Topicos em Inteligência Artificial: Aprendizado de Máquina para Ciências com Regressão Simbólica

November 2023

1 Sobre o professor

- **Professor:** Fabrício Olivetti de França
- **E-mail:** folivetti@ufabc.edu.br
- **Site:** https://folivetti.github.io/

2 Objetivos

Aprendizado de Máquina para Ciências (ou em inglês, *Scientific Machine Learning* - SciML) é uma área emergente que tem por objetivo principal a busca por um modelo preditivo que seja aderente a certas características de interesse e forneça informações relevantes para o fenômeno estudado. Essa área reduz o foco do modelo criado apenas para predição e estimula o uso do modelo como parte de uma investigação científica para a extração de conhecimento. A Regressão Simbólica é conhecida como um modelo de regressão não-linear cujas características permitem a criação de modelos interpretáveis, incorporando conhecimento de domínio, e que permitem a análise de incertezas por padrão. Esse curso tem o objetivo principal de ensinar sobre os conceitos principais de Regressão Simbólica contextualizados em SciML e como utilizar tal ferramenta para extração de conhecimento de dados diversos.

3 Metodologia

Aulas expositivas dos conteúdos abordados pelo professor com demonstrações do uso das ferramentas de regressão. A avaliação da disciplina será feita através da escrita de um artigo científico aplicando os conceitos da disciplina em uma base de dados de interesse.

4 Planejamento Semanal

4.1 Semana 01

1. Introdução à disciplina
2. Conceitos básicos de aprendizado de máquina

4.2 Semana 02

1. Feriado de carnaval

4.3 Semana 03

1. Análise de Regressão e Modelos Paramétricos
2. Regressão Simbólica

4.4 Semana 04

1. Regressão simbólica: Programação Genética
2. Regressão Simbólica: outras abordagens

4.5 Semana 05

1. Ferramentas de regressão simbólica: SRBench, srtree-opt
2. Gráficos de avaliação de modelo

4.6 Semana 06

1. Função distribuição e verossimilhança
2. Exemplos de dados de diferentes distribuições

4.7 Semana 07

1. Otimização não-linear dos coeficientes
2. Implementando um algoritmo de otimização

4.8 Semana 08

1. Validação de Modelo
2. Seleção de Modelo

4.9 Semana 09

1. Simplificação de Modelo
2. Integrando conhecido pré-existente

4.10 Semana 10

1. Feriado (08 de abril)
2. Extraindo informações do modelo

4.11 Semana 11

1. Incertezas: intervalos de confiança e predição
2. Incertezas: profile likelihood

4.12 Semana 12

1. Apresentação dos resultados dos artigos
2. Apresentação dos resultados dos artigos

4.13 Semana 13 (reposição)

1. Considerações finais (30 de abril - terça-feira)
2. Revisão final dos artigos (03 de maio - sexta-feira)

5 Referências Bibliográficas

References

[1] Guilherme Seidyo Imai Aldeia and Fabrício Olivetti de França. "Measuring feature importance of symbolic regression models using partial effects". In: *Proceedings of the Genetic and Evolutionary Computation Conference*. 2021, pp. 750–758.

[2] Douglas Bates. "Nonlinear regression analysis and its applications". In: *Wiley Series in Probability and Statistics* (1988).

[3] Brian C Falkenhainer and Ryszard S Michalski. "Integrating quantitative and qualitative discovery: the ABACUS system". In: *Machine Learning* 1 (1986), pp. 367–401.

[4] Fabricio Olivetti de Franca and Gabriel Kronberger. "Prediction Intervals and Confidence Regions for Symbolic Regression Models based on Likelihood Profiles". In: *arXiv preprint arXiv:2209.06454* (2022).

[5] Fabricio Olivetti de Franca and Gabriel Kronberger. "Reducing Overparameterization of Symbolic Regression Models with Equality Saturation". In: *Proceedings of the Genetic and Evolutionary Computation Conference*. 2023, pp. 1064–1072.

[6] Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.

[7] Andrew Gelman, Jennifer Hill, and Aki Vehtari. *Regression and other stories*. Cambridge University Press, 2020.

[8] Donald Gerwin. "Information processing, data inferences, and scientific generalization". In: *Behavioral Science* 19.5 (1974), pp. 314–325.

[9] Gareth James et al. *An introduction to statistical learning*. Vol. 112. Springer, 2013.

[10] John R Koza. *Genetic programming: A paradigm for genetically breeding populations of computer programs to solve problems*. Vol. 34. Stanford University, Department of Computer Science Stanford, CA, 1990.

[11] Gabriel Kronberger et al. "Shape-constrained symbolic regression—improving extrapolation with prior knowledge". In: *Evolutionary computation* 30.1 (2022), pp. 75–98.

[12] William La Cava et al. "Contemporary symbolic regression methods and their relative performance". In: *arXiv preprint arXiv:2107.14351* (2021).

[13] Pat Langley. "Data-driven discovery of physical laws". In: *Cognitive Science* 5.1 (1981), pp. 31–54.

[14] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[15] Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 1999.

[16] Patryk Orzechowski, William La Cava, and Jason H Moore. "Where are we now? A large benchmark study of recent symbolic regression methods". In: *Proceedings of the Genetic and Evolutionary Computation Conference*. 2018, pp. 1183–1190.

[17] Paul Roback and Julie Legler. *Beyond multiple linear regression: applied generalized linear models and multilevel models in R*. CRC Press, 2021.

[18] Max Willsey et al. "Egg: Fast and extensible equality saturation". In: *Proceedings of the ACM on Programming Languages* 5.POPL (2021), pp. 1–29.