ISSN:0975-9646

Modelling of Weather Data Using Gaussian Linear Regression

Vamsi Krishna G

Computer Science and Engineering,
GITAM University, Visakhapatnam, India

***Abstract*— In this paper, a model is proposed using Gaussian regression for forecasting the weather conditions. The main advantage of this model is that it can be understood very easily to the common farmer. The weather parameters that are considered for predicting analysis include precipitation, Wind speed, mean average temperature. These features play a vital role as they are the key indicators for identifying the outcomes. The developed model initially considered as simple regression model which are then portrayed onto a Gaussian regression model. The developed model helps to predict the vital parameters that decide possible weather changes.**

***Keywords*— Weather forecasting, multiple regression, linear regression, Gaussian regression, Time Series.**

I INTRODUCTION

Weather forecasting plays a significant role in the human lives. Weather prediction also has an equivalent role since, the weather prediction attributes can be very much useful for agriculturists and weather dependent industries. Weather has a significant role in sectors varying from defense, shipping, aerospace and navigation etc. These feature predictions can be used to warn the public about the natural disasters and the consequent conditions well ahead to the public so that necessary safety mechanisms can be considered.

Weather forecasting is generally carried out in a 2-phased manner i.e., by considering the weather data through images acquired from the satellites and by the prediction of variables along with the wind speed, precipitation, humidity, temperature and other several factors which decide the abnormal weather changes. Many models have been showcased in the literature of which majority are inclined towards the usage of time-series data [1] [2] [3], the time series models also help to understand relevant feedback from the stock marketing sectors and agricultural sectors. Many models have been utilized for weather prediction of which some include linear regression [4] [5] [6], time series vector models [7] [8] [9], artificial neural networks and based on data mining techniques. However, most of the information pertaining to the weather data exhibit nonlinear hidden patterns having a huge dimensionality space. It is not flexible to model the data using regression functions [10] [11], therefore parametric models are focused in the literature. However, these models also witnessed the disadvantages because of the rigid nature of atmospheric pressure structures, and hence could not be effective in modeling the weather data. To overcome these disadvantages, it is necessary to develop procedures based on Non-Parametric Regression Models, such as Gaussian Regression Models. These models help to identify non-linear patterns more effectively and can manage the huge dimensional data. Therefore in this paper an attempt is made towards this direction by using statistical models.

Statistical methods are most vigorous for identifying the atmospheric patterns since the atmospheric patterns exhibit a non-linear dynamic system and hence cannot be predicted efficiently by using deterministic models.

Regression models help to find the relationships between variable i.e., between predictor variable and other variables needed for identification of changes in the weather. Conventional methods assume these errors as independent random variables having zero mean and also considered to be variance, hence, with these assumptions most of the works in this area are considered by using Gaussian mixture distributions. In this paper, the Gaussian Regression model is utilized by considering the advantages of Gaussian Mixture models.

The rest of the paper is organized as follows section-2 highlights about the Gaussian regression procedure, in section-3 the data set considered is discussed and section-4 of the paper deals with brief introduction about linear regression. Section-5 elucidates methodology together with results derived and concludes section-6 summarizes the paper.

II GAUSSIAN REGRESSION PROCEDURE

The Gaussian regression process is given by using exponential square model given by

$$k(x,x') = \delta_f^2 \exp\left[\frac{-(x-x')^2}{2l^2}\right] \quad \text{---(1)}$$

Where $k(x,x')$ denotes the covariance

δ denotes the allowable covariance

l^2 denotes the separations parameter which decides the deviations between x and x'

The main advantage of Gaussian Regression Models is that Non parametric methodologies are more suitable for estimation and regression.

It explains the relationship between are variable and other and helps to identify the changes.

III DATA SET CONSIDERED

The data set considered for this work is collected from Indian meteorology department. The data set is considered for a period of 24 months from January 2012 to December 2014. The data considered before applying to the model need to be preprocessed for which the following procedures are adapted.

Remove the missing data, duplicate data and covert the data into a form suitable for analysis and this process is called data cleaning.

The data set considered has 10 attributes and the type of the data along with the descriptive is presented in the following table.

Attribute	Type	Description
Year	numeric	Year is considered
Month	numeric	Month is considered
Wind speed	numeric	Kilometers
Radiation	numeric	Amount of radiation
Evaporation	numeric	Amount of evaporation
Minimum temperature	numeric	Monthly minimum temperature
Maximum temperature	numeric	Monthly maximum temperature
Rainfall	numeric	Aggregate of rainfall during month
Sunshine	numeric	Amount of sunshine
Cloud form	numeric	Amount of mean temperature

Data transformation:

The data that is considered is converted into comma separated value for processing.

IV LINEAR REGRESSION

The main advantage of considering the prediction models are that they help in identifying the hidden features and enhance the capability of adapting to situations. In this paper we have considered a Gaussian regression model which mainly focused on the preparation of covariance matrices given by

$$k=\begin{bmatrix} k(x_1,x_2) & k(x_1,x_2) & ..... & k(x_1,x_n) \\ k(x_2,x_1) & k(x_2,x_2) & ..... & k(x_2,x_n) \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ k(x_n,x_1) & k(x_n,x_2) & ..... & k(x_n,x_n) \end{bmatrix}$$

Where $k(x_1,x_2)$ the covariance among the points. These covariances in our particular case have been obtained from linear regression model with together time series models. These data is given as input to Gaussian regression procedure specified in section-2 of the paper. For each of these data sets the procedure is repeated and initial estimates are identifies from these estimates the best variable are determined for which the equation considered includes

$$\sum x_1 y = \sum x_1^2 b_0 + \sum x_1 x_2 b_1$$
$$\sum x_2 y = \sum x_1 x_2 b_0 + \sum x_2^2 b_1$$
$$\sum x_3 y = \sum x_3 x_1 b_0 + \sum x_3 x_2 b_1 + \sum x_3^2 b_2$$

TABLE 1

S.No	y_1	Min x_1	Max x_2	Mean x_3	x_1x_2	x_1x_3	x_2x_3	x_1^2	x_2^2	x_1^2
1	Wind speed	79.33	188.78	134.913	14975.9	14975.9	25468.7	6293.24	35637.9	18201.5
2	Evaporation	1.7	10.9	4.128	18.53	7.01	44.99	2.89	118.81	17.04
3	Cloud Form	7	7	7	49	49	49	49	49	49
4	Radiation	7.6	43.08	13.081	327.408	99.41	563.52	57.76	1855.88	171.11
5	Sunshine	1.5	7.9	5.07	11.85	7.6	40.05	2.25	62.41	25.7
6	Min Temp	21.1	30.9	23.157	651.99	488.61	715.55	445.21	954.81	536.24
7	Max Temp	26.8	38.4	31.93	1029.12	855.72	1226.11	718.24	1474.56	1019.52
8	Rainfall	0	373.4	120.7	0	0	45069.4	0	139428	14568.5
9	Year	2000	2009	0	4018000	0	0	4000000	4036081	0
10	Month	1	12	0	12	0	0	1	144	0

yw=79.33min+188.73max+134.913mean
ye=1.7min+10.9max+4.128mean
ycf=7min+7max+7mean
yr=7.6min+43.08max+13.081
ys=1.5min+7.9max+5.07mean
ymin=21.1min+30.9max+23.157mean
ymax=26.8min+38.4max+31.93mean
yr=0min+373.4max+120.7mean
yy=2000min+2009max+0mean
ym=1min+12max+0mean

From the time series analysis it can be identified that the values minimum, maximum, mean are estimated to be 0.239, -0.274, 0.623 respectively using which the above equations are estimated as

yw=79.33(0.239)+188.73(-0.274)+134.913(0.623)
=18.95-51.71+84.05
=51.29

Similarly the values are calculated for all the nine equations and the values obtained are as follows

Ye=-0.01, Ycf=4.116, Yr=-1.85, Ys=1.34, Ymin=11, Ymax=15.77,Yr=-27.12, Yy=-72.4, Ym=-3.049

From the table it could be absorbed that the impact of wind pressure on the other pattern is high which signifies that for effective prediction wind speed should be taken into consideration. The above data is given as input to the Gaussian regression model given by

$$k(x,x') = \delta_f^2 \exp\left[\frac{-(x-x')^2}{2l^2}\right]$$

Where l is the length parameter.

x, x' are the minimum and maximum allowable coefficients.

The data obtained from the above least square analysis is given as input to Gaussian regression process to assess the impact

$$\begin{bmatrix} 51.29 & -0.01 & 4.116 \\ -1.85 & 1.34 & 11 \\ 15.77 & -27.12 & -72.4 \end{bmatrix}$$

This model helps to predict the impact of weather given the other factors.

TABLE 2

S.No	x1	x2	x3	$\left(\frac{x_1-\mu_1}{n}\right)^2$	$\left(\frac{x_2-\mu_1}{n}\right)^2$	$\left(\frac{x_3-\mu_1}{n}\right)^2$
1	79.33	188.78	134.913	1245.38	155.5	116.4
2	1.7	10.9	4.128	711.28	1038.77	13.91
3	7	7	7	742.56	86459.52	11.62
4	7.6	43.08	13.081	746.38	821.39	7.50
5	1.5	7.9	5.07	708.62	1060.80	13.17
6	21.1	30.9	23.157	830.59	900.60	2.62
7	26.8	38.4	31.93	867.30	851.47	0.40
8	0	373.4	120.7	701.19	64.64	84.82
9	2000	2009	0	61851.69	3598609	17.55

Evaluation Metrics:

To evaluate the model we have considered various metrics like correlation coefficient and mean square error. The correlation coefficient is considered to be best when it tends towards 1(one). The mean square error is the sum of difference between each value and its corresponding computed value. The results derived by using the above metrics for the data set considered are

TABLE 3

Performance Measure	Training Data Result	Test Data Result
MSE	0.5326	0.7210
RMSE	-0.732	-0.345

V METHODOLOGY

The data acquired is given as input to the Linear Regression Model, presented in section 4 of the paper. The modeled data thereof is given as input to the Gaussian Regression process, as presented in Table-2.From the above table-2, it could be understood that the sum of deviations, about the mean are more for the data value 79.3, where the precipitation, wind speed and mean temperature are more. This data can be used for further usage, where in case of prediction, the values of the attributes can be given as input to the Gaussian Regression process and the outputs derived can be used to derive possible conclusions.

VI CONCLUSION

In this paper the methodology is presented for effective prediction of the weather conditions. This model uses the initial results using linear regression and the results derived thereof are given as input to the Gaussian regression model. From the results derived it can be understood that the effect of wind and precipitation together with temperature helps to predict the possibility of weather changes. This model can be used for feature identification of the weather based on the three input parameters wind speed, precipitation and mean temperature. It could be also identified that the sum of deviation about the mean is high when the parameters are high which signifies the changes in weather.

REFERENCES

1. Agarwal, B.L., 1991, Basic Statistics, Second Edition, New Delhi, Wiley Eastern Ltd.
2. Chatfield, C., 1994, The Analysis of Time Series- An Introduction, Fourth Edition, London, Chapman and Hall.
3. Holmström, L., Koistinen, P., Laaksonen, J. and Oja, E., Neural and Statistical Classifiers: Taxonomy and Two Case Studies, Jan.1997, IEEE Trans. Neural Networks, vol. 8, No. 8, pp 5-15.
4. Montgomery, D. C. and Lynwood, A. J., 1996, Forecasting and Time Series Analysis, New York, McGraw-Hill.
5. Mathur, S., Dec. 1998, Stock Market Forecasting Using Neural Networks-An MBA Project Report Submitted to School of Management Studies, IGNOU, New Delhi.

6. Mathur, S., Shukla, A.K. and Pant, R.P., Agra, Sept. 22-23, 2001, International Conference on Optimization Techniques and its Applications in Engineering and Technology, A Comparative Study of Neural Network and Regression Models for Estimating Stock Prices.
7. Roadknight, C.M., Balls, G.R., Mills, G.E. and Palmer-Brown, D., Modeling Complex Environmental Data, July 1997, IEEE Trans. Neural Networks, vol. 8, No. 4, pp852-861.
8. Shukla, M.C., and Gulshan, S.S., 1980, Statistics, Fourth Edition, New Delhi, S.Chand and Company.
9. MacKay, D. (1998). In C.M. Bishop (Ed.), Neural networks and machine learning. (NATO ASI Series, Series F, Computer and Systems Sciences, Vol. 168, pp. 133166.) Dordrecht: Kluwer Academic Press.
10. Rasmussen, C. and C. Williams (2006). Gaussian Processes for Machine Learning. MIT Press.
11. Sivia, D. and J. Skilling (2006). Data Analysis: A Bayesian Tutorial (second ed.).Oxford Science Publications.