Anomaly Detection

JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY

jiit

विद्या तत्व ज्योतिसम:

Artificial Intelligence Project

Submitted to - Dr. Satish Chandra

15103297	**Jitesh Pabla**	**B8**
15103311	**Vaibhav Sharma**	**B8**
15103332	**Sajal Subodh**	**B8**

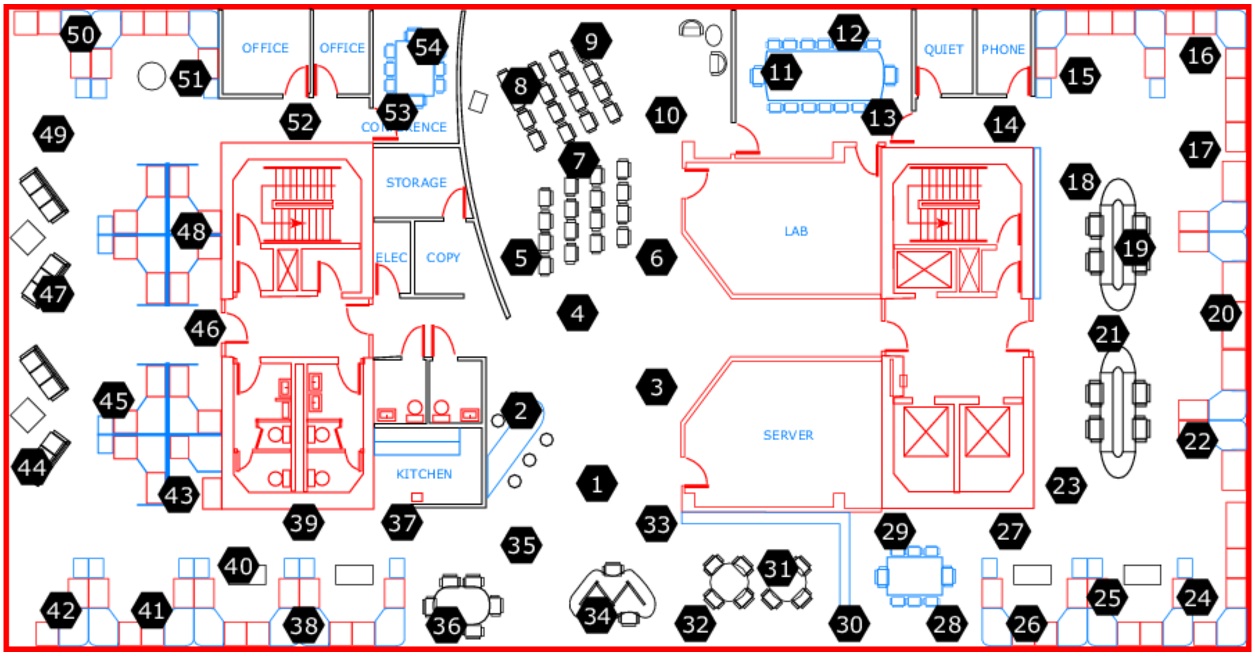
Problem Statement

The aim of this project is to detect anomalies and outliers in the Intel Lab Data.

About Intel Lab Data

Intel Lab Data contains the information about data collected from 54 sensors deployed in the Intel Berkeley Research lab between February 28th and April 5th, 2004.
Mica2Dot sensors with weather boards collected timestamped topology information, along with humidity, temperature, light and voltage values once every 31 seconds. Data was collected using the TinyDB in-network query processing system, built on the TinyOS platform.

The arrangement of the 54 sensors is as shown below :



It includes a log of about 2.3 million readings collected from these sensors. The file is 34MB gzipped, 150MB uncopressed. The schema is as follows:

date:yyyy-mm-dd	time:hh:mm:ss.xxx	epoch:int	moteid:int	temperature:real	humidity:real	light:real	voltage:real

In this case, epoch is a monotonically increasing sequence number from each mote. Two readings from the same epoch number were produced from different motes at the same time. There are some missing epochs in this data set. Moteids range from 1-54; data from some motes may be missing or truncated. Temperature is in degrees Celsius. Humidity is temperature corrected relative humidity, ranging from 0-100%. Light is in Lux (a value of 1 Lux corresponds to moonlight, 400 Lux to a bright office, and 100,000 Lux to full sunlight.) Voltage is expressed in volts, ranging from 2-3; the batteries in this case were lithium ion cells which maintain a fairly constant voltage over their lifetime; note that variations in voltage are highly correlated with temperature.

Software Requirements:

1. Python 3 and above versions
2. Packages used for data handling : NumPy, Pandas
3. Packages used for data visualisation : Matplotlib, Seaborn
4. Deep Learning Library used : Keras (Using TensorFlow Backend)
5. IPython Notebook : Jupyter Notebook

Methods Used :

We have done our analysis on temperature data of sensor with moteId 1. It involved :

1. Data - Preprocessing

2. Simple Moving Average Method

A moving average is a technique to get an overall idea of the trends in a data set; it is an average of any subset of numbers. The moving average is extremely useful for **forecasting long-term trends**. For example, if you have sales data for a twenty-year period, you can calculate a five-year moving average, a four-year moving average, a three-year moving average and so on. Stock market analysts will often use a 50 or 200 day moving average to help them see trends in the stock market and (hopefully) forecast where the stocks are headed.

Anomaly Detection using SMA :

For a given rolling window of time period, the mean and standard deviation of the entries are calculated. If the next entry in the dataset falls between the **mean ± 2 * standard deviation,** it is considered normal else Anomaly.

3. Statistical approach involving deviation from the mean

This is the most simplest way of classifying anomalies from the data points. The mean and standard deviation of each day were calculated. For each day, if any data point within that day falls outside the range of **mean ± 3 * standard deviation**, it is classified as Anomaly.

4. Density Based Spacial Clustering

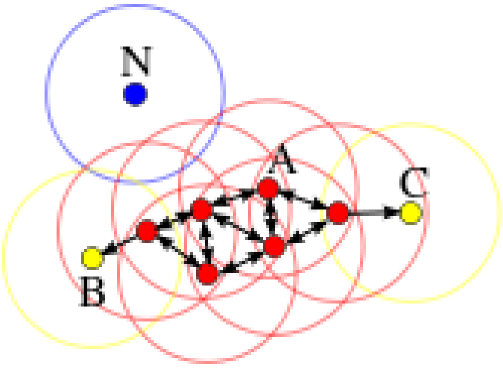
Density-based spatial clustering of applications with noise (**DBSCAN**) is a data clustering algorithm proposed by Martin Ester, Hans-Peter Kriegel, Jörg Sander and Xiaowei Xu in 1996.[1] It is a density-based clustering algorithm: given a set of points in some space, it groups together points that are closely packed together (points with many nearby neighbors), marking as outliers points that lie alone in low-density regions (whose nearest neighbors are too far away). DBSCAN is one of the most common clustering algorithms and also most cited in scientific literature.

Anomaly Detection using DBSCAN :

Consider a set of points in some space to be clustered. For the purpose of DBSCAN clustering, the points are classified as *core points*, (*density*-)*reachable points* and *outliers*, as follows:

- A point p is a core point if at least minPts points are within distance ε (ε is the maximum radius of the neighborhood from p) of it (including p). Those points are said to be *directly reachable* from p. By definition, no points are *directly reachable* from a non-core point.
- A point q is reachable from p if there is a path $p_1, ..., p_n$ with $p_1 = p$ and $p_n = q$, where each p_{i+1} is directly reachable from p_i (all the points on the path must be core points, with the possible exception of q).
- All points not reachable from any other point are outliers.

Now if p is a core point, then it forms a *cluster* together with all points (core or non-core) that are reachable from it. Each cluster contains at least one core point; non-core points can be part of a cluster, but they form its "edge", since they cannot be used to reach more points.



5. Long Short Term Memory(LSTM):

Long short-term memory (**LSTM**) block or network is a *simple* recurrent neural network which can be used as a building component or block (of hidden layers) for an eventually bigger recurrent neural network. The LSTM block is itself a recurrent network because it contains recurrent connections similar to connections in a conventional recurrent neural network.

An LSTM block is composed of four main components: a **cell**, an **input gate**, an **output gate** and a **forget gate**. The cell is responsible for "remembering" values over arbitrary time intervals; hence the word "memory" in LSTM. Each of the three *gates* can be thought as a "conventional" artificial neuron, as in a multi-layer (or feedforward) neural network: that is, they compute an activation (using an activation function) of a weighted sum. Intuitively, they can be thought as *regulators* of the flow of values that goes through the connections of the LSTM; hence the denotation "gate". There are connections between these gates and the cell. Some of the connections are recurrent, some of them are not.

The expression *long short-term* refers to the fact that LSTM is a model for the *short-term memory* which can last for a *long* period of time. There are different types of LSTMs, which differ among them in the components or connections that they have.

Anomaly Detection using LSTM :

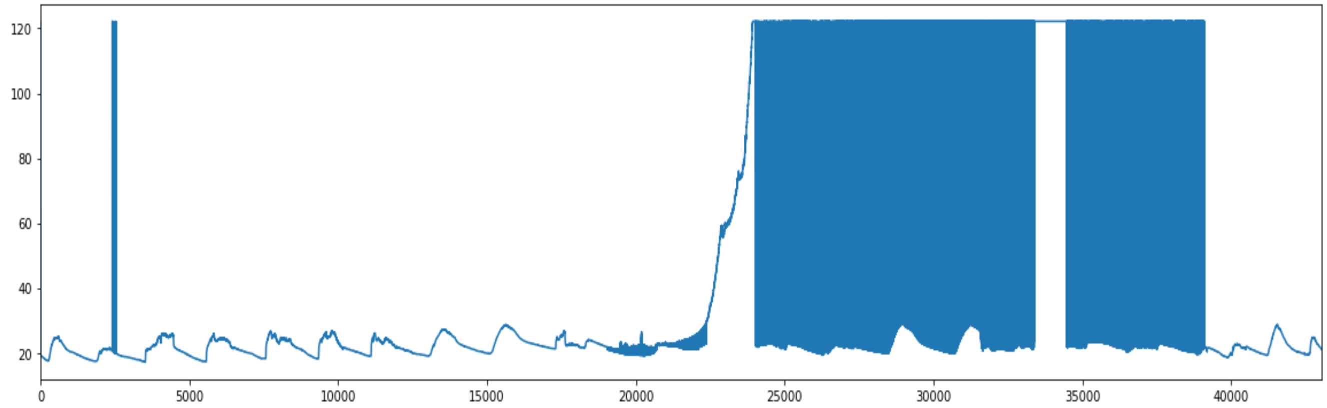
LSTMs to build a **prediction model,** i.e. given current and past values, predict next few steps in the time-series. Then, error in prediction gives an indication of anomaly. For example, if prediction error is high, then it indicates anomaly.

Another way is to directly use LSTM as a **classifier** with two classes: normal and anomalous.

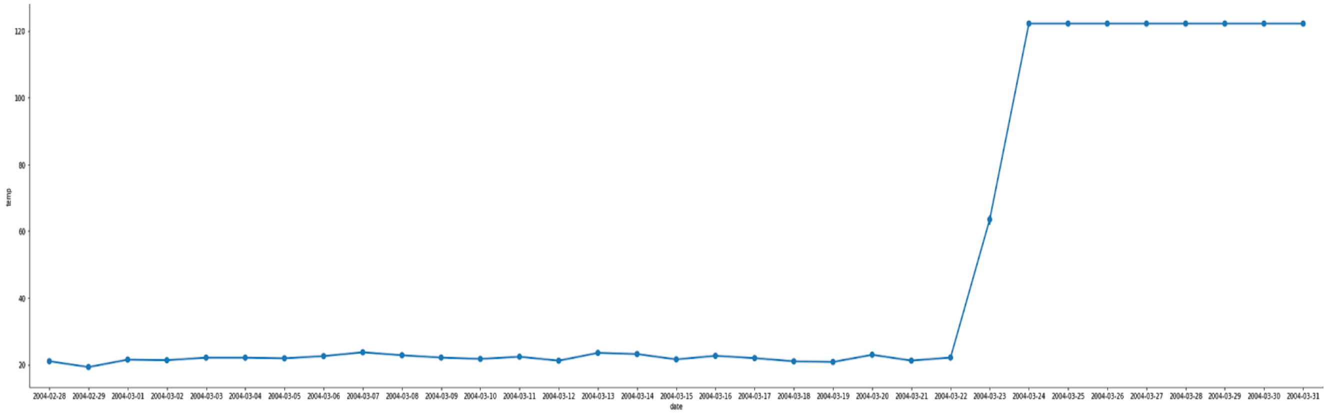
Prediction model based approach is better when anomalous instances are not easily available whereas a classifier based approach is more suitable when there are sufficient labeled instances of both normal and anomalous instances.

Analysis and Findings :

The temerature data for the moteId 1 is as :

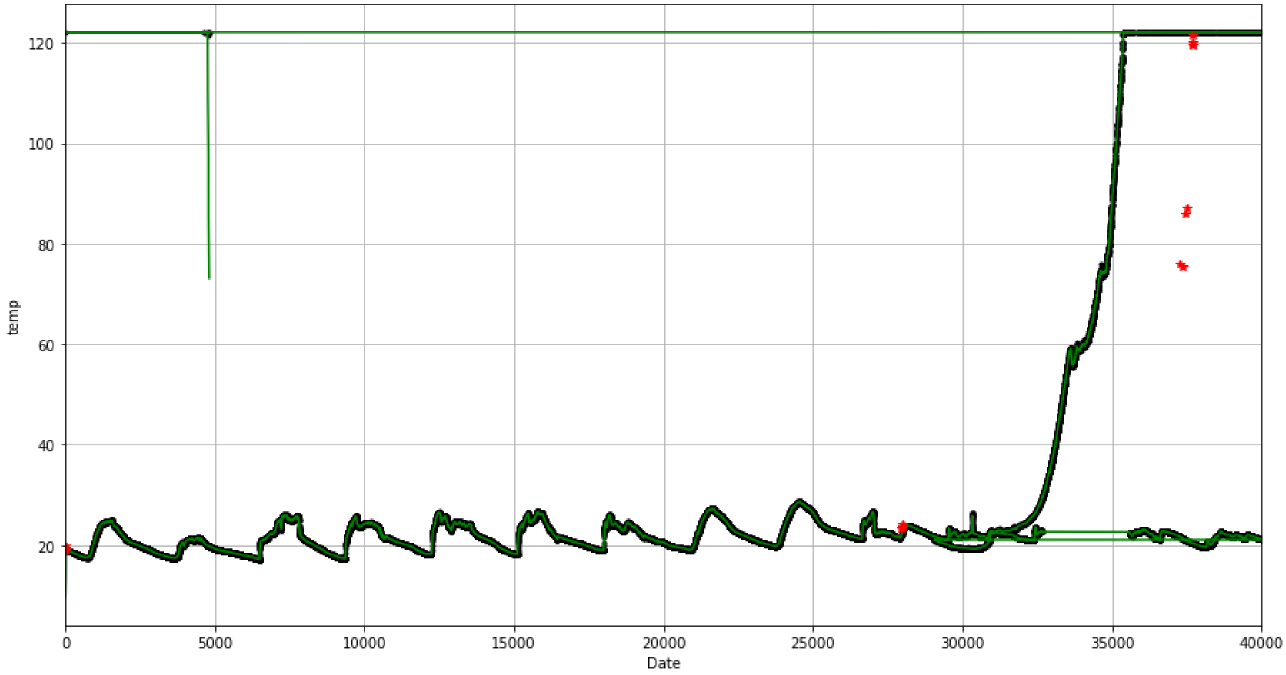


Sorting the data by date and time:



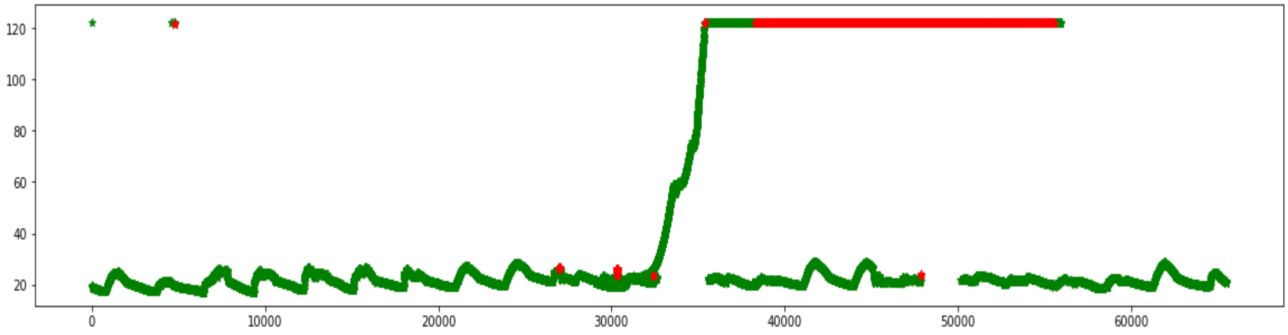
Using SIMPLE MOVING AVERAGE Method :

Window size used : 10



The Red Points are Anomalies.

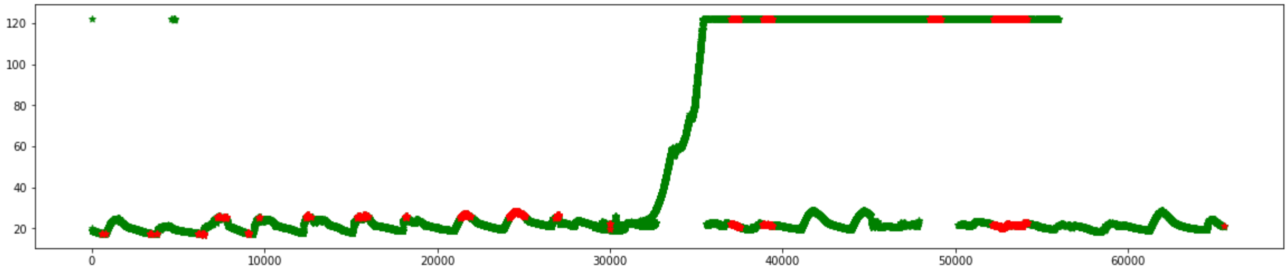
Using Statistical Way :



Total Classified Anomalies : 4808

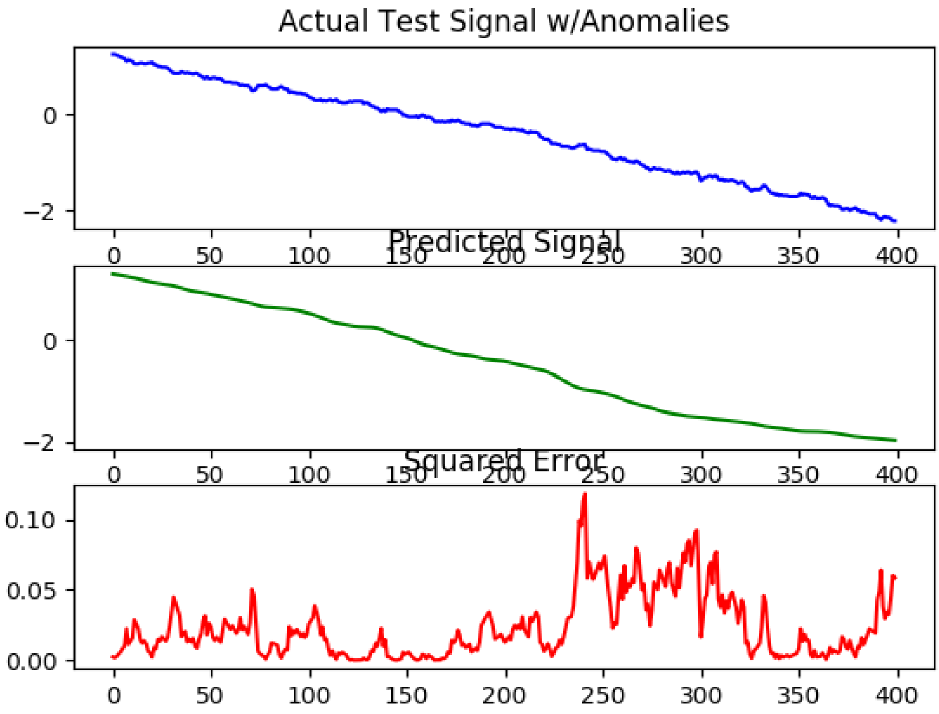
Using DBSCAN :

Each value is normalised and then fed into DBSCAN classifier with eps = 0.15 and min_sample = 450



Using LSTM :

Visible(Input) Layer : 1 node
Hidden Layer1 : 64 nodes
Hidden Layer2 : 256 nodes
Hidden Layer3 : 100 nodes
Output Layer : 1 Node


Actual Test Signal w/Anomalies

Predicted Signal

Squared Error