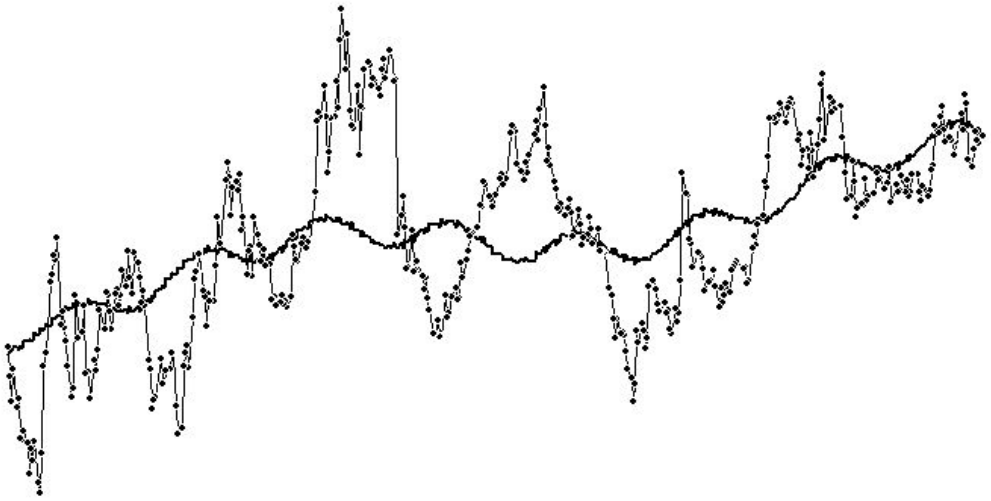
initiative for
interdisciplinary
research

http://www.i2r.org

Genetic Regression



By Wojciech Gryc (wojciech@i2r.org)
- - - http://www.i2r.org/

PROGRAM INTRODUCTION

Genetic Regression is a program designed to fit sine curves and multiples thereof to data sets. It accepts text-based files with x and y values separated by a set of spaces, like so:

3 4.00
7 3.75
9 8.23
...and so on.

Save the data in *data.txt* within the same directory as the JAR file, and the program will automatically detect it and load the data.

The final result will be displayed in the following format:

A: 3.4002 B: 5.2421 C: 6.2312 D: 0.0923
A: 1.3242 B: 3.2423 C: 5.2421 D: 5.2342

Average Difference: 0.003424

The above output can be interpreted as sine curves of the form:

A*sin(Bx+C) – D

Each sine curve is multiplied by the other sine curves, and the *Average Difference* is the average absolute difference between the genetically regressed function and the data point at the corresponding x value.

OPTIONS WITHIN THE PROGRAM

Clicking *Evolve Curves* will automatically start analysis with the program's default settings, which are 1000 generations and one sine curve multiple.

To change the settings above, go to *Options – Regression Settings* and you will be able to modify the number of generations and the number of sine curve multiples to use.

Using *Options – Debug* will begin analyzing the curve and rather than finding the answer, will return the lowest *Average Difference* of each generation to show how the evolution is progressing.

THEORY BEHIND THE PROGRAM

The program analyzes data and fits a multiple of sine curves onto the data through the use of a genetic algorithm. When the data is first loaded, the program finds the range and domain of the data and then finds random numbers corresponding to the A, B, C, and D coefficients for each sine equation. The sine curves are then multiplied to get a function, like so:

$$\mathbf{f(x) = [A_1*\sin(B_1x+C_1)+D_1] * [A_2*\sin(B_2x+C_2)+D_2] * \ldots * [A_n*\sin(B_nx+C_n)+D_n]}$$

...where *n* is the number of sine curves chosen.

Once the randomly generated coefficients are applied into the functions, each function runs through the data. The best functions – chosen according to lowest *Average Difference* – are then taken and are evolved using a genetic algorithm.

This algorithm works by taking the *Average Difference* and by modifying each sine curve's A, B, C, and D values by a fraction of the *Average Difference.* This is done a number of times, and as the *Average Difference* gets smaller, is a form of simulated annealing.

Average Differences decrease through an exponentially decreasing pattern. The first few generations drastically lower their *Average Difference*, but subsequent generations have much smaller improvements.

MORE INFORMATION

If you are interested in learning more about the program or have questions or comments, please e-mail the program's creator, Wojciech Gryc, at wojciech@i2r.org or visit http://www.i2r.org/.