Homework 4 Example Formatting

This document is provided as an example (mainly so that you know what kind of formatting I'm expecting). The papers used in this example are not necessarily related to classification.

Part A

Paper 1

1. Citation information:

Patryk Orzechowski, William La Cava, and Jason H. Moore. 2018. Where are we now? a large benchmark study of recent symbolic regression methods. In Proceedings of the Genetic and Evolutionary Computation Conference (GECCO '18). Association for Computing Machinery, New York, NY, USA, 1183–1190. https://doi.org/10.1145/3205455.3205539

2. Abstract

In this paper we provide a broad benchmarking of recent genetic programming approaches to symbolic regression in the context of state of the art machine learning approaches. We use a set of nearly 100 regression benchmark problems culled from open source repositories across the web. We conduct a rigorous benchmarking of four recent symbolic regression approaches as well as nine machine learning approaches from scikit-learn. The results suggest that symbolic regression performs strongly compared to state-of-the-art gradient boosting algorithms, although in terms of running times is among the slowest of the available methodologies. We discuss the results in detail and point to future research directions that may allow symbolic regression to gain wider adoption in the machine learning community.

3. How did you find the paper?

Your response describing how you found this paper goes here.

Paper 2

1. Citation information:

William La Cava, Lee Spector, and Kourosh Danai. 2016. Epsilon-Lexicase Selection for Regression. In Proceedings of the Genetic and Evolutionary Computation Conference 2016 (GECCO '16). Association for Computing Machinery, New York, NY, USA, 741–748. https://doi.org/10.1145/2908812.2908898

2. Abstract

Lexicase selection is a parent selection method that considers test cases separately, rather than in aggregate, when performing parent selection. It performs well in discrete error spaces but not on the continuous-valued problems that compose most system identification tasks. In this paper, we develop a new form of lexicase selection for symbolic regression, named epsilon-lexicase selection, that redefines the pass condition for individuals on each test case in a more effective way. We run a series of experiments on real-world and synthetic problems with several treatments of epsilon and quantify how epsilon affects parent selection and model performance. epsilon-lexicase selection is shown to be effective for regression, producing better fit models compared to other techniques such as tournament selection and age-fitness Pareto optimization. We demonstrate that epsilon can be adapted automatically for individual test cases based on the population performance distribution. Our experiments show that epsilon-lexicase selection with automatic epsilon produces the most accurate models across tested problems with negligible computational overhead. We show that behavioral diversity is exceptionally high in lexicase selection treatments, and that epsilon-lexicase selection makes use of more fitness cases when selecting parents than lexicase selection, which helps explain the performance improvement.

3. How did you find the paper?

Your response describing how you found this paper goes here.

Paper 3

1. Citation information:

La Cava, W., Orzechowski, P., Burlacu, B., França, F. O. de, Virgolin, M., Jin, Y., Kommenda, M., & Moore, J. H. (2021). Contemporary Symbolic Regression Methods and their Relative Performance. Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks.

2. Abstract

Many promising approaches to symbolic regression have been presented in recent years, yet progress in the field continues to suffer from a lack of uniform, robust, and transparent benchmarking standards. In this paper, we address this shortcoming by introducing an open-source, reproducible benchmarking platform for symbolic regression. We assess 14 symbolic regression methods and 7 machine learning methods on a set of 252 diverse regression problems. Our assessment includes both real-world datasets with no known model form as well as ground-truth benchmark problems, including physics equations and systems of ordinary differential equations. For the real-world datasets, we benchmark the ability of each method to learn models with low error and low complexity relative to state-of-the-art machine learning methods. For the synthetic problems, we assess each method's ability to find exact solutions in the presence of varying levels of noise. Under these controlled experiments, we conclude that the best performing methods for real-world regression combine genetic algorithms with parameter estimation and/or semantic search drivers. When tasked with recovering exact equations in the presence of noise, we find that deep learning and genetic algorithm-based approaches perform similarly. We provide a detailed guide to reproducing this experiment and contributing new methods, and encourage other researchers to collaborate with us on a common and living symbolic regression benchmark.

3. How did you find the paper?

Your response describing how you found this paper goes here.

Part B

I chose to read [insert your chosen paper title here].

1. Why did you choose to read this paper?

Your response here.

2. In your own words, what is the paper about?

Your response here.

3. To the best of your understanding, what classification algorithms are used in the paper?

Your response here.