Multivariate Polynomial Regression in Data Mining: Methodology, Problems and Solutions

Priyanka Sinha

Abstract— Data Mining is the process of extracting some unknown useful information from a given set of data. There are two forms of data mining – predictive data mining, descriptive data mining. Predictive data mining is the process of estimation of the values based on the given data set. This can be achieved by regression analysis on the given data set. In this paper, we are specifying the various methods that can be used typically Polynomial Regression Technique, the various forms of implementing analysis, problems and possible solutions.

Index Terms— Data Mining, Prediction, Regression, Polynomial Regression, Multivariate Polynomial Regression.

1 Introduction

With the increasing use of computers in our data to day life, there is a continuous growth in the data. These data contain a large amount of known and unknown knowledge which could be utilized for various applications like for [1],[2] knowledge extraction, pattern analysis, data archaeology and data dredging. This extraction of unknown useful information is achieved by the process of Data Mining.

Data mining is considered as an instrumental development in analysis of data with respect to various sectors like production, business and market analysis. There are two forms of data mining namely: Predictive Data Mining and Descriptive Data Mining.

Descriptive data mining is the process of extracting the features from the given set of values. Predictive Data Mining is the process of estimating or predicting future values from an available set of values. Predictive data mining uses the concept of regression for the estimation of future values.

2 Regression

Regression is the method of estimating a relationship from the given data to depict the nature of data set. This relationship can then be used for various computations like for the forecasting future values or for computing if there exists a relation amongst the various variables or not.[3.]

Regression analysis is basically composed of four different stages:

1. Identification of dependent and independent variables.
2. Identification of the form of relationship among the variables like linear, parabolic, exponential, etc. by means of scatter diagram between dependent and independent variables.
3. Computation of regression equation for analysis.
4. Error analysis to understand how good the estimated model fits the actual data set.

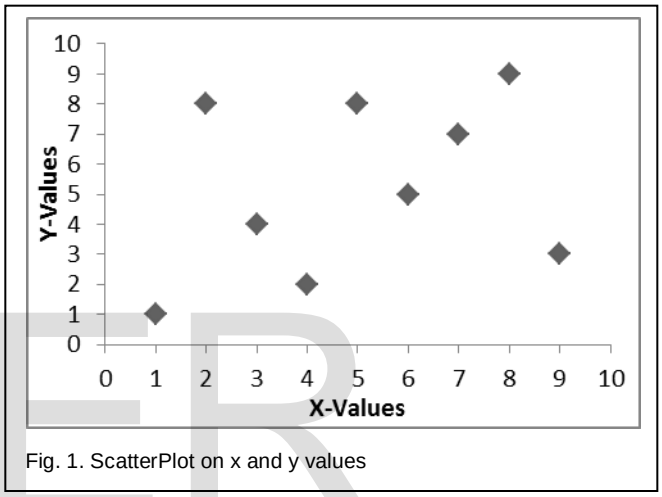

Fig. 1. ScatterPlot on x and y values

There can be 'n' number of ways of joining the given points in scatter graph. The idea of plotting the regression curve is that the diversion within the various points should be minimal. But if we simply compute the diversion to be minimal, i.e., $\sum_{i=1}^{n}(y - \hat{y})$ is minimal, we can again have n number of possibilities as the negative and the positives cancel out. So just computing minimal diversion is not appropriate.

There are mainly two methods for finding the best fit curve, namely, Method of Least Square and Method of Least Absolute Value. In Method of Least Square Value(LSV), the sum of the squares of the diversions is taken to be minimum. This sum is referred to as Sum of Square of Error.

$$SSE = \sum_{i=1}^{n}(y - \hat{y})^2 = \text{minimum.} \quad (1)$$

Another method of finding approximate regression curve is the method of Least Absolute Value (LAV). Here, it I s assumed that $\sum_{i=1}^{n}|y - \hat{y}|$ is minimum. This method has a drawback that finding a line that satisfies this equation is difficult. Furthermore, there may be no unique LAV Regression Curve.

There are various methods of Regression Analysis like: Simple Linear Regression, Multivariate Linear Regression, Polynomial Regression, Multivariate Polynomial Regression, etc.

In Linear Regression, a linear relationship exists between the variables. The linear relationship can be amongst one response variable and one regressor variable called as simple

- *Priyanka Sinha is currently pursuing Masters degree program inSoftware Technology in Velloe Institute of Technology,India, PH-91-9629785836. E-mail:er.priyakasinha@gmail.com*

linear regression or between one response variable and multiple regression variable called as multivariate linear regression. The linear regression equation is of the form:

$y = \beta_0 + \beta_1 x$ for simple linear regression (2)

$y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$ for multivariate linear regression (3)

The linear regression curve is of the form:

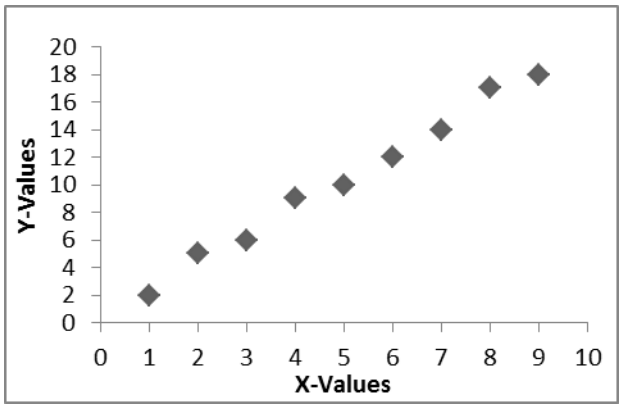

Fig. 2. ScatterPlot in Linear Regressionon x and y values

Quadratic Regression is the regression in which there is a quadratic relationship between the response variable and the Regressor variable. Quadratic equation is a special case of polynomial linear regression where the nature of the curve can be predicted. The equation is of the form:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 \quad (4)$$

and regression curve is a parabolic curve.

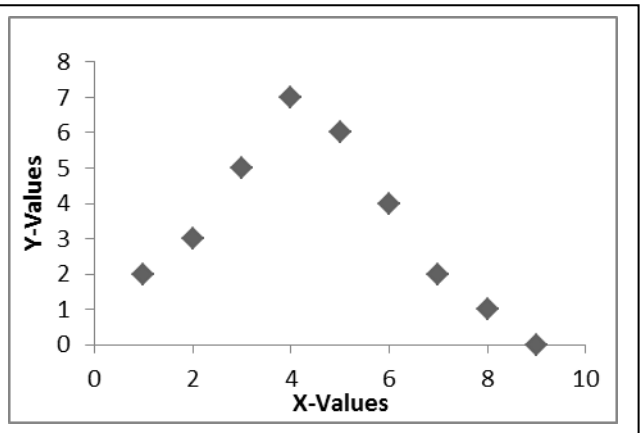

Fig. 2. ScatterPlot in Polynomial Regressionon x and y values

In Polynomial Regression, the relationship between the response and the Regressor variable is modelled as the n^{th} order polynomial equation. The nature of regression graph prediction is not possible in this case. The form is:

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i {x_i}^i \quad (5)$$

3 Polynomial Regression

Polynomial Regression is a model used when the response variable is non-linear, i.e., the scatter plot gives a non-linear or curvilinear structure.[3]

General equation for polynomial regression is of form:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_k x^k + \varepsilon \quad (6)$$

To solve the problem of polynomial regression, it can be converted to equation of Multivariate Linear Regression with k Regressor variables of the form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + \varepsilon \quad (7)$$

Where, $x_i = x^i$ (8)

ε is the error component which follows normal distribution $\varepsilon_i \sim N(0,\sigma^2)$

The equation can be expressed in matrix form as :

$$Y = X\beta + \varepsilon \quad (9)$$

Where, X, Y, β, ε are the vector matrix form representations which can be expanded as:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$ is the vector of observations (10)

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}$$ is the vector of parameters (11)

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$ is the vector of errors (12)

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k-1} \\ 1 & x_{21} & \cdots & x_{2k-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk-1} \end{pmatrix}$$ is the vector array or variables (13)

Estimation of parameters is done by Least Square Method. Assuming the fitted regression equation as:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \cdots + \hat{\beta}_k x_k \quad (14)$$

Then, by Least Square Method, minimum error is represented as :

$$SS_{RES} = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \cdots + \hat{\beta}_k x_k\right)^2 \quad (15)$$

The matrix representation for the above equation is given as:

$$SS_{RES} = Y'Y - 2\hat{\beta}X'Y + \hat{\beta}X'X\hat{\beta} \quad (16)$$

By partial differentiation with respect to the regression

equation model parameters, we have k independent normal equations which can be solved for solution of parameters, which is represented as:

$$\hat{\beta} = (X'X)^{-1}X'Y \quad (17)$$

Substitution of $Y = X\hat{\beta}$ gives error as:

$$SS_{RES} = Y'Y - \hat{\beta}X'Y = \sum_{i=1}^{n} e_i^2 \quad (18)$$

TABLE 1
ANOVA TABLE FOR REGRESSION PARAMETERS

ANOVA Table				
Source of Variation (Src)	Degree of Freedom (DF)	Sum of Square (SS)	Mean Square (MS)	F-Statistics
Regression (Reg)	k-1	SSReg	MSReg= SSReg/(k-1)	F = MSReg/ MSRes
Residual (Res)	n-k	SSRes	MSRes= SSRes/(n-k)	
Total (Tot)	n-1	SSTot	MSTot= SSTot/(n-1)	

Here, Degree of freedom for the regression equation is (n-k).Significance of the regression equation can be estimated by means of Analysis of Variance table called ANOVA table.

This Multiple Linear Regression model can be used to compute the Polynomial Regression Equation.

4 MULTIVARIATE POLYNOMIAL REGRESSION

Polynomial Regression can be applied on single Regressor variable called Simple Polynomial Regression or it can be computed on Multiple Regressor Variables as Multiple Polynomial Regression[3],[4]. A Second Order Multiple Polynomial Regression can be expressed as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon \quad (19)$$

Here,

β_1, β_2 are called as linear effect parameters.
β_{11}, β_{22} are called as quadratic effect parameters.
β_{12} is called as interaction effect parameter.

The Regression Function for this is given as:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 \quad (20)$$

This is also called as the Response Surface.
This can again be represented in Matrix form as:

$$Y = \beta X + \varepsilon \quad (21)$$

The parameter for the given equation can be computed as:

$$\hat{\beta} = (X'X)^{-1}X'Y \quad (22)$$

And the Computed Regression Equation is represented as:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_{11} x_1^2 + \hat{\beta}_{22} x_2^2 + \hat{\beta}_{12} x_1 x_2 \quad (23)$$

5 PROBLEMS WITH MULTIVARIATE POLYNOMIAL REGRESSION

The major issue with Multivariate Polynomial Regression is the problem of Multicolinearity. When there are multiple regression variables, there are high chances that the variables are interdependent on each other. In such cases, due to this this relationship amongst variables, the regression equation computed does not properly fit the original graph.

Another problem with Multivariate Polynomial Regression is that the higher degree terms in the equation do not contribute majorly to the regression equation. So they can be ignored. But if the degree is each time estimated and decided I required or not, then each time all the parameters and equations need to be computed.

6 SOLUTIONS FOR MULTIVARIATE POLYNOMIAL REGRESSION PROBLEMS

Multicolinearity is a big issue with Multivariate Polynomial Regression as it restricts from the proper estimation of regression curve. To solve this issue, the Polynomial Equation can be mapped to a higher order space of independent variables called as the feature space. There are various methods for this like: Sammon's Mapping, Curvilinear Distance Analysis, Curvilinear Component Analysis, Kernel Principle Component Analysis, etc. These methods transform the related regression variables into independent variables which results in better estimation of the regression curve.

The solution to the problem of computation of parameter each time for increase in order can be solved by computation using Orthogonal Polynomial Representation as:

$$y = \alpha_0 P_0(x) + \alpha_1 P_1(x) + \cdots + \alpha_k P_k(x) + \varepsilon \quad (24)$$

7 CONCLUSION

Data Mining in real time problems consist of variety of data sets with different properties. The prediction of values in such problems can be done by various forms of regression. The Multivariate Polynomial Regression is used for value prediction when there are multiple values that contribute to the estimation of values. These may be related to each other and can be converted to independent variable set which can be used for better regression estimation using feature reduction techniques.

REFERENCES

[1] Han, Jiawei and Kamber, Micheline. (2001). Data Mining Concepts & Techniques, Elsevier

[2] Fayyad, Usama, Pietetsky-Shapiro, Gregory, and Symth, Padharic. Knowledge Discovery and Data Mining: Towards a Unifying Framework (1999). KDD Proceedings, AAAI

[3] Gatignon, Hubert. (2010). Statistical Analysis of Management Data Second Edition. New York: Springer Publication.

[4] Kleinbaum, David G, Kupper, Lawrence L, and, Muller, Keith E. Applied Regression Analysis and Multivariable Methods 4th Edition. California: Thomson Publication.