**LINEAR REGRESSION AND CURVE FITTING BY METHOD OF LEAST SQUARES**

1. **INTRODUCTION:**

The method of least squares for fitting a curve assumes that the best fit curve for a given dataset is the one which has the minimal sum of deviations from the actual value, i.e., least square error from a given set of data.

This method helps us arrive at an approximate equation which governs the values of the given dataset, such that any new datapoint obtained has the highest probability of fitting into the given curve, i.e., the new value already lies on the given curve or in its vicinity.

**2. IMPLEMENTATION AND METHODOLOGY:**

Suppose the dataset initially consists of *n values say (x*1, y1), *(x2*, y2), …….., *(xn*, yn), where x is the independent variable and y is the dependent variable on x.

We then plot the graph of y=f(x) for all the existing datapoint and try to observe the nature of the curve. We determine if the resulting graph closely resembles linearity, or it is quadratic or cubic (or in general the degree of the polynomial if in case it is), or exponential or logarithmic.

Depending on our analysis of the curve we determine the nature of the curve fitting function F(x).

In case the graph is linear, F(x)=ax + b, where a, b are real/complex constants.

In case the graph is quadratic, F(x)=ax2+ bx+ c, where a, b, c are real/complex constants and so on.

**Hence the plotting of graph of a dataset having n points takes a total time of**

**O(n)————— (1)**

Sometimes the resulting function F(x) exhibits two traits, i.e., it is a linear combination of two or more functions.

F(x)=F1(x)+F2(x)+F3(x)+………

Let the deviation in the value of Y=F(x) from the actual value y=f(x) be denoted by the letter ei , the subscript representing the error e for datapoint i.

Then,

e1 = y1 - F(x1)

e2 = y2 - F(x2)

e3 = y3 - F(x3)

.

.

.

en = yn - F(xn)

According to the method of least squares the best fitting curve has a property that,

should have minimum value for all possible choices of F(x) that satisfy the dataset values.

**3. LEAST SQUARE METHOD FOR** **LINEAR DATASET:**

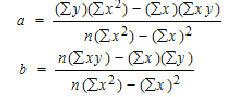
For a linear set of datapoint the fitting function F(x) will be of the form,

F(x) = ax + b, where a and b are real/complex constants.

To satisfy the property of minimum deviation, the sum of all values of [ yi - F(xi) ] must be minimum.

Implies that the first derivative of the above function with respect to the independent variable must be zero.

**This leads us to calculate the above value for all n datapoints which again take a time of O(n)—————————— (2)**

The constants a and b can also be alternatively calculated by the formula directly

**In this method we calculate the sum of all values of x as well as y both which take a time of the order of n**

In both cases the total time needed to calculate a and b and hence arrive at the best fit curve is of the order n.

Therefore the total time taken

T = n + n = 2n ———————- from (1) and (2).

Hence **T(n) = 2n = O(n)**. The total time needed for this regression method is of the order n.

-GIRIDHAR K. SHANBHAG

-1RV18CS056

-CSE / 4B