Unknown Signal

Elan Virtucio

1 Introduction

It is very useful to be able to model a given set of data points to an appropriate degree of accuracy. This can allow for predictions to be made for an output given some input.

In this instance, a set of data points is given which follows an unknown signal. The task given was to reconstruct this signal and alongside it, calculate the sum squared error or residual sum of squares (RSS) to give an idea of how well the model represents the data. There are different segments of this signal and each one can be modelled by either a linear function, a polynomial function of a given degree or some other function that is unknown.

Although it's possible to model a set of data, ultimately, the results will be highly dependent on the accuracy and correlation of the data and therefore problems may arise such as overfitting. The aims of this project were to minimise these effects and address the limitiations of modelling a data set.

2 Implementation

The program is *lsr.py* and it takes in Comma Separated Value (CSV) files consisiting two columns for the x and y data points respectively. The files can contain one or multiple line segements and each line segments consists of 20 data points. The segments are split up and for each one, a model is fitted and it's RSS is calculated. The RSS of each line segment are summed up to produce the total RSS.

The regression method used was the matrix form of the Least Squares Method (LSM) and to account for any form of overfitting, the use of a k-fold cross-validation (CV) was used. There are two implementation that can be used, a fixed k-fold or a random k-fold. For example, if $k = 5$, in a fixed k-fold, the parts are split up like so, $[[0,1,2,3],[4,6,7,8],\dots,[16,17,18,19]]$ where the numbers represent the index of the data point. Whereas, in a random k-fold, the parts can be split up like so, $[[8,3,19,7],[10,2,14,4],\dots,[15,1,0,8]]$, and this can differ per run.

The program iterates through a list of defined models. These models are a list object of their own and contains the properties that'll allow it to model a data set using LSM. These properties are it's name, the relevant function required to extend the X vector with the relevant feature vectors, where X consists of the x data points, and additionally, the equation to use. The data is modelled using each of this models and the RSS is calculated for each one or to be more precise, the CV error, which corresponds to the average RSS value that was calculated during the CV process. The model that produced the minimum error gets chosen as the model to use for that data set.

Table 1: Models

Name	X	$\hat{y}$
Linear	$\begin{bmatrix}1 & x_0 \\ \vdots & \vdots \\ 1 & x_{N-1}\end{bmatrix}$	$a_0 + a_1 x_i$
Polynomial	$\begin{bmatrix}1 & x_0 & \dots & x_0^d \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N-1} & \dots & x_{N-1}^d\end{bmatrix}$	$a_0 + a_1 x_i + \cdots + a_d x_i^d$
Exponential	$\begin{bmatrix}1 & e_0^x \\ \vdots & \vdots \\ 1 & e_{N-1}^x\end{bmatrix}$	$a_0 + a_1 e_i^x$
Sine	$\begin{bmatrix}1 & sin(x_0) \\ \vdots & \vdots \\ 1 & sin(x_{N-1})\end{bmatrix}$	$a_0 + a_1 sin(x_i)$
Cosine	$\begin{bmatrix}1 & cos(x_0) \\ \vdots & \vdots \\ 1 & cos(x_{N-1})\end{bmatrix}$	$a_0 + a_1 cos(x_i)$

[1] $N = number\ of\ data\ points\ \therefore\ N = 20.$

[2] $a_i \in A = (X^T X)^{-1} XY \ \mid\ Y = \begin{bmatrix} y_0 \\ \vdots \\ y_{N-1}\end{bmatrix}$

[3] $d \in \mathbb{N} \mid 2 \leq d \leq 10$

Table 1

3 Result

4 Conclusion