Causal Machine Learning in Healthcare

Anonymous Author(s)
Affiliation
Address
email

Abstract

There is significant interest in discovering causal relationships to optimize the treatment of patients in healthcare. In the era of big data, techniques from causal machine learning such as interpretable models and specialized model architectures increasingly support domain experts in the process of exploring these relationships.

1 Introduction

Causal inference is core to medicine. In this setting, we generally have some covariates (e.g., age, gender, images) about a patient and want to answer the counterfactual [15] question: Which treatment would lead to the best outcome? The state of the art approach to answer this question are randomised controlled trials (RCTs) [9] in which patients are assigned to the intervention or the comparator group at random. If the sample size is large enough, the act of randomization ensures that potential confounders (measured and unmeasured) are balanced between the groups which allows the attribution of differences in the outcome to the intervention. However, researchers face several challenges when conducting these trials:

- **Representativeness/Generalisation**: The study is only applicable to large groups in the real world if the original population of the trial is representative. There may be biases in the population (e.g., because of the recruitment process) that decrease generalization.
- **Costs/Resources**: RCTs are very expensive and require experts and manual labor. In 2013, the average per-patient costs were estimated at $36,500 per trial phase and developing a new medicine required an investment of around $2.6 billion [3].
- **Multiple Treatments**: We often want to compare more than one treatment.
- **Measuring Outcomes**: For some diseases, measuring the outcome can be hard, for instance because the effects are only observed after some years.
- **Ethical Issues**: Not treating a patient can be unethical in some settings.

A second central application of causal inference in healthcare is the discovery of interventions that could be used as new treatment options. Currently, this is mainly done with experiments that are analyzed and visualized, leading to new insights and experiments to further refine a hypothesis. The problem with this approach is that it is manual and largely driven by domain experts: Someone needs to come up with good hypotheses, prioritise them, design the experiments, potentially merge the evidence with other experiments, and decide if the results are representative.

A key challenge when applying causal inference techniques to healthcare is dealing with complexity. The causal generative process of the human body is very sophisticated and the causal relations span multiple scales of resolution, from reactions at the molecular level to symptoms of the body as a whole. Furthermore, because of the previously mentioned challenges with RCTs, addressing these problems by collecting more data is often not feasible.

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

2 Estimating Causal Treatment Effects from Observations

2.1 Framework

We are in the Rubin-Neyman Potential Outcomes Framework [18] where the counterfactual outcomes $Y = [y_0 \dots y_k]^T$ are the outcomes that are (or would be) observed after applying one of k treatments $t_0, \dots, t_k$. We use t to denote which treatment is assigned to an individual. The population consists of N cases with pre-treatment covariates X and we are usually interested in estimating:

- Average Treatment Effect:

$$ATE_{i,j} = \mathbb{E}[y_{t_j} - y_{t_i}] = \sum_{k=1}^{N} \left(y_{t_j}(k) - y_{t_i}(k)\right)$$

- Individual Treatment Effect/Conditional Average Treatment Effect:

$$ITE_{i,j} = \mathbb{E}[y_{t_j} - y_{t_i} \mid X]$$

2.2 Quasi-Experimental Studies

In quasi-experimental studies, we try to infer causal effects from non-randomised experiments. While we can control for observed confounding, we cannot do so for hidden (unmeasured) confounders. For that reason, the degree of evidence for causal effects is generally lower than in RCTs.

One type of quasi-experimental studies are case-control studies. The outcomes across two groups are compared based on a potential causal factor and we control for observed confounding by matching cases with similar controls. Matching by comparing the covariates can be infeasible because X is high-dimensional in many settings, so a balancing score $b(X)$ is often used in practice. The treatment effects can only be identified if certain assumptions hold [11, 17, 19]:

- **Conditional Independence Assumption**: $Y \perp\!\!\!\perp t \mid b(X)$ (with the special case $b(X) = X$), meaning that the assignment of the treatment is independent of the outcome, given the balancing score.
- **Common Support Assumption**: $0 < P(t = 1 \mid X) < 1 \, \forall X$, i.e. every unit has a chance of receiving each treatment.
- **Stable Unit Treatment Value Assumption (SUTVA)**: The values of all outcomes Y are not affected by any t (note that which value we observe in the study is obviously affected by t, but the statement is about the whole vector which is partially unobserved), which implies that there is no interference between units.

These assumptions are generally untestable [28], but Pearl introduced a simple graphical test that can be applied to the causal graph (which we need to construct with domain knowledge) for testing if a set of variables is sufficient for identification [14].

2.3 Counterfactual Regression

Given the observational data, we want to train a counterfactual estimator that allows us to predict (in Pearl's *do*-notation [15]):

$$f(X, t) = p(Y \mid X, do(t = T))$$

One approach is to learn individual models for the different treatments (which can result in asymptotically consistent/unbiased estimates, e.g. using the Double/Debiased Machine Learning approach introduced by Chernozukov et al. [4]), but this introduces additional variance because the control and treated distributions (i.e. $p(x \mid t = 0)$ and $p(x \mid t = 1)$) usually differ. Shalit et al. [25] upper bound this source of variance using an Integral Probability Metric (IPM) between the two distributions. Based on this bound, they introduce the Counterfactual Regression (CFR) and Treatment-Agnostic Representation Network (TARNet) models (with the difference that TARNet ignores the IPM term when calculating the loss) which consists of shared base layers (learning non-linear representations of the input data) and two separate "heads" to estimate the outcome under treatment/control. The goal of these networks is to minimize the factual loss and the IPM distance at the same time.

Schwab et al. [22] extend TARNet to the multiple treatment setting with k head networks. Furthermore, they introduce the mini-batch augmentation method Perfect Match that imputes the unobserved counterfactual outcomes by the outcomes of the nearest neighbors (using a balancing score to measure distances). This approach constructs virtually randomised minibatches that approximate a randomised experiment.

Dose-Response Networks [24] are a further extension of the described model architecture where the range of dosages is discretized into buckets and a separate head layer is used for every bucket. The number of buckets allows to tradeoff predictive performance and computational requirements.

For the evaluation of counterfactual regression models that estimate the ITE, the precision in estimating heterogenous effects (PEHE) is often used, defined as (for binary treatments) [10]:

$$\epsilon_{\text{PEHE}} = \frac{1}{N}\sum_{k=1}^{N}\left(\mathbb{E}_{y_j(k)\sim\mu_j(k)}\left[y_1(k) - y_0(k)\right] - \mathbb{E}\left[f(X^{(k)},1) - f(X^{(k)},0)\right]\right)^2$$

Where μ_0 and μ_1 are the underlying outcome distributions, which are generally not known. There are different techniques to estimate the PEHE, such as data simulation or substituting the expectation by the outcomes of a similar individual according to a distance such as the Mahalanobis distance [20].

3 Causal Explanation Models

We are often not only interested in the prediction of a model, but we also want to know which inputs caused this prediction (i.e., calculate feature importance scores for the different inputs). This is especially important in healthcare because the interpretation of the output and the further steps that are taken can depend a lot on the contributing factors (in settings where humans and machine learning algorithms cooperate). Furthermore, it can generally be beneficial for model debugging as it allows to reason about the discovered patterns and judge their reasonableness.

3.1 Attentive Mixture of Experts Model

One approach is to train machine-learning models that learn to jointly produce accurate predictions and estimations of the feature importance, for instance attentive mixture of experts (AME) models [23]. The basic idea is to distribute the features among experts (neural networks with their own parameters/architectures, outputting their topmost feature representation h_i and their contribution c_i for a given sample) and use attentive gating networks (one per expert) for assigning weights to the experts. The individual attentive gating networks take the feature representation and contribution of every expert as input (i.e., $(h_1, c_1, \ldots, h_p, c_p)$ for p experts) and output an attention factor a_i. Because the features are split across experts, there is no information leakage across them and the network can only increase the contribution of a feature by increasing the expert's attention factor. However, there is generally no guarantee that weights accurately represent feature importance [29] and the networks may collapse towards a minima where very few or only one expert is used [2, 27].

Schwab et al. address this problem by introducing an objective function that measures the mean Granger-causal error (MGE). In the Granger-causality framework, X causes[1] Y if the prediction of Y is better when using all available information instead of all available information except X [7]. Based on that definition, the (normalized) decrease in error associated with adding an expert's information is measured and the Granger-causal objective is the Kullback-Leibler divergence between this decrease and the models attention factors a_i. With this additional objective function, there is incentive (tuneable with a hyperparameter which controls the contribution of the Granger-causal objective) for the network to learn attention factors that correspond to the Granger-causal attributions.

3.2 Comparison

An alternative approach for feature importance estimation is to model the impact of local pertubations on the prediction [1]. The LIME (Local Interpretable Model-agnostic Explanations) algorithm

[1] Note that the term causality may be misleading in this context. Because of that, some researchers use the term "predictive causality", meaning a variable contains useful information for predicting another [5]. Granger himself later used the word "temporal relation" instead of causality [8].

does this by sampling in a local region and fitting an interpretable model (e.g., a sparse linear model) to these samples, which can help understanding and validating the corresponding prediction [16]. With multiple LIME explanations, the model as a whole can be examined. SHAP (SHapley Additive exPlanations) calculates the local feature importance using Shapley values [26], the marginal contribution towards the reduction in prediction error [13]. While both of these approaches are model-agnostic, their sampling-based nature is computationally demanding. AME shows similar estimation accuracy for the feature importances with significantly lower computational requirements. Furthermore, the associations identified by the AME model with a properly tuned MGE/MSE tradeoff were consistent with those reported by domain experts, which was not the case for the other evaluated models.

However, there are some limitations to AME models. The model structure is fixed, which can result in worse predictive performance for certain tasks. Moreover, as the MSE/MGE is jointly optimized, the MSE generally increases when more importance is given to the MGE, meaning there is a tradeoff between predictive performance and accurate importance estimation. Furthermore, the direction of the influence (positive or negative) is not inferred and with many features (and therefore experts, if a one-to-one mapping is used), the optimization can become intractable.

3.3 CXPlain Model

CXPlain addresses the issue of the fixed model structure and increasing MSE that arises when using AME by training a separate explanation model and allowing arbitrary predictive models [21]. The explanation model treats the predictive models as blackboxes and calculates its outputs with and without each input feature. Note that a different strategy for obtaining the predictions without a feature is needed than in AME models as the predictive model now is arbitrary and cannot be modified. This can be accomplished by masking the feature with zeroes, replacing it with the mean value, or using more sophisticated masking schemes. Given these outputs, the (normalized) decrease in error is calculated for every input feature and the Kullback-Leibler divergence between this decrease and the models importance scores a_i is used as the objective function like in AME models. However, this objective function is now optimized individually, the task of producing feature importance estimates is therefore transformed into a supervised learning task with a Granger-causal objective function.

Because some feature importance estimates may themselves be very unreliable [30], CXPlain additionally provides uncertainty estimates for each feature importance estimate. It uses bootstrap resampling for that, i.e. training the explanation model on different subsets of the data (possibly containing duplicates) and using the importance scores of the runs to construct confidence intervals.

As AME, CXPlain provided more accurate feature importance estimates than LIME and SHAP, while being model-agnostic and still computationally efficient. Even though the approach works with arbitrary models, the accuracy of the estimates does depend on the predictive model and some model architectures seem to be better suited for explanation models.

4 Discussion

Although the problem statement of causal machine learning in healthcare is conceptually similar to the problems that were addressed by other researchers in the seminar series, the complexity seems to be much higher. Many of the other applications we have seen were evaluated on datasets with a few factors of variation and a relatively simple causal graph, such as robotics [6] or abstract reasoning on non-convoluted images [12]. Because of the high complexity in the healthcare domain with very complicated relations and many causal factors, the current approaches seem to follow the pragmatic, task-solving based approach that was also mentioned by Francesco Locatello: Instead of trying to infer all of the causal relationships (which may be very hard or even impossible to do for humans in healthcare), it seems like the goal is often to find useful, potentially causal relations that are helpful for solving tasks (which often involve humans).

In my opinion, it will be very interesting to see if we ever achieve a point where we are able to autonomously infer the causal graph in domains with such a high complexity and have enough confidence in the estimate to act upon it without human involvement. This would open up completely new possibilities such as cheap, personalized medicine and treatment procedures.

References

[1] Philip Adler et al. "Auditing Black-Box Models for Indirect Influence". en. In: *Knowledge and Information Systems* 54.1 (Jan. 2018), pp. 95–122. ISSN: 0219-3116. DOI: 10.1007/s10115-017-1116-3.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural Machine Translation by Jointly Learning to Align and Translate". In: *ICLR* (2015).

[3] "Biopharmaceutical Industry-Sponsored Clinical Trials: Impact on State Economies". In: *Pharmaceutical Research and Manufacturers of America* (Mar. 2015).

[4] Victor Chernozhukov et al. "Double/Debiased Machine Learning for Treatment and Structural Parameters". en. In: *The Econometrics Journal* 21.1 (2018), pp. C1–C68. ISSN: 1368-423X. DOI: 10.1111/ectj.12097.

[5] Francis X. Diebold. *Elements of Forecasting*. en. Thomson/South-Western, 2007. ISBN: 978-0-324-35904-6.

[6] Muhammad Waleed Gondal et al. "On the Transfer of Inductive Bias from Simulation to the Real World: A New Disentanglement Dataset". In: *arXiv:1906.03292 [cs, stat]* (Nov. 2019). arXiv: 1906.03292 [cs, stat].

[7] C. W. J. Granger. "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods". In: *Econometrica* 37.3 (1969), pp. 424–438. ISSN: 0012-9682. DOI: 10.2307/1912791.

[8] Clive Granger and Paul Newbold. *Forecasting Economic Time Series*. Elsevier Monographs. Elsevier, 1986.

[9] Eduardo Hariton and Joseph J. Locascio. "Randomised Controlled Trials—the Gold Standard for Effectiveness Research". In: *BJOG : an international journal of obstetrics and gynaecology* 125.13 (Dec. 2018), p. 1716. ISSN: 1470-0328. DOI: 10.1111/1471-0528.15199.

[10] Jennifer L. Hill. "Bayesian Nonparametric Modeling for Causal Inference". In: *Journal of Computational and Graphical Statistics* 20.1 (Jan. 2011), pp. 217–240. ISSN: 1061-8600. DOI: 10.1198/jcgs.2010.08162.

[11] Michael Lechner. "Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption". en. In: *Econometric Evaluation of Labour Market Policies*. Ed. by Michael Lechner and Friedhelm Pfeiffer. ZEW Economic Studies. Heidelberg: Physica-Verlag HD, 2001, pp. 43–58. ISBN: 978-3-642-57615-7. DOI: 10.1007/978-3-642-57615-7_3.

[12] Francesco Locatello et al. "Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations". In: *International Conference on Machine Learning*. 2019, pp. 4114–4124.

[13] Scott M. Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". en. In: *Advances in Neural Information Processing Systems* 30 (2017), pp. 4765–4774.

[14] Judea Pearl. "[Bayesian Analysis in Expert Systems]: Comment: Graphical Models, Causality and Intervention". In: *Statistical Science* 8.3 (1993), pp. 266–269. ISSN: 0883-4237.

[15] Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. 1st. USA: Basic Books, Inc., 2018. ISBN: 978-0-465-09760-9.

[16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. New York, NY, USA: Association for Computing Machinery, Aug. 2016, pp. 1135–1144. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939778.

[17] Paul R. Rosenbaum and Donald B. Rubin. "The Central Role of the Propensity Score in Observational Studies for Causal Effects". en. In: *Biometrika* 70.1 (Apr. 1983), pp. 41–55. ISSN: 0006-3444. DOI: 10.1093/biomet/70.1.41.

[18] Donald B. Rubin. "Causal Inference Using Potential Outcomes". In: *Journal of the American Statistical Association* 100.469 (Mar. 2005), pp. 322–331. ISSN: 0162-1459. DOI: 10.1198/016214504000001880.

[19] "Practical Implications of Modes of Statistical Inference for Causal Effects and the Critical Role of the Assignment Mechanism". In: *Matched Sampling for Causal Effects*. Ed. by Donald B. Rubin. Cambridge: Cambridge University Press, 2006, pp. 402–425. ISBN: 978-0-521-67436-2. DOI: 10.1017/CBO9780511810725.033.

[20] Alejandro Schuler et al. “A Comparison of Methods for Model Selection When Estimating Individual Treatment Effects”. In: *arXiv:1804.05146 [cs, stat]* (June 2018). arXiv: 1804.05146 [cs, stat].

[21] Patrick Schwab and Walter Karlen. “CXPlain: Causal Explanations for Model Interpretation under Uncertainty”. en. In: *Advances in Neural Information Processing Systems* 32 (2019), pp. 10220–10230.

[22] Patrick Schwab, Lorenz Linhardt, and Walter Karlen. “Perfect Match: A Simple Method for Learning Representations For Counterfactual Inference With Neural Networks”. In: *arXiv:1810.00656 [cs, stat]* (May 2019). arXiv: 1810.00656 [cs, stat].

[23] Patrick Schwab, Djordje Miladinovic, and Walter Karlen. “Granger-Causal Attentive Mixtures of Experts: Learning Important Features with Neural Networks”. In: *arXiv e-prints* 1802 (Feb. 2018), arXiv:1802.02195.

[24] Patrick Schwab et al. “Learning Counterfactual Representations for Estimating Individual Dose-Response Curves”. In: *AAAI Conference on Artificial Intelligence*. 2020.

[25] Uri Shalit, Fredrik D. Johansson, and David Sontag. “Estimating Individual Treatment Effect: Generalization Bounds and Algorithms”. en. In: *International Conference on Machine Learning*. PMLR, July 2017, pp. 3076–3085.

[26] Lloyd S. Shapley. “A Value for N-Person Games”. In: *Contributions to the Theory of Games (AM-28)*. Vol. 2. 1953, pp. 307–317.

[27] Noam Shazeer et al. “Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer”. In: *arXiv:1701.06538 [cs, stat]* (Jan. 2017). arXiv: 1701.06538 [cs, stat].

[28] Richard Stone. “The Assumptions on Which Causal Inferences Rest”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 55.2 (1993), pp. 455–466. ISSN: 0035-9246.

[29] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. “Axiomatic Attribution for Deep Networks”. In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML’17. Sydney, NSW, Australia: JMLR.org, Aug. 2017, pp. 3319–3328.

[30] Yujia Zhang et al. “"Why Should You Trust My Explanation?" Understanding Uncertainty in LIME Explanations”. In: *arXiv:1904.12991 [cs, stat]* (June 2019). arXiv: 1904.12991 [cs, stat].