Symbolic Regression Summary Report

Hridoy Rahman
February 16, 2024

1 Summary

In Symbolic Regression, the idea is to create mathematical expressions that accurately model the relationship between the input and the output variables provided in a training dataset. In other words, the symbolic regression process finds a mathematical expression that best fits with the provided dataset. In genetic programming, SR (Symbolic Regression) is to find a perfect fit of the data points provided in a dataset and search for a function from a possible search space of s-expression (Symbolic expression) that fits the given data points, which is derived from a set of predefined functions and terminals. Furthermore, symbolic regression aims to find the expression that best fits the points and aims to find a perfect or closest match between the predicted and the actual output values.

Objective	Search for a function based on the independent variable and dependent variable, in symbolic form which fits a given set of data points of 10 (x,y) where the target function is $x^2 + 3x + 4$
Terminal Set	X (Independent Variable)
Function Set	+, -, *, (Protected Division) %, SIN, Ephemeral
Fitness Cases	provided dataset with 21 data points where $X_i \in [-10, 10]$
Raw Fitness	The fitness is the dependent value acquired from the generated S-expression minus the independent variable's (x) actual dependent value (Y_i) from the actual dataset.
Standardized Fitness	same as raw fitness
Hits	Number of times, "the value of the dependent variable of s-expressions minus the value of the provided dependent variable is $<= 0.1$ "
Success Predicate	An S-expressions that score 21 hits.

Table 1: Table of Symbolic Regression

1.1 Fitness Evaluation & Processed Data

The data is processed by taking 21 points within this [-20,20] interval. Then, the function $x^2 + 3x + 4$ is applied to get Y_i for each X_i from the dataset. The fitness formula minimizes

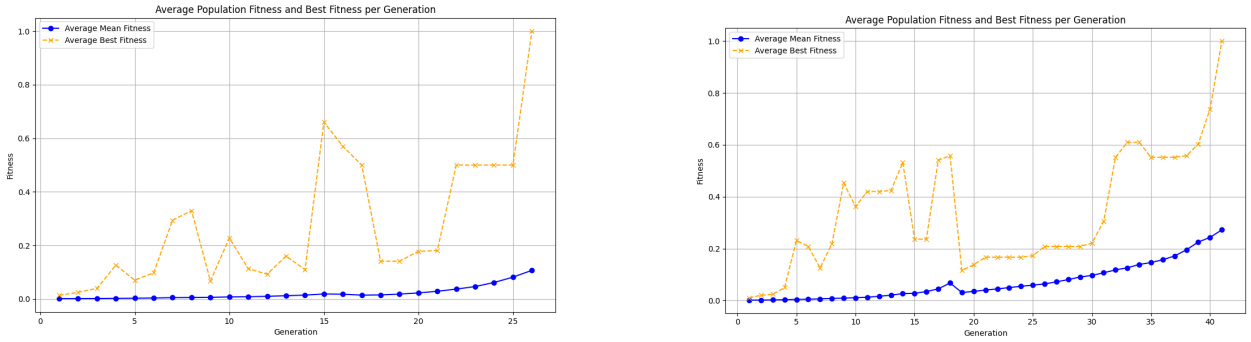
error by getting the dependent value of s-expressions by subtracting the value from the target value provided in the dataset Y_i; this is applied for each data point provided in the dataset. The evaluation is to check if the s-expression is close to the dependent value with respect to the target value. It finds the number of hits it gets by checking if the expected result by subtracting the generated S-expressions dependent value is $<= 0.1$. This is checked for each of the data points from the dataset. A successful predicate is determined when a generated s-expression hits all the data points provided in the dataset.

1.2 Table of GP Parameters

Parameter	Value
Population Size	1024
Generation Size	51
Crossover & Mutation Rate (Variation 1 without Elitism)	90% & 10%
Crossover & Mutation Rate (Variation 2 without Elitism)	100% & 100%
Crossover & Mutation Rate (Variation 3 with Elitism)	90% & 10%
Crossover & Mutation Rate (Variation 4 with Elitism)	100% & 100%
Selection Method	Tournament Selection
Tournament Size	3
Elites	2
Runs (Jobs)	10

Table 2: Table of Parameters Used

1.3 Parameter Tuning



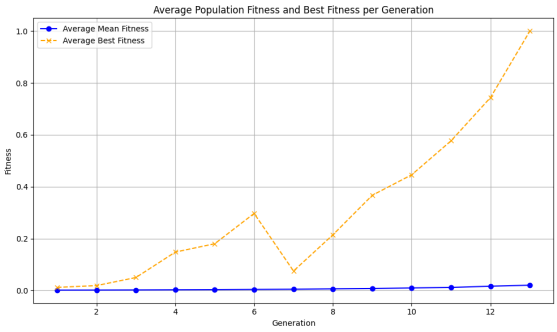
(a) Parameters: 90% & 10% without Elitism

(b) Parameters: 100% & 100% without Elitism

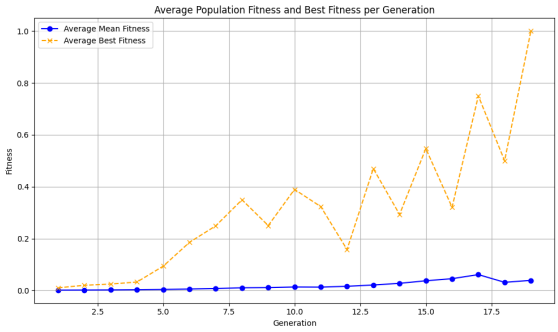
Figure 1: Average of Mean and best Fitness plotted in a graph across 10 runs

The parameters experimented with are provided in the table of GP parameters, where four variations are used. Variations 1 and 2, with crossover and mutation probability rates of 90%, and 10% and 100% & 100%, are used to run the program, and no elitism was used in these two variations. Variation 1 provides better results than variation 2. Variation 1 converges towards the solution rather early than variation 2. In 26 generations, it finds a

maximum value of 1 for fitness, which indicates the best solution possible. Each run had 21 hits on each parameter variation, meaning it always found the solution. However, the mean fitness of variation 1 is less than variation 2. Furthermore, {90% and 10% with two elitism}, had the smallest run and best average score of mean and best fitness. Variation 4 with elitism fluctuates significantly over the generations to find the best fit. Finally, of all the parameter variations, variation 3 had the best results with respect to other parameters.



(a) Parameters: 90% & 10% with Elitism



(b) Parameters: 100% & 100% with Elitism

Figure 2: Average of Mean and best Fitness plotted in a graph across 10 runs