Data-Driven and Physics-Inspired Machine Learning

UROP Report U081470

Saumya Shah

ABSTRACT

TK

Keywords: Machine Learning, Symbolic Regression, AI Feynman

CONTENTS

1 INTRODUCTION

AI Feynman, a symbolic regression algorithm, is a physics-inspired symbolic regression algorithm developed by Silviu-Marian Udrescu and Max Tegmark in 2020

2 BACKGROUND

TK

3 RELATED WORK

- Silviu-Marian Udrescu and Max Tegmark. AI Feynman: A physics-inspired method for symbolic regression. Technical report, 2020
    - Existing solutions use genetic algorithms or sparse regression.
    - This paper uses NNs to find simplifying factors like symmetry or separability in the dataset.
    - 6 simplifying assumptions/properties are used:
        - *Units:* variables have known physical units
        - *Low-order polynomials:* f is/composed of low-order polynomial(s)
        - *Composition:* f is composed of a small set of elementary functions
        - *Smoothness:* f is continuous
        - *Symmetry:* some variables in f are symmetric
        - *Separability*: f can be separated into sum/product of variables
    - Full algorithm (recursive in nature):

        1. *Dimensional analysis:* Reduces the number of dimensions and simplifies data (makes the model depend on as few variables as possible).
        2. *Polynomial fit:* Polynomial coefficients are calculated by solving a system of linear equations and testing if the RMSE is lower than the threshold.
        3. *Brute force:* Expressions of increasing complexity are generated by brute force and tested till the error drops below a certain threshold.
        4. *NN-based transformations:* Tests for translational symmetry, separability, equality of variables, and other transformations like power, log, etcetera

    - The key improvement of AI Feynman over Eureqa is its ability to decompose the problem into simpler sub-problems with fewer variables.

- Zi-yu Khoo, Abel Yang, Jonathan Sze Choong Low, and Stéphane Bressan. Celestial Machine Learning: From Data to Mars and Beyond with AI Feynman. pages 469–474. 2023

    - Physics-inspired symbolic regression methods like AI Feynman leverage properties of f like symmetry and separability.

    - Four variations of AI Feynman were compared

        * No additional bias
        * Observational bias (replacing angular values with their sines/cosines)
        * Inductive bias (search space restriction)
        * Both observational and inductive bias

    - For experiments 1 and 3, none of the equations on the Pareto frontier matched the orbital equation for Mars.

    - For experiments 2 and 4, 3 out of 9 equations matched.

    - Experiment 4 (combining inductive and observational biases) was best suited to rediscover the orbital equation of Mars.

- Zi-Yu Khoo, Gokul Rajiv, Abel Yang, Jonathan Sze Choong Low, and Stéphane Bressan. Celestial Machine Learning: Discovering the Planarity, Heliocentricity, and Orbital Equation of Mars with AI Feynman. pages 201–207. 2023

    - This paper extends AI Feynman to discover heliocentricity and planarity of Mars' orbit by adding biases.

    - Experimental Setup

        * An inductive bias is built in by restricting the search space to trigonometric, polynomial, and radical functions.
        * An observational bias is embedded by replacing angular values with their sine and cosine.
        * The description length serves as a measure for fit and parsimony.
        * There is a log-scaled penalty on absolute loss (for optimising fit) and on real numbers, operators, and variables (for optimising parsimony)

    - Experiment 1

        * Three sets of observations, corresponding to the three reference frames are created for both coordinate systems (Cartesian and polar) and are used as inputs to AI Feynman.
        * For the Cartesian coordinates, several equations were identified that matched a known equation.

        - For the polar coordinates, none of them matched any known equations. However, one of the equations uses a similar attempt (using angular width) to a known equation.
        - Both coordinate systems preferred heliocentric equations, suggesting a higher parsimony in that reference frame.
    - Experiment 2
        - Principal component analysis is first used to project the data into 3d, 2d, and 1d spaces, which are then used as inputs to AI Feynman.
        - None of the equations match the known equation forms. However, two of the equation use a square root to fit $r(t)$ similar to a known equation.
        - Most equations only use one to two eigenvectors suggesting a planar relationship.
    - Experiment 3
        - Knowledge regarding the heliocentricity and planarity of Mars' orbit are embedded as observational biases.
        - Some of the equations suggest a circular orbit due to low eccentricty.
        - After correcting for a vertical shift of focus, Kepler's first law was obtained from both the Cartesian and polar datasets.

- Pablo Lemos, Niall Jeffrey, Miles Cranmer, Shirley Ho, and Peter Battaglia. Rediscovering orbital mechanics with machine learning. *Machine Learning: Science and Technology*, 4(4):045002, 2023
    - The authors assume Newton's 2nd and 3rd laws as inductive biases and simulate orbital dynamics of solar system bodies using a graph network. Then, they use symbolic regression using an open-source analogue of Eureqa to find an analytic expression for Newton's law of gravitation.
    - Data of 31 bodies were used from the Horizons On-Line Ephemerys System between 1980 and 2013.
    - The procedure was composed of two stages: training a graph-network-based simulator on observed data, and then performing symbolic regression.
    - Graph-Network-Based Simulator
        - The input is a graph where the nodes represent the celestial bodies and relationships between two bodies are represented as edges.
        - Each node contained a trainable scalar V analogous to mass, while each edge stored the spatial displacement vector between the two corresponding bodies.
        - The model computed interactions for each body using a function to calculate the values for the edges with a trainable parameter θ, analogous to a force.
        - θ and v are trained using gradient descent using the error between the true and predicted acceleration.
        - The graph network uses a 3 layer MLP model implemented in Tensorflow.
        - A random three-dimensional rotation was applied to the input graph to prevent biases and to encourage learning rotational equivariance.
    - Symbolic Regression
        - A dataset of force function inputs was created and SR was used to fit an explicit formula for the force function.
        - The PySR library was used, which uses a tree search algorithm to produce a set of candidate equations. This was done by assigning a score as the ratio of the increase in accuracy and the increase in complexity.

  - Results
    - ∗ The simulator acheived a high degree of accuracy for the accceleration, with an error of 0.2% on validation data. However, there was a greater uncertainty of 9.1% for the mass.
    - ∗ The equivalence principle holds that for bodies which have negligible gravitational influence, their masses are difficult to estimate accurately. As such, these bodies had a higher error in their mass.
    - ∗ The experiment was able to discover Newton's law of gravity as well as a value very close to the gravitational constant.
    - ∗ After discovering the form of the interactions between bodies, the edge function was replaced by the discovered expression and retrained to to estimate the mass and gravitational variables, The mean percent error in mass improved to 1.6%.
  - This was only possible due to the inductive biases of Newton's 2nd and 3rd laws.

- Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. 113(15):3932–3937, 2016
  - Symbolic regression using genetic programming is expensive and may be prone to overfitting.
  - Sparsity techniques efficiently identify the relevant terms to reduce the search space. The resulting model thus balances parsimony with accuracy while also avoiding overfitting.
  - SINDy leverages the fact that physical systems only have a few relevant terms to promote sparsity.
  - *Process*:
    1. A time history of the state is collected and its derivative is numerically approximated. Gaussian noise is added, and the data are organised into two matrices.
    2. A library of candidate non-linear functions for the state is created for each column of the state matrix.
    3. A sparse regression problem is set up to determine the vector of coefficients that determine which non-linear terms are in f.
    4. LASSO or other alternative techniques are applied, depending on the context of the problem, to calculate the vector of coefficients.
  - Domain knowledge can be used to help choose appropriate variables and non-linear functions, and to exploit simplifying properties like symmetry.
  - For the Lorenz system (a system with chaotic dynamics), the model accurately reproduces attractor dynamics, correctly identifies the relevant terms, and determines the coefficients to within 0.03% of the true value.
  - The authors further generalise the SINDy method to an example of vortex shedding (fluid dynamics). For the logistic map, parameters are identifed to within 0.1
  - SINDy is robust to measurement noise and unavailability of derivative measurements.
  - Significant challenges remain in the correct choice of measurement coordinates and the choice of sparsifying function basis.

4 METHODOLOGY

TK

5 PERFORMANCE EVALUATION

TK

6 CONCLUSION

TK

REFERENCES

[1] Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. 113(15):3932–3937, 2016.

[2] Zi-Yu Khoo, Gokul Rajiv, Abel Yang, Jonathan Sze Choong Low, and Stéphane Bressan. Celestial Machine Learning: Discovering the Planarity, Heliocentricity, and Orbital Equation of Mars with AI Feynman. pages 201–207. 2023.

[3] Zi-yu Khoo, Abel Yang, Jonathan Sze Choong Low, and Stéphane Bressan. Celestial Machine Learning: From Data to Mars and Beyond with AI Feynman. pages 469–474. 2023.

[4] Pablo Lemos, Niall Jeffrey, Miles Cranmer, Shirley Ho, and Peter Battaglia. Rediscovering orbital mechanics with machine learning. *Machine Learning: Science and Technology*, 4(4):045002, 2023.

[5] Silviu-Marian Udrescu and Max Tegmark. AI Feynman: A physics-inspired method for symbolic regression. Technical report, 2020.

APPENDICES

TK