On evaluating counterfactual explanations for ML models in distributed systems

Elianna Douka
elidouka@di.uoa.gr
National and Kapodistrian University of Athens
Athens, Greece

Dimitrios Tomaras
tomaras@aueb.gr
Athens University of Economics and Business
Athens, Greece

Dimitrios Gunopulos
dg@di.uoa.gr
National and Kapodistrian University of Athens
Athens, Greece

ABSTRACT

The imminent need to interpret the output of a Machine Learning model with counterfactual (CF) explanations – via small perturbations to the input – has been notable in the research community. However, it is quite important to evaluate the performance of such explanations, when such models are being trained in a distributed setting. This work transfers the knowledge from interactive systems and proposes metrics such that CF examples can be evaluated and be useful for unbiased model training. Our preliminary results illustrate the significance and the benefits of our approach.

ACM Reference Format:
Elianna Douka, Dimitrios Tomaras, and Dimitrios Gunopulos. 2024. On evaluating counterfactual explanations for ML models in distributed systems. In . ACM, New York, NY, USA, 1 page.

1 INTRODUCTION

Training Machine Learning (ML) models in distributed systems can affect their performance can bias their output. Pre-processing data transformations may inherently impact models being learnt in a distributed setting and this phenomenon is exacerbated in such systems. CF explanations [2] have emerged as a means of analyzing and interpreting ML models, as it is crucial to understand model behavior, without running disruptive live tests on the models. However, this emerges the risk of inaccuracies in model estimations due to biases or incorrect assumptions made by the practitioner who examines the model. A crucial question to answer is how to evaluate counterfactual explanations for models trained using a distributed setting. Existing works have focused on policy learning e.g. the relevance of a targeted ad to a user, which has been used to estimate and evaluate the online performance of a given policy or even find and learn a new policy that improves upon an existing policy. The question we aim to answer in this work is the ability to transfer this knowledge from the policy domain to the CF explanations domain.

2 RESEARCH PROBLEM

We aim to answer how to evaluate counterfactual explanations for models trained in distributed systems. Several metrics have been proposed such as Inverse Propensity Score[1] and Doubly Robust. We focus on the IPS metric on a novel type of advertising data, as it can be used to estimate the effects of treatments or actions in observational studies where random assignment is not possible, mitigating selection bias. Propensity score is the probability of assignment to a particular treatment given a set of observed covariates. Inverse Propensity Scoring weights each instance by the inverse of its propensity score. IPS adjusts the importance of each action's reward based on how likely it was to be chosen by the behavior policy, aiming to estimate what the reward would have been under a different target policy. Therefore, it can help reduce bias in estimating treatment effects from observational data and allows for the comparison of outcomes under different treatment strategies without needing randomized trials. Our preliminary results evaluate and explain bias in advertising data have illustrated the following: In individuals with high propensity scores (more likely to click), we might expect higher weights for treated individuals and lower weights for untreated individuals. Thus, in the original data the weighting process has relatively evenly distributed influence across the range of propensity scores (Figure 1). However, in the counterfactual dataset, we have an indication that a few individuals with high propensity scores have extremely large weights compared to the rest of the population and therefore the training procedure over the distributed setting requires to mitigate the bias.

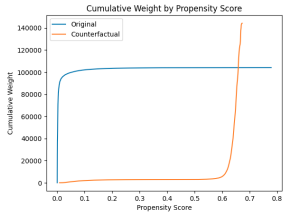

Figure 1: Cumulative weight propensity score

ACKNOWLEDGMENTS

This research has been financed by the European Union through the EU ICT-48 2020 project TAILOR (No. 952215), the H2020 AutoFair project (No. 101070568) & the Horizon Europe CoDiet project (No. 101084642).

REFERENCES

[1] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983), 41–55.
[2] Sandra Wachter and et al. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference'17, July 2017, Washington, DC, USA
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM