LEARNING THE GOVERNING EQUATIONS OF ECOLOGICAL DYNAMICS

Markus Bauer,
Samson Mont,
Nour Rahal-Arabi,
Madelyn Smith,
Jinshui Wang

2021-2022

Markus Bauer

B.S. Computer Science - Data Science

Role: Machine Learning Developer

Focus: Balanus and Mytilus Growth Rate

Samson Mont

B.S. Computer Science - Artificial Intelligence

Role: Machine Learning Developer

Focus: Prey Density and Temperature
Predicting Feeding Rate

Nour Rahal-Arabi

B.S. Computer Science - Cybersecurity

Role: Machine Learning Developer

Focus: Symbolic Regression Output Equation Interpretation/Parsing

Madelyn Smith

B.S. Computer Science - Custom Applied Plan

B.S. Sustainability

Role: Lead Researcher

Research Focus: Relevant Work (Symbolic Regression in Ecological Studies)

Jinshui Wang

B.S. Computer Science - Machine learning

Role: Machine Learning Developer

Focus: Growth rate of mussel

Our Task: Answer a fundamental question in Ecology

How and why do species abundances vary over space and time?

PROJECT DESCRIPTION

Create a machine learning model to describe abundance dynamics as influenced by predator-prey interactions in the Oregon marine intertidal which is easy to understand and accurate, based on time-series data from a three-year ecological field experiment

An Introduction to gplearn

A genetic programming package in python

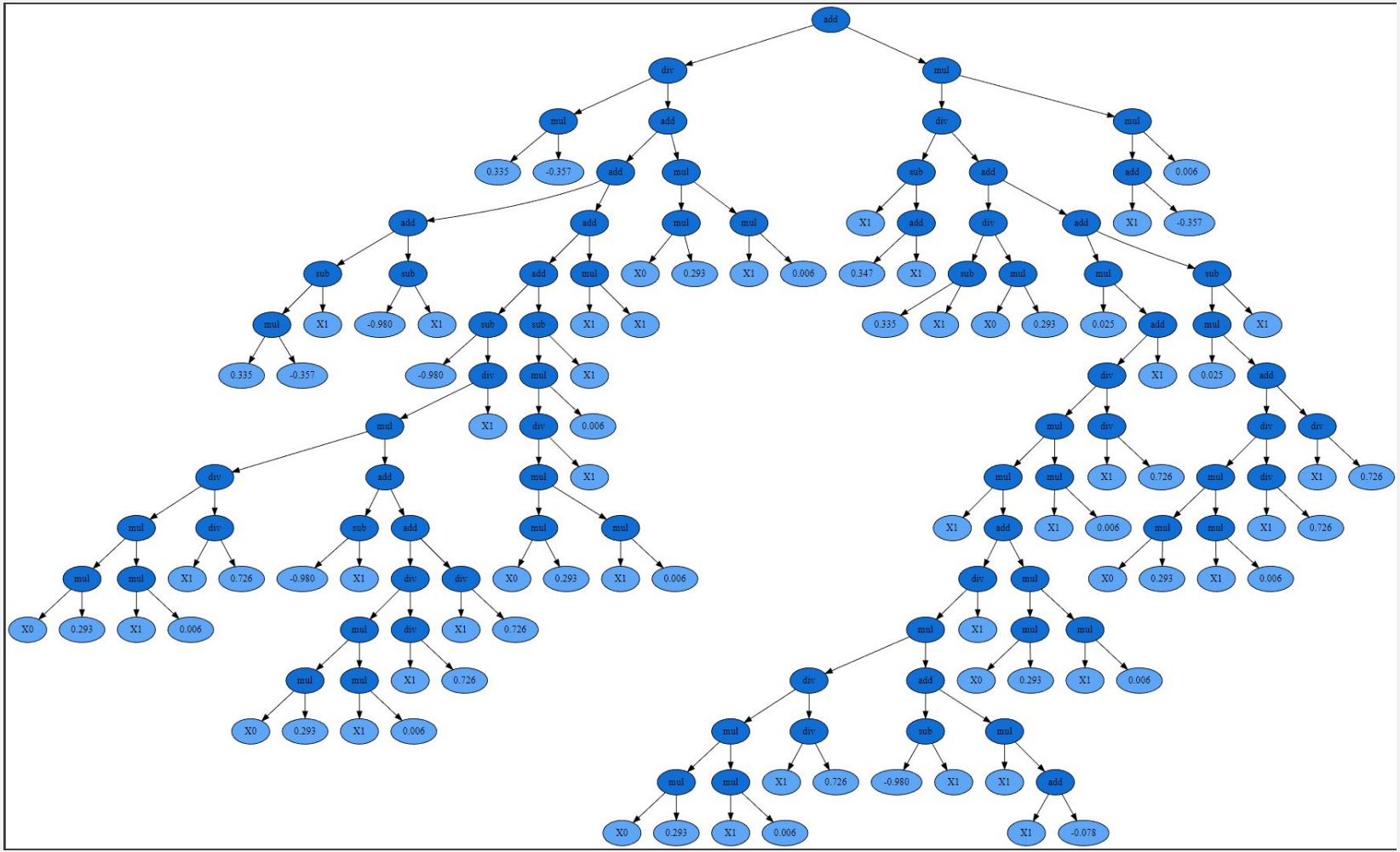
Symbolic Regressor

The Tool of Choice

```
classgplearn.genetic.SymbolicRegressor(*,
population_size=1000, generations=20,
tournament_size=20, stopping_criteria=0.0,
const_range=(- 1.0, 1.0), init_depth=(2, 6),
init_method='half and half', function_set=('add',
'sub', 'mul', 'div'), metric='mean absolute error',
parsimony_coefficient=0.001, p_crossover=0.9,
p_subtree_mutation=0.01,
p_hoist_mutation=0.01,
p_point_mutation=0.01, p_point_replace=0.05,
max_samples=1.0, feature_names=None,
warm_start=False, low_memory=False,
n_jobs=1, verbose=0, random_state=None)
```

1/3 Parsimony Coefficient

- Most important parameter – and the most frustrating to work with
- Encourages the model to pick smaller programs to avoid equation bloat
- High Parsimony means short equations
  - Can get "stuck"
- Must be relative to the results, high y values will need much higher parsimony
  - Was a problem during our project


add
div
mul
mul
add
div
mul
0.335
-0.357
add
mul
sub
add
mul
mul
add
sub
sub
add
mul
mul
X1
add
div
add
X1
-0.357
sub
sub
add
mul
X0
0.293
X1
0.006
0.347
X1
sub
mul
mul
sub
mul
-0.980
X1
-0.980
X1
sub
sub
X1
X1
0.335
X1
X0
0.293
0.025
add
mul
X1
0.335
-0.357
-0.980
div
mul
X1
div
X1
0.025
add
mul
X1
div
mul
div
div
div
mul
mul
X1
0.726
mul
div
X1
mul
div
add
sub
add
mul
mul
X1
add
X1
0.006
mul
mul
X1
0.726
X0
0.293
X1
0.006
X0
0.293
X1
0.006
mul
div
X1
0.726
div
mul
mul
X1
mul
mul
X0
0.293
X1
0.006
mul
div
X1
0.726
div
add
mul
mul
X0
0.293
X1
0.006
mul
div
sub
mul
mul
mul
X1
0.726
-0.980
X1
X1
0.726
X0
0.293
X1
0.006
add
X1
-0.078

2/3 Trials and Runtime

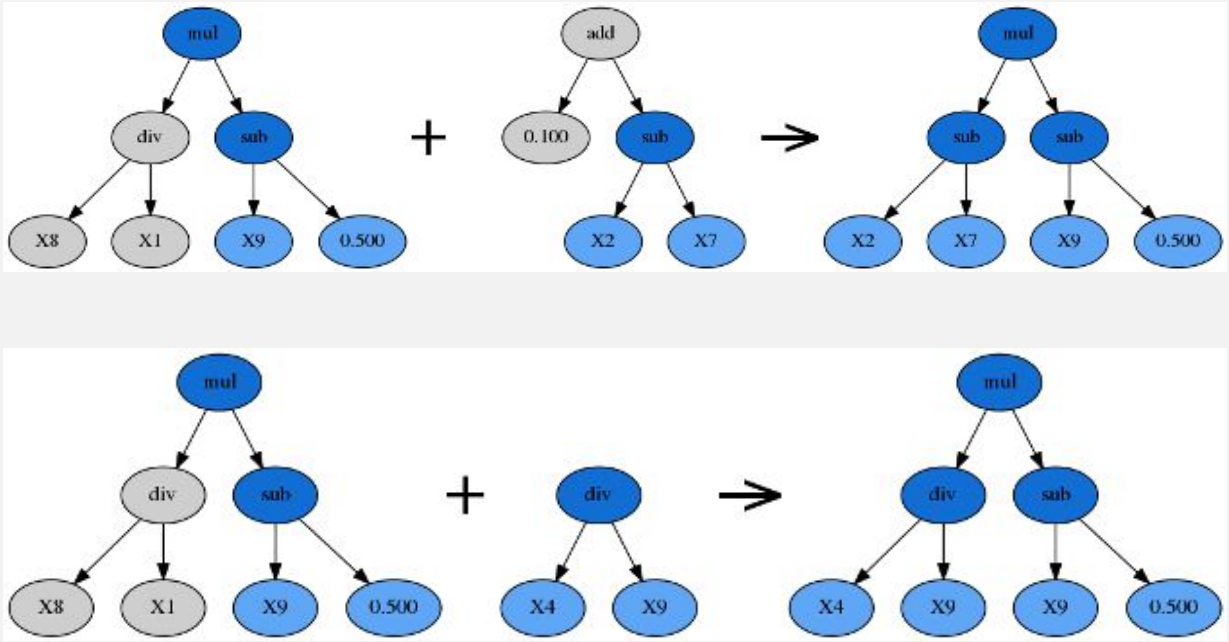
- Stopping Criteria – If the model reaches this threshold it will stop executing
- Generations – The number of generations to run if the stopping criteria is never met
- Population Size – The number of equations created per generation
- Tournament Size – A subset of the population where the fittest equation is selected to move on to the next generation

3/3 Evolution (4-P's)

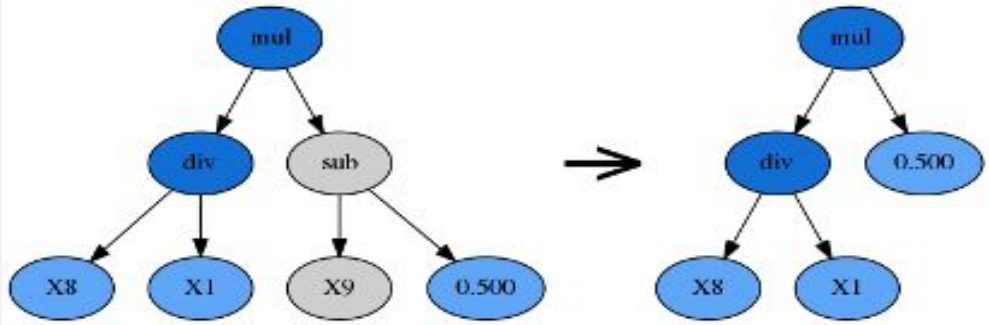
Run on Tournament Winners

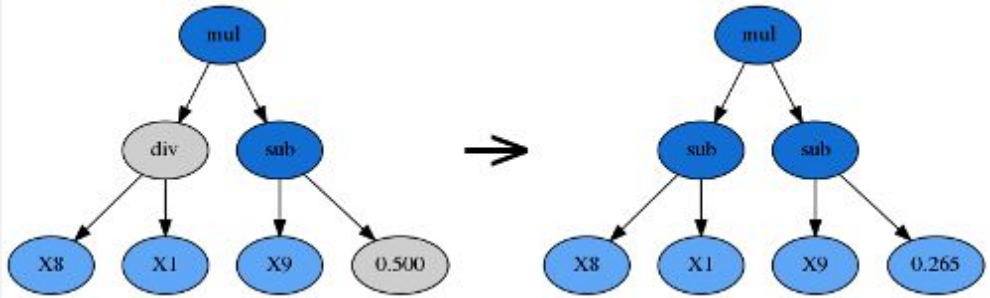
- Crossover – Running two tournaments and replacing a branch from the first with one from the second
- Subtree Mutation – Replaces branches with randomly generated subtrees
- Hoist Mutation – Deletes sections of a subtree to fight bloat
- Point Mutation – Selects random nodes and replaces them with random functions and terminals
- Reproduction – If the above parameters do not sum to 1, the remainder is the probability of the selected winner advancing unchanged to the next generation

Crossover vs Subtree Mutation



Hoist vs Point Mutation





Honorable Mentions

- Function Set –The functions that the regressor uses
- Low Memory – If set to true, will not store previous equations when no longer in use
  - Very helpful for running through Google Colab which had memory issues when we would run long tests
- N_Jobs – Allows for concurrent runs of the fitness function, only helpful with large function sets or input variables
- Warm Start – If true, will continue running where the previous function stopped
  - Didn't get it to work in colab, but would have been very helpful in saving overnight runs that timed out

Conclusion and Future Development

- Another regression
- Another development tool
- Extract immediate equation
- Produce more graph or visual

Thank you.