UNKNOWN SIGNALS RECONSTRUCTION USING LEAST SQUARE METHOD

I. Introduction

The aim of this report is to demonstrate how we can train a system using a set of data to gives us a model which we can use to make predictions on similar or different data. In this case study we analyse an unknown signal.

Using the given training data, the approach taken here is to make use of the Least Squares Method (LSM) a regression analysis that would help us minimise the square error of the residuals taken from the results of every single equation fitting the data. This would serve as a mean to narrow down our assumption of what regression functions would best fit the data. We also make use of Cross-Validation (CV) as a way to check how sensible the square errors (SSE) are to a subset of points from the data and hence through their comparison, emphasise on the correctness of the best line of fit.

II. Program Methodology

Given a set of data of x- coordinates and y- coordinates; using an analysis from their graphs, and properties such as how y changes according to x, we can factor out a list of plausible regression functions which would match the data of the unknown signal by matching it to known functions that we know such as linear, polynomial and periodic functions.The next step is a process of trial and error. Given that the signal is made up of 3 functions only. We test functions that are likely to fit the data and only use those with minimum SSE.

A line segment would consists of 20 different points; we split the data such that we can treat each line segment case by case. Using the least square method, we then compute an estimate of the regression line on which these data points would lie.

The Linear and polynomial least square regression can be calculated using the below formula.

$$Y_{(N\times1)} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix} X_{(N\times(K+1))} = \begin{pmatrix} 1 & X_1^1 & \dots & X_1^K \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_N^1 & \dots & X_N^k \end{pmatrix}, A_{((K+1)\times1)} = \begin{pmatrix} A_0 \\ A_1 \\ \vdots \\ A_K \end{pmatrix}$$

$$\text{Given that } K = \begin{cases} 1, if \text{ linear polynomial} \\ c, if \text{ degree c-polynomial} \\ where\ c = \text{2,3,4....} \end{cases}$$

Using the above formula we get the coefficient vector $A_LS = (X^TX)^{-1}X^Ty$ given a regression line $y_i = a_0 + a_1x + a_2x_i^2 + \ldots + a_kx_i^k$. With this same general process , we can extend this matrix form to accommodate periodic functions such as Sin(x). This would look like this .

$$Y_{N\times 1} = \begin{pmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ Y_N \end{pmatrix} \quad X_{(N\times 2)} = \begin{pmatrix} 1 & \sin(x_1) \\ \vdots & \vdots \\ 1 & \sin(x_N) \end{pmatrix}$$

Using the above formula the coefficient vector results to $A_LS = (X^TX)^{-1}X^Ty$ giving a regression line $y_i = a_0 + a_1 \sin(x_i)$

We then use these regression lines to calculate the total reconstruction error using the

Squared Sum Error (SSE) method sum up over each data data segment.

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

III. Results and Analysis

An interesting trend is to notice how the square error is affected by a regression line or order K along the different train data . From the table below you would notice how the square error increases notably the spike in SSE when k= 4 for ***adv_3.csv.*** In addition, as ***Basic_3.csv*** is only made of a single data segment whose points demonstrate a parabola shape.It is then safe to make the assumption that a certain value k would be ideal to avoid any case of overfitting and under-fitting of the data.The variations in SSE are further investigated in the difference between the graphs of ***adv_3.csv*** with k= 3 and k= 4 in figure 1 and 2.

Looking at both figures we can generally agree that with k = 3, the model fits the data more than when k = 4. If we dive more on the last segment of each plot; We realise that while the regression curve on the last segment in figure 3 matches well with the data, the regression line in figure 4 missed the general trend of the data points. This would indicate that a high value for k such as 4, would lead to under-fitting of the data. Using cross validation we can convince ourselves that the analysis hold.

Train_data	SSE with k= 3 (optimal)	SSE with k=2	SSE with k = 4	SSE with k = 5
basic_3.csv	1.438×10^{-18}	15.743	2.601×10^{-12}	4.505×10^{-08}
adv_3.csv	979.592	986.583	91487.029	92004.166

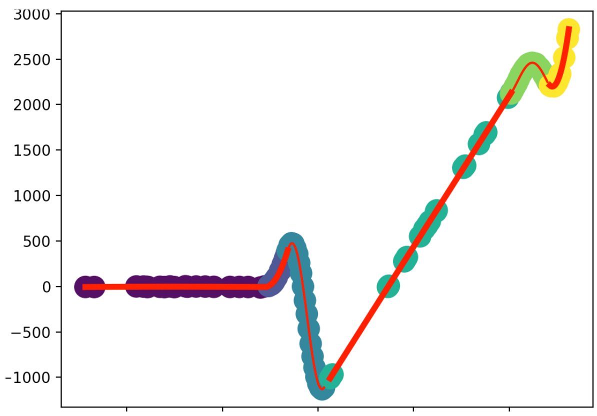

Figure 1 adv_3.csv with k=3

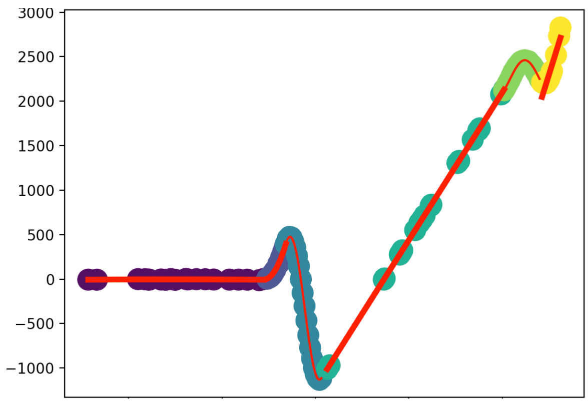

Figure 2 adv_3.csv with k = 4

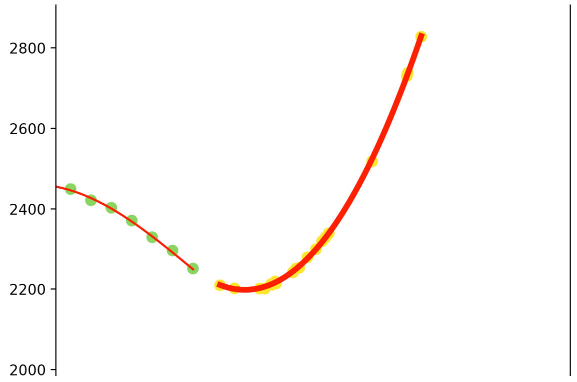

Figure 3 adv_3.csv , last segment with k =3

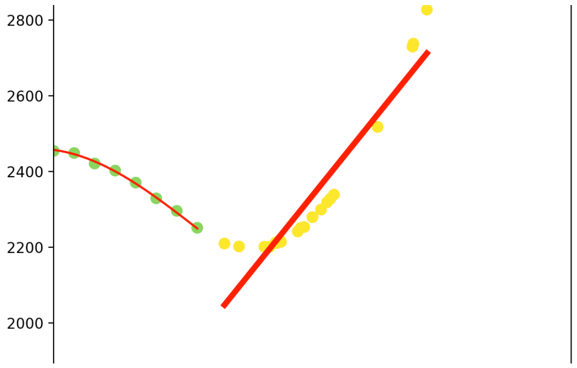

Figure 4 adv_3.csv, last segment with k=4

Applying the same analysis to the training data showing a periodic regression, we can infer that the optimal regressions for the model are linear , cubic and sinusoidal.Some of the results using this conclusion can be shown on the table below.

train_data	Optimal SSE	Type of Regression per segment
basic_1.csv	1.688×10^{-28}	Linear
basic_2.csv	6.215×10^{-27}	Linear,linear
basic_4.csv	1.385×10^{-12}	Linear, cubic polynomial
basic_5.csv	1.05×10^{-25}	Sine
adv_2.csv	3.650	Sine, linear, sine
adv_3.csv	979.592	Linear,cubic,sine,linear,sine,cubic
noise_2.csv	797.917	Linear, cubic

When looking at the sum of the advanced and noise training data, we notice that the model returns a high reconstruction error for the data. The validity of the model to return such results can be based on the fact that each observation does not affect the process of obtaining the next observation. As we train the model it picks up suitable regression functions for the data. A high SSE would simply mean that the model could not find the best match for the data .Given that these data are simply noise and could not represent anything constructive under the regression functions we used; Having an overall low SSE would have suggested that the generalisation of the model is unreliable. We want the model to be able to pick up the trend of the data, not to match exactly all the data points given.

IV. Conclusion

The linear, cubic and sinusoidal regressions gives us an overall generalist model which minimises overfitting and under-fitting and hence suitable for use on new data. A possible extension is the regularisation of the data to improve the accuracy of the model.This is a form of regression, that constrains the coefficient estimates towards zero hence discouraging learning a more complex model, so as to avoid overfitting.