Contribution ID: **45** Type: **not specified**

Explanation Groves –Analyzing the Trade-off between Complexity and Adequacy of an Explanation

Wednesday, 15 May 2024 15:50 (20 minutes)

The increasing popularity of machine learning in many application fields has led to an increasing demand in methods of explainable machine learning as they are e.g. provided by the R packages DALEX (Biecek, 2018) and iml (Molnar, 2018). A general process to ensure the development of transparent and auditable machine learning models in industry (TAX4CS) is given in Bücker et al. (2021).

In turn, comparatively few research has been dedicated to the limits of explaining complex machine learning models (cf. e.g. Rudin, 2019, Szepannek and Lübke, 2023). In the presentation, explanation groves (Szepannek and von Holt, 2024) will be introduced. Explanation groves extract a set of understandable rules in order to explain arbitrary machine learning models. In addition, the degree of complexity of the resulting explanation can be defined by the user. In consequence, they provide a useful tool to analyze the trade off between the complexity of a given explanation on one hand and how well it represents the original model on the other hand.

After presenting the method some results on real world data will be shown. A corresponding R package xgrove is available on CRAN (Szepannek, 2023) and will be briefly demonstrated.

Biecek P (2018). DALEX: Explainers for Complex Predictive Models in R. Journal of Machine Learning Research, 19(84), 1-5. https://jmlr.org/papers/v19/18-416.html.

Bücker, M.; Szepannek, G., Gosiewska, A. Biecek, P. (2021): Transparency, Auditability and eXplainability of Machine Learning Models in Credit Scoring, Journal of the Operation Research Society, DOI: 10.1080/01605682.2021.1922098.

Molnar C, Bischl B, Casalicchio G (2018). "iml: An R package for Interpretable Machine Learning."JOSS, 3(26), 786. DOI:10.21105/joss.00786.

Rudin, C (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1, 206–215. DOI:10.1038/s42256-019-0048-x

Szepannek G (2023). xgrove: Explanation Groves. R package version 0-1-7. https://CRAN.R-project.org/package=xgrove.

Szepannek, G., von Holt, B.-H. (2024): Can't See the Forest for the Trees –Analyzing Groves for Random Forest Explanation, Behaviormetrika, DOI: 10.1007/s41237-023-00205-2.

Szepannek, G., Luebke, K. (2022): Explaining Artificial Intelligence with Care – Analyzing the Explainability of Black Box Multiclass Machine Learning Models in Forensics, Künstliche Intelligenz, DOI : 10.1007/s13218-022-00764-8.

Type of presentation

Contributed Talk

Primary author: SZEPANNEK, Gero (Stralsund University of Applied Sciences)

Presenter: SZEPANNEK, Gero (Stralsund University of Applied Sciences)

Session Classification: Contributed session

Track Classification: Spring Meeting