ISSN: 2586-761x(Print) Vol. 03, No. 01, June 2020
ISSN: 2635-5817 (Online)
**International Journal of Advanced Social Sciences**
Source: http://ictaes.org
Manuscript received: April 15, 2020; Revised: May 25, 2020; Accepted: June 5, 2020

# Weather Prediction Using Multivariable Linear Regression

Sujan Shrestha[1], Ashutosh Baral[2], Panas Tiwari[3], Shreyash Shrestha[4], Surendra Shrestha[*5]

[1,2,3,4]*Department of Electronics, Communication and Information Engineering, Kathmandu Engineering College, Tribhuvan University, Nepal, 44614,*
[5]*Depatment of Electronics and Computer Engineering, Pulchwok Campus, Institute of Engineering, Tribhuvan University, Nepal, 44700*

*sujan.shrestha@keckist.edu.np[1], baralashutosh0@gmail.com[2], panas.tiwary18@gmail.com[3], Shreyashshrestha22@gmail.comcom[4], surendra@ioe.edu.np[5]*

***Abstract***

*This paper is carried on heuristic prediction of temperature using multivariable linear regression. Vast research in the field of machine learning and artificial intelligence has given rise to enormous efficiency in learning algorithms, hence in this paper we explore the application for weather forecasting. The scope of this paper is restricted to predict maximum and minimum temperature of the next day, given weather data of the past five days. Multivariable linear regression algorithm is used to find the pattern between future and past temperature conditions..*

***Keywords:*** *machine learning, artificial intelligence, multivariable linear regression, weather prediction*

## 1. Introduction

Weather forecasting is the task of predicting the state of the atmosphere at a future time and a specified location. Traditionally, this has been done through physical simulations in which the atmosphere is modeled as a fluid. The present state of the atmosphere is sampled, and the future state is computed by numerically solving the equations of fluid dynamics and thermodynamics. However, the system of ordinary differential equations that govern this physical model is unstable under perturbations, and uncertainties in the initial measurements of the atmospheric conditions and an incomplete understanding of complex atmospheric processes restrict the extent of accurate weather forecasting to a 10 days period, beyond which weather forecasts are significantly unreliable. Machine

Corresponding Author : Surendra Shrestha
Author's affiliation : Dept. of Electronics and Computer Engineering, Pulchwok Campus, IoE, TU, Nepal
Email: surendra@ioe.edu.np
Copyright © ICT-AES

learning, on the contrary, is relatively robust to perturbations and doesn't require a complete understanding of the physical processes that govern the atmosphere. Therefore, machine learning may represent a viable alternative to physical models in weather forecasting [1].

Machine learning is the ability of a computer to learn without being explicitly programmed. It allows machines to find hidden patterns and insights. In this process, a model based on labeled training data can be built. The model is then used for mapping new examples. So, based on the observed weather patterns from the past, a model can be built and used to predict the weather [2].

Multivariable linear regression algorithm is used to create two models for both minimum and maximum temperature. This model takes the minimum and maximum temperature of the past five days and predict the minimum and maximum temperature of the next day.

## 2. Methodology

Methodology consists of collection of data, filtering of data, selection and processing of data, creating models, predicting the result. The program of this paper is written in Python programming language. After completing all these processes, we will be able to create a simple machine learning model that predicts future temperature with the help of past data.

The first step of any machine learning is the collection of the data. As we use past data to predict the future temperature the quality is one of the most important things in this process. In our case we used the dataset collected from the office of the Department of Hydrology and Meteorology, Kathmandu [3].

**Table 1:** Sample collected data

| | gsid | index_no | station | district | year | month | day | max_temp | min_temp |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 366 | 1030 | Kathmandu Airport | Kathmandu | 2009 | 1 | 1 | 19.8 | 4.8 |
| **1** | 366 | 1030 | Kathmandu Airport | Kathmandu | 2009 | 1 | 2 | 19.2 | 5.8 |
| **2** | 366 | 1030 | Kathmandu Airport | Kathmandu | 2009 | 1 | 3 | 20.6 | 6.8 |
| **3** | 366 | 1030 | Kathmandu Airport | Kathmandu | 2009 | 1 | 4 | 20.5 | 7.8 |
| **4** | 366 | 1030 | Kathmandu Airport | Kathmandu | 2009 | 1 | 5 | 18.8 | 8.8 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **4012** | 366 | 1030 | Kathmandu Airport | Kathmandu | 2019 | 12 | 27 | 15.7 | 1.2 |
| **4013** | 366 | 1030 | Kathmandu Airport | Kathmandu | 2019 | 12 | 28 | 16.3 | 0.6 |
| **4014** | 366 | 1030 | Kathmandu Airport | Kathmandu | 2019 | 12 | 29 | 18.1 | 1.0 |
| **4015** | 366 | 1030 | Kathmandu Airport | Kathmandu | 2019 | 12 | 30 | 18.0 | 1.2 |
| **4016** | 366 | 1030 | Kathmandu Airport | Kathmandu | 2019 | 12 | 31 | 17.4 | 1.5 |

Data contains minimum and maximum temperature from 2009 to 2019 from the station Kathmandu Airport. The originally collected data contains the base information (year, month, day, max_temp, min_temp) that we require. These data can't simply be fit to the machine learning model because we might not have the full information that

we require for the model. We also have to filter for any kind of null values in the dataset.

We use the past five days temperature as input so we also need to create new columns in the dataset for input and output values. Table 3 shows the final data after the filtration process. Here "max_1" and "min_1" refers to the maximum and minimum temperature of the past one day respectively. These past temperatures are used as an input to the model. "max_temp" and "min_temp" refers to today's maximum and minimum temperature respectively that is to be predicted. The model we use is a supervised model. So, we need to train it with the help of a labeled dataset.

**Table 2:** Sample dataset after filtration

| | year | month | day | max_temp | min_temp | max_1 | max_2 | max_3 | max_4 | max_5 | min_1 | min_2 | min_3 | min_4 | min_5 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 2009 | 1 | 7 | 21.0 | 10.8 | 20.6 | 18.8 | 20.5 | 20.6 | 19.2 | 9.8 | 8.8 | 7.8 | 6.8 | 5.8 |
| **1** | 2009 | 1 | 8 | 20.4 | 11.8 | 21.0 | 20.6 | 18.8 | 20.5 | 20.6 | 10.8 | 9.8 | 8.8 | 7.8 | 6.8 |
| **2** | 2009 | 1 | 9 | 22.2 | 12.8 | 20.4 | 21.0 | 20.6 | 18.8 | 20.5 | 11.8 | 10.8 | 9.8 | 8.8 | 7.8 |
| **3** | 2009 | 1 | 10 | 21.5 | 13.8 | 22.2 | 20.4 | 21.0 | 20.6 | 18.8 | 12.8 | 11.8 | 10.8 | 9.8 | 8.8 |
| **4** | 2009 | 1 | 11 | 22.3 | 14.8 | 21.5 | 22.2 | 20.4 | 21.0 | 20.6 | 13.8 | 12.8 | 11.8 | 10.8 | 9.8 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **4006** | 2019 | 12 | 27 | 15.7 | 1.2 | 15.8 | 16.7 | 17.5 | 18.8 | 20.3 | 0.9 | 2.9 | 3.2 | 3.0 | 4.2 |
| **4007** | 2019 | 12 | 28 | 16.3 | 0.6 | 15.7 | 15.8 | 16.7 | 17.5 | 18.8 | 1.2 | 0.9 | 2.9 | 3.2 | 3.0 |
| **4008** | 2019 | 12 | 29 | 18.1 | 1.0 | 16.3 | 15.7 | 15.8 | 16.7 | 17.5 | 0.6 | 1.2 | 0.9 | 2.9 | 3.2 |
| **4009** | 2019 | 12 | 30 | 18.0 | 1.2 | 18.1 | 16.3 | 15.7 | 15.8 | 16.7 | 1.0 | 0.6 | 1.2 | 0.9 | 2.9 |
| **4010** | 2019 | 12 | 31 | 17.4 | 1.5 | 18.0 | 18.1 | 16.3 | 15.7 | 15.8 | 1.2 | 1.0 | 0.6 | 1.2 | 0.9 |

A key assumption required by the linear regression technique is that you have a linear relationship between the dependent variable and each independent variable. One way to assess the linearity between independent variables, which for now will be the maximum and minimum temperature and the other independent variables is to calculate the Pearson correlation coefficient.

The Pearson correlation coefficient is a measurement of the amount of linear correlation between equal length arrays which outputs a value ranging -1 to 1. Correlation values ranging from 0 to 1 represent increasingly strong positive correlation, whereas the correlation values from 0 to -1 are said to be inversely, or negatively, correlated in that when the values of one series increase the corresponding values in the opposite series decrease but, as changes in magnitude between the series become equal (with opposite direction) the correlation value will approach -1 [4].

So, the value of Pearson correlation coefficient between independent and dependent variables i.e., input and output variables of the model shown in the Table 3.

**Table 3:** Pearson correlation coefficient between variables

| | max_temp | min_temp |
|---|---|---|
| day | -0.020711 | 0.002424 |
| year | -0.096786 | 0.008871 |
| month | -0.031606 | 0.169745 |
| max_temp | 1.000000 | 0.777798 |
| min_temp | 0.777798 | 1.000000 |
| max_1 | 0.903494 | 0.795043 |
| max_2 | 0.858019 | 0.798739 |
| max_3 | 0.839023 | 0.799135 |
| max_4 | 0.820886 | 0.795439 |
| max_5 | 0.808487 | 0.793486 |
| min_1 | 0.774965 | 0.974303 |
| min_2 | 0.773575 | 0.960598 |
| min_3 | 0.768552 | 0.951772 |
| min_4 | 0.763908 | 0.945742 |
| min_5 | 0.755552 | 0.940240 |

So, as input to our model the variables that have correlation coefficient more than 0.7 is used.

## 3. Linear regression algorithm

Linear Regression model predicts value based on independent variables. Linear regression refers to the mathematical technique of fitting given data to a function of a certain type. It is best known for fitting straight lines [5]. In this paper, we explain the theory behind linear regression and illustrate this technique with real world data set. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used. Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.
The linear regression hypothesis is shown below in equation 1.

$$h_\theta(x) = \theta_0 \times x_0 + \theta_1 \times x_1 + \ldots + \theta_n \times x_n \quad (1)$$

where:

$h_\theta(x)$ = output of model when x in the input for the fixed value of θ

x = independent variable

$\theta_0$ = bias

$\theta_{1\text{-}n}$ = weight

n = number of independent variables

Least-square or the cost equation has been used in the algorithm of linear-regression method. The least-squares

cost equation is shown in equation 2:

$$J_\theta(x) = \frac{1}{2} \times n \times [\sum (h_\theta(x) - y)^2] \quad (2)$$

where:

$J_\theta(x)$ = least-square error (cost) when x is input for fixed value of θ

y = labeled output of training dataset

Gradient descent algorithm is used for optimizing the cost function in this method. The gradient descent equation is shown in equation 3 and equation 4:

$$\theta_0 = \theta_0 - \alpha \times (1/m) \times [\sum (h_\theta(x) - y)] \quad (3)$$

$$\theta_{i-n} = \theta_{i-n} - \alpha \times (1/m) \times [\sum (h_\theta(x) - y)] \times x_{i-n} \quad (4)$$

where:

m = total number of rows in training dataset

α = learning rate

## 4. Result and Discussions

Two models were created using multivariable linear regression to predict both the maximum and minimum temperature of today using the temperature of the past five days. In our process we split the dataset in two parts. One for the training set and another for the testing set. 80% of total data is used for training the model and remaining 20% is used for testing and validation of the model.

The result of linear regression model for predicting maximum temperature is shown in Figure 1.

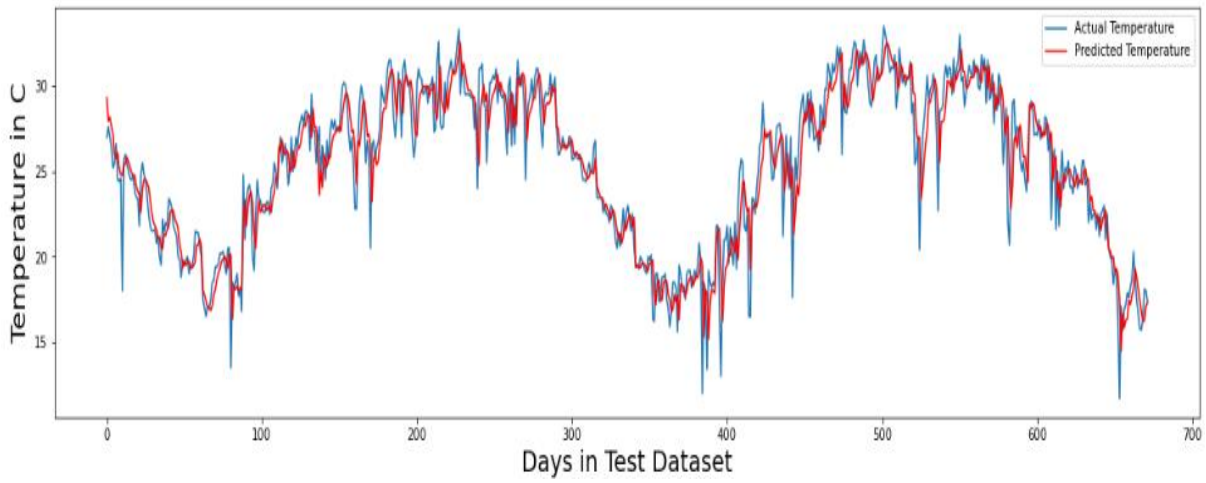



**Figure 1**. Actual maximum temperature, predicted maximum temperature for test data.

The root mean square error for training and testing dataset are:
For training dataset: 1.85 ºC, For testing dataset: 1.67 ºC.

The cost of the hypothesis equation gets conversing after every iteration during training of the dataset. The iteration vs cost function during the training process of the maximum temperature model is shown in Figure 2.

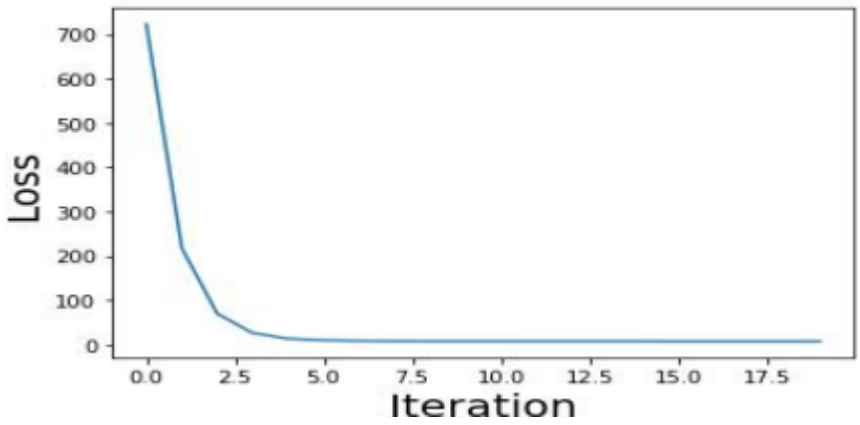



**Figure 2.** Iteration vs Loss (Maximum temperature).

From the Figure 2, the loss is conversing with iteration and that the model is actually learning from the training dataset. The reason to perform gradient descent algorithm is to find out the best value of bias and weight such that the hypothesis equation predicts the best result for the given input values. Hence the value of weight and bias for the model that predicts maximum temperature is:

w = [ 0.60454594, 0.097754, 0.11203822, 0.06905639, 0.10877939, -0.028793, 0.04597889, 0.00543061, 0.03686152, -0.05159367]

b = 0.05502594864669722

The result of linear regression model for predicting minimum temperature are shown in Figure 3:

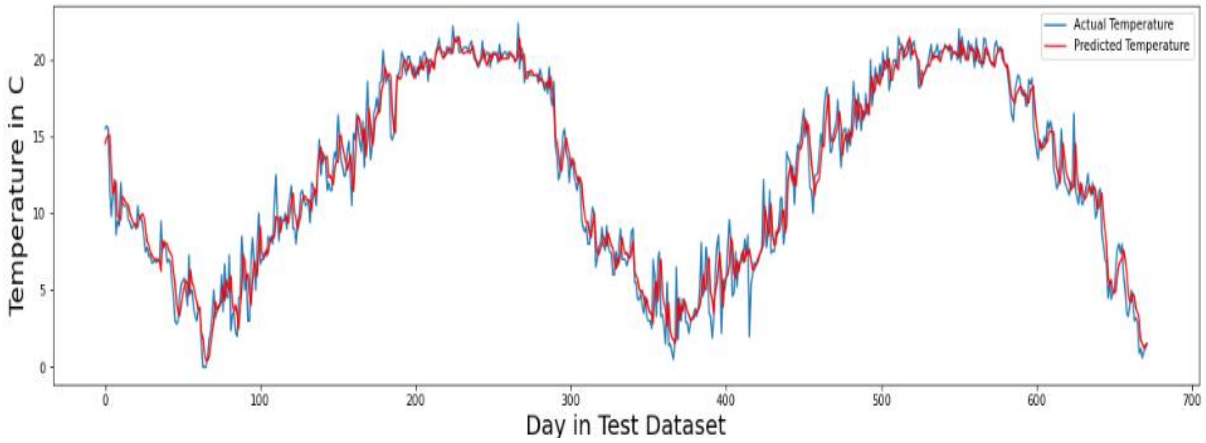



**Figure 3.** Actual maximum temperature, predicted maximum temperature for test data

The predicted temperature is relating to the graph of the actual temperature in the training data set.

The root mean square error for training and testing dataset are:

For training dataset: 1.86 ºC, for testing dataset: 1.30 ºC.

The relation between iteration vs loss of the model prediction minimum temperature is given in Figure 4.

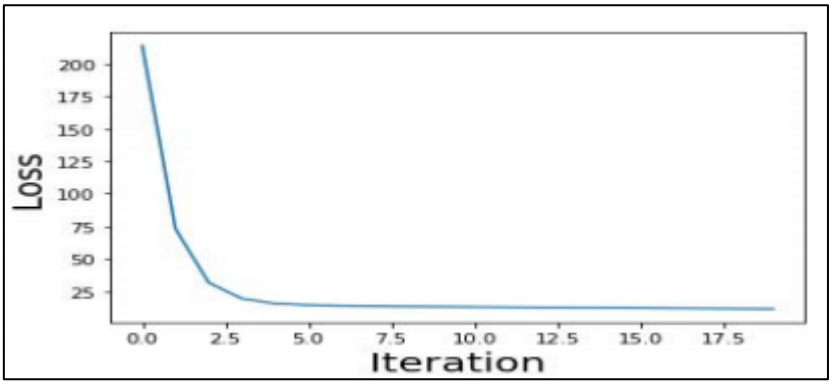



**Figure 4.** Iteration vs Loss (Minimum temperature)

From the Figure 4, the loss is conversing with iteration and that the model is actually learning from the training dataset.

Finally, the value of weight and bias of this model are:

Weight = [0.0909464250, -0.000611824493, -0.00382323843, -0.00374283709, -0.0320568418, 0.653501724, 0.149437993, 0.386362461, 0.0611421409,0.0686111882]

bias = -0.0483537050

## 5. Conclusion

From the results obtained, we can conclude that linear regression algorithm can predict maximum and minimum temperature of the next day with the root mean square error of 1.6 ºC and 1.3 ºC respectively. We used a gradient descent algorithm that calculates the best value of weight and bias required for the model. After 10 epochs, we didn't see significant change in loss so we can stop as soon as we reach stable state. We can also increase the accuracy of the model by using a dataset that contains more parameters like pressure, humidity, wind speed, wind direction and many more.

## References

[1] M. Holmstrom, D. Liu, C. Vo "Machine Learning Applied to Weather Forecasting", Stanford University (2016).

[2] S. Gupta, Indumathy, G. Singhal, "Weather Prediction Using Normal Equation Method and Linear regression Techniques", International Journal of Computer Science and Information Technologies (2016).

[3] Department of Hydrology and Meteorology, Kathmandu, Nepal. http://dhm.gov.np/

[4] A. McQuistan "Using Machine Learning To Predict Weather"https://stackabuse.com/using-machine-learning-to-predict-the-weather-part-1/, https://stackabuse.com/using-machine-learning-to-predict-the-weather-part-2/ (2016).

[5] M. Hyang "Theory of Implementation of linear regression" International Conference on Computer Vision, Image and Deep Learning (CVIDL), 30 Nov, 2020.