Volume 3, Issue 4, April-2016, pp. 162-164 ISSN (O): 2349-7084

International Journal of Computer Engineering In Research Trends

Available online at: www.ijcert.org

Data Trend Analysis by Assigning Polynomial Function For Given Data Set

Dhaneesh T,
Student, Department of Computer Science, Christ University, Bangalore

ABSTRACT:-This paper aims at explaining the method of creating a polynomial equation out of the given data set which can be used as a representation of the data itself and can be used to run aggregation against itself to find the results. This approach uses least-squares technique to construct a model of data and fit to a polynomial. Differential calculus technique is used on this equation to generate the aggregated results that represents the original data set.

Keywords - Curve Fitting, Trend Analysis in Data, Data Analytics

1. INTRODUCTION

Data analytics (DA) could be defined as science of examining raw data with the purpose of drawing conclusions about the information contained within.

Applications of Data Analytics or Trend Analysis are:

- Useful in computing customer satisfaction metrics for various products in the market.
- Used for identifying trends in the market over a period of time specific to a region on the globe
- To forecast stock market movement based on earlier trends that are identified etc…

2. EXISTING MODELS

Existing models of Data Analysis requires a vast storage space to accommodate the volume of data. This works aptly for the enterprise which has focus over variety of metrics and grows dynamically over a short period of time. However for the organizations that has a clear focus on predefined metrics from the data which is their concern for analysis, it would not be feasible for them to support the storage space requirement which increases the cost.

3. FITTING DATA TO A POLYNOMIAL EQUATION: THE APPROACH

Consider we take a data set and represent them in the vector form of (x_i, y_i) pair, where x - the independent variable and y is the dependent variable. We assume that the data that we would want to represent as an equation fits into a polynomial of nth order. Hence the problem for us to fit this data into the equation would be to find the coefficients c_i for every pair of (x_i, y_i),

$$y_i = c_1 + c_2 x_i + c_3 x_i^2 + \cdots + c_{n-1}$$

This is the general equation for $(n+1)$ unknowns. We know that to find the $(n+1)$ unknowns, we need $(n+1)$ equations and each one would arise from an (x, y), (independent, dependent) pair. Thus now the task would to solve this set of equations to find the coefficients and represent the data.

Our assumption about the data to be fit is that each data points that we derive from the given dataset is in the format of independent and dependent pair (x_i, y_i). For example, consider if we want to fit the data about the yield that a farmer gets from a crop depending on the amount of fertilizer that the farmer uses. This is a slow growing metrics which we would want to store

and run analytics over a large period of time. Here in this example, the independent variable x will be fertilizer usage in tones and the crop yield in tones will be the dependent variable. The graph for this example would be as given below in the Fig 3.1:

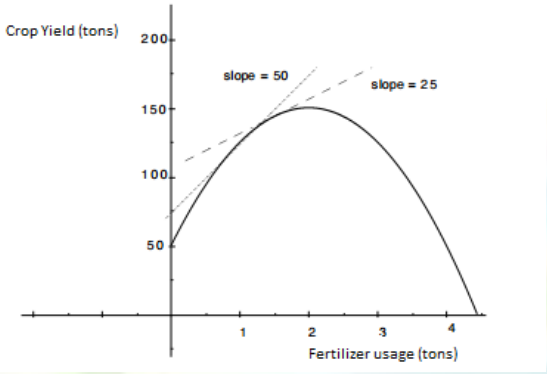

Fig 3.1: Sample Data Fit to a Curve

Our initial task would be to generate the polynomial equation for the dataset that represents the curve that is shown in the above figure. Once we have the equation, we can run differential equation on that to identify the crop yield of the farmer corresponding to the amount of the fertilizer that the farmer used. Decision about the metrics selection still rests with the requirement and is out of scope of this paper.
Once the metrics are chosen and the polynomial equation are derived, then the same can be stored with a minimum space requirement than storing the whole dataset itself.

4. DIFFERENTIAL CALCULUS: FINDING THE SLOPE AT A GIVEN POINT

Differential calculus, in practicality, is about describing in a precise fashion the ways in which related quantities change. This fundamentally lays down the idea about comparison and analytics of data.

Given a function $f(x)$, the derivative of the function is denoted as $f'(x)$ where $f'(x)$ is the rate of change of the function with respect to the variable x.

Example

Let's take a scenario where a production facility, which is capable of producing 60,000 artifacts a day, decides to store and manage their production costs summary. We look at two metrics here which are the number of artifacts produced each day and the cost of production for that day. Using the data fit method that is described earlier; the cost versus number of artifacts produced is mapped to the cost function as given below:

$$C(x) = 250{,}000 + 0.08x + \frac{200{,}000{,}000}{x}$$

Instead of storing the entire dataset the facility decides to store this cost function and run analytics over this. Let's try to find a business case where the task would be find the number of artifacts that they should produce in a day to minimize the production cost.

The criteria here would be to minimize the cost subject to the fact that the value of x, the number of artifacts that the facility can produce in a day, should be in the range $0 \leq x \leq 60{,}000$.

The first set of derivatives for the above cost function would be:

$$C'(x) = 0.08 - \frac{200{,}000{,}000}{x^2}$$

$$C''(x) = \frac{400{,}000{,}000}{x^3}$$

Thus the critical points of the cost function is given by:

$$C'(x) = 0 \quad \rightarrow \quad 0.08 - \frac{200{,}000{,}000}{x^2} = 0$$

Solving the above equation we get the value of x.

$$0.08x^2 = 200{,}000{,}000$$

$$x^2 = 2{,}500{,}000{,}000$$

$$x = \pm\sqrt{2{,}500{,}000{,}000} = \pm 50{,}000$$

From the above value, it could be clearly stated that the negative value for x can be omitted. Thus this concludes that the facility should produce 50,000 artifacts in a day to minimize the cost of production. This value is purely based on the cost function that is generated by a sample production detail. Thus variation in the cost function would result in a different value.

5. RESULTS AND DISCUSSIONS

We have presented a general idea about how a polynomial equation that is generated from the given

dataset would solve the problem of data analysis. Scilab was used to demonstrate the method of data fit to a curve using Least Squares Method to reduce the noise factor while considering a vast data set. The sum of the squares of the offset values are used instead of the offset absolute values itself because this allows the residuals to be treated as a continuous differentiable quantity and thus reducing the noise. A sample curve that was generated using the least squares method is shown below in the Fig 5.1

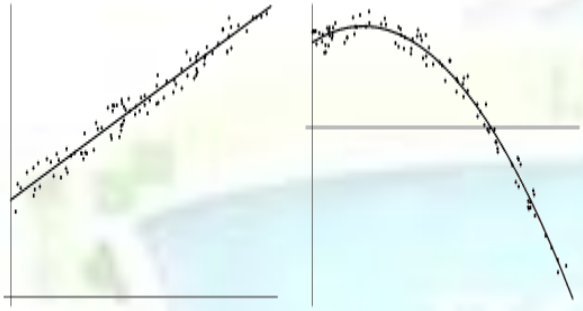

Fig 5.1: Sample Curve generated using Least Squares Method

6. CONCLUSION

Scilab tool was used to demonstrate the creation of polynomial equation for the given data set.Future enhancements for this idea will include usage of Machine Learning algorithms to generate the polynomial equations for the given data set. This will help us in creating a model that can update the existing equation due to merging of more data with the archived data for the same metrics.

REFERENCES

[1] Haijun Chen, "A SPECIAL LEAST SQUARES METHOD FOR CURVE FITTING", Measurement and Control Group, Dept. Electrical Engineering, Eindhoven University of Technology, Postbus 513, 5600 MB Eindhoven, The Netherlands.

[2] Aimin Yang, "The Research on Parallel Least Squares Curve Fitting Algorithm" College of Science Hebei Polytechnic University Tangshan, Hebei Province, 063009 China.

[3] Junyeong Yang and Hyeran Byun "Curve Fitting Algorithm Using Iterative Error Minimization for Sketch Beautification", Dept. of Computer Science, Yonsei University, Seoul, Korea, 120-749

[4] G. Taubin, "An improved algorithm for algebraic curve and surface fitting", Proc. Fourth 658-665, Berlin, Germany.

