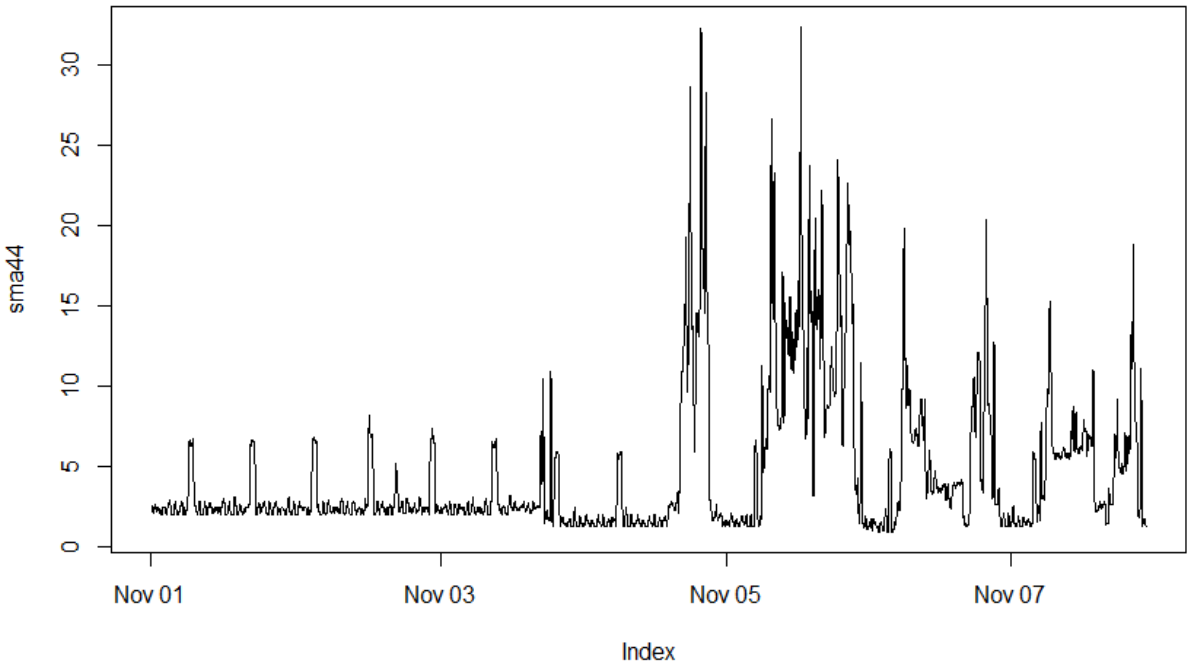


Mean:3.286136

Median: 2.26

SD: 2.931347

Week 44:



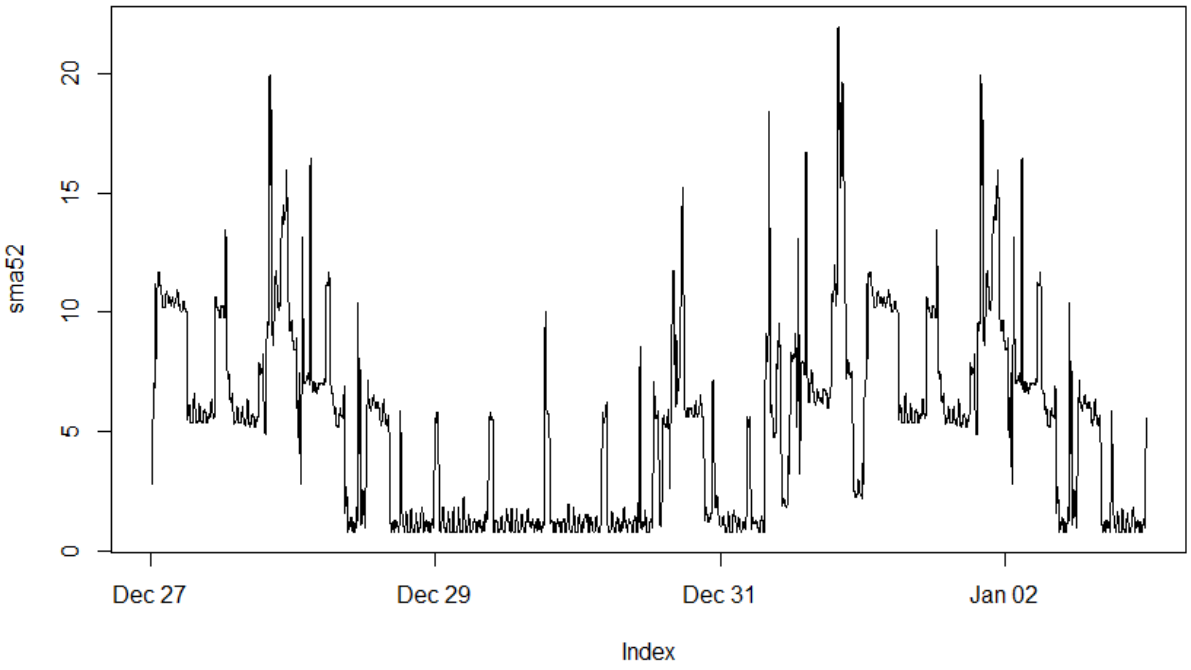
Mean: 4.506277

Median: 2.42

SD: 4.491641

Median here as you can see is around 2.42 because of the data being weirdly cut, also SD score is normal.

Week 52:



Mean: 5.196453

Median: 5.48

SD: 3.915592

As you can see at the plot is questionable for 3 days.

3.

Hidden Markov model:

What is the value of the global intensity at the certain time of the day? Does it match the model? Any anomalies we can view?

Summary:

Given here the observed states are what the value of the global intensity is in each set time of the day.

Probability pi matrix would be the probability of this specific value being close the model range value we are looking at.

The probability state is the probability that a certain value for the global intensity is within that range at a specific time of the day.

How to get the probabilities in this context.

Probability of past events:

(Global intensity scoring values, mean, median, sd, lm)

State transition probabilities:

P (global intensity values this week| past event week)

Observable states:

are the values of the global intensity given the time of the day on the model week 1-52 excluding anomalies in the dataset.