Research Working Papers

Understanding Models and Model Bias with Gaussian Processes

by: Thomas R. Cook and Nathan M. Palmer

June 15, 2023

Using counterfactual reasoning and Gaussian process models can help detect bias in machine learning models.

RWP 23-07, June 2023

Despite growing interest in the use of complex models, such as machine learning (ML) models, for credit underwriting, ML models are difficult to interpret, and it is possible for them to learn relationships that yield de facto discrimination. How can we understand the behavior and potential biases of these models, especially if our access to the underlying model is limited? We argue that counterfactual reasoning is ideal for interpreting model behavior, and that Gaussian processes (GP) can provide approximate counterfactual reasoning while also incorporating uncertainty in the underlying model's functional form. We illustrate with an exercise in which a simulated lender uses a biased machine model to decide credit terms. Comparing aggregate outcomes does not clearly reveal bias, but with a GP model we can estimate individual counterfactual outcomes. This approach can detect the bias in the lending model even when only a relatively small sample is available. To demonstrate the value of this approach for the more general task of model interpretability, we also show how the GP model's estimates can be aggregated to recreate the partial density functions for the lending model.

JEL classifications: C10, C14, C18, C45

Article Citations

- Cook, Thomas R., and Nathan M. Palmer. 2023. "Understanding Models and Model Bias with Gaussian Processes." Federal Reserve Bank of Kansas City, Research Working Paper no. 23-07, June. Available at https://doi.org/10.18651/RWP2023-07

Related Research

- Bertrand, M., and E. Duflo. 2017. "Field Experiments on Discrimination." *Handbook of Economic Field Experiments*, vol. 1, pp. 309–393. Available at https://doi.org/10.1016/bs.hefe.2016.08.004

- Cook, Thomas R., Greg Gupton, Zach Modig, and Nathan M. Palmer. 2021. "Explaining Machine Learning by Bootstrapping Partial Dependence Functions and Shapley Values." Federal Reserve Bank of Kansas City, Research Working Paper no. 21-12, November. Available at https://doi.org/10.18651/RWP2021-12
- DiNardo, John, Nicole M. Fortin, and Thomas Lemieux. 1995. "Labor Market Institutions and the Distribution of Wages, 1973–1992: A Semiparametric Approach." National Bureau of Economic Research, working paper no. 5093, April. Available at https://doi.org/10.3386/w5093
- Rasmussen, Carl Edward, and Christopher K. I. Williams. 2006. *Gaussian Processes for Machine Learning, Volume 2*. Cambridge, MA: MIT Press.

Author

Thomas R. Cook

Data Scientist

Tom Cook is a Data Scientist in the Economic Research Department of the Federal Reserve Bank of Kansas City. He joined the bank in August 2016 after completing his PhD in Political Science at the University of Colorado. He also holds an MPA from DePaul University, and a BA in Philosophy and Political Science from the University of Iowa. His substantive research interests are the roles of time and information transmission in political and economic strategic behavior. Methodologically, his research at the bank focuses on the development of machine learning, neural networks, and advanced statistical models for use in economic research.