Causal Representation Learning in Healthcare Summary

Zixin Shu
Department of Computer Science
ETH Zürich
Switzerland
zixshu@student.ethz.ch

Jannik Gut
Department of Computer Science
ETH Zürich
Switzerland
jgut@student.ethz.ch

Abstract

This seminar explains the basic concepts of causal inference such as Pearl's causal ladder, counterfactual outcomes and treatment effects and the motivation of applying causal representation learning in healthcare. Also, the reason for using the case-control study and its assumptions are presented. Recent works on counterfactual representation learning (TARNet and DRNet) and causal explanation models(AME network) are introduced and discussed as well.

1 Introduction and motivation

Causal inference is a process of identifying and understanding the relationship between the cause and its effects. In many cases, we encounter the problem of choosing among various medical treatments for a given patient. Causal inference can help to choose among these treatments. Randomised controlled trials are often used to answer the question "Which treatment would lead to the best outcome"[1]. It is an experiment design aiming to reduce the certain source of bias when testing the effectiveness of new treatments by comparing the outcome of two groups of patients(treated and untreated control). Randomisation ensures potential confounders, including both measured and unmeasured, are independently identically distributed across the two groups.

However, there are many questions, which need to be solved, when using this approach. For example, which patient to include/exclude, and if the selected groups are representative. The people selected in the trial may not be the ones that are treated in reality. Additionally, how to generalise the result from the experiment is also a question that needs to be answered. In practice, besides choosing between developed treatments, we are also interested in interventions that can be new treatment options. Causality provides a language to formalise those questions and can be a method for reasoning about the potential answers.

2 Causality in the medical setting

Pearl's causal ladder is the hierarchy that structures three types of questions in causality with increasing difficulty to answer[2]. Following is the causal ladder and the example questions in the medical setting. It is worth mentioning that for the analysis at the intervention and counterfactual level, we do not only rely on observed data but also on domain knowledge, experimental data and assumptions. Thus, supervised learning, diagnostic algorithms and digital biomarkers belong to the association level.

1. ASSOCIATION(Seeing): we are interested in the association between observed data, e.g. patient presents fatigue and persistent coughing- how likely is he to have lung cancer?

Lecture notes for the *Causal Representation Learning* course at ETH (fall semester 2020).

2. INTERVENTION(doing): we want to learn about what-will-happen-conditioning on some factors for a given case, e.g. will administering chemotherapy improve this patient's chances of survival? At this level, the treatments are the ones that exist in the observed world.

3. COUNTERFACTUALS(imagining): we would like to investigate what will happen if we apply some new treatments, e.g. what would have happened if we had administered immunotherapy in addition to chemotherapy? At this level, we think of the potential new treatment options.

Since the human body is a very complex system, it can be challenging to inference the causal generative process with potentially hundreds or even thousands of nodes. Also, how do the interactions between molecules and cells give rise to symptoms and phenomena spanning multiple scales of resolution? Furthermore, a controlled experiment can sometimes be either impractical, unethical or extremely costly[3], e.g. end-stage cancer with available efficacious treatment options. In those cases, collecting more data also does not help, which is relatively different from other machine learning cases.

3 Basic concepts

3.1 Counterfactual outcomes(potential outcomes)

We have $Y = \{y_0, y_1, y_2..., y_k\}$ as the outcomes that we either have, or would, observe for a given case after applying one of k treatments $t_0, \ldots, t_k$ (Rubin-Neyman Potential Outcomes Framework)[4]. We then call the difference between outcomes y_i and y_j for a given case the *individual causal effect* of treatment t_i in relation to treatment t_j. For example, in the binary treatment case($k = 2$), we have $Y = \binom{y_0}{y_1}$. y_0 is the outcome(e.g. survival/ death) after t_0(treatment 0) and y_1 is the outcome after t_1(treatment 1). Then $y_1 - y_0$ is the causal effect of treatment t_1 in relation to treatment t_0[5].

3.2 Treatment effect

In medical applications, we are interested in estimating

Average treatment effect: $ATE_{i,j} = E[y_{t_j} - y_{t_i}] = \sum_{i=0}^{N}(y_{t_j}(i) - y_{t_j}(i))$

Individual treatment effect(conditional average treatment effect): $ITE_{i,j} = E[y_{t_j} - y_{t_i}|X]$ where N is the amount of cases in the population and X is the number of covariates which are pretreated.

3.3 Quasi-experimental design and case control studies

Quasi-experimental design is an alternative type of empirical study that aims to infer causal effects from non-randomised experiments[4]. Understanding quasi-experimental design can help us understand methods for causal inference from observational data. When randomisation is impractical or unethical, we can use quasi-experimental design by only matching certain variables across groups. The degree of evidence of quasi-experimental design for causal effects is generally lower than that of randomised experiments. The main issue is the unknown distribution of the unmeasured confounders. Since the confounder is a common cause of the two nodes we are interested in analysing but not in the causal pathway, the unmeasured confounder causes a spurious association. The bias when it comes to the analysis of the outcome effects can come from the unmatched distribution of unmeasured confounders between two groups.

In case-control studies, outcomes across two groups are compared based on the presence or absence of a potential causal factor(not randomised). Observed confounding can be controlled for by matching cases with similar or equivalent controls, on the other hand, hidden confounding is not controlled for[6]. As an example, Doll and Hill's famous matched case-control study demonstrates the causal effect of smoking on lung carcinoma. Since it is not ethical to "force" the people to smoke, they matched the patients in the experiment to compare the effects of smoking.

To conduct a case-control study, we need to meet several necessary assumptions:
(1) **No unmeasured confounding**: $Y \perp\!\!\!\perp t|X$ (given X, treatment is conditionally independent of potential outcomes).
(2) **Common support assumption**:$0 < P(t = i) < 1$ for all X(for all X, it must be possible to be

assigned to a treatment group).
(3) **Stable Unit Treatment Value Assumption (SUTVA)**: Observed outcomes in units do not affect outcomes in any other outcome[7]. Those assumptions ensure the validity of the result from a case case-control study.

3.4 Backdoor criterion

The assumption for case-control studies are generally untestable and cannot be verified from observational data alone. Pearl provides a simple graphical criterion that implies the independence assumption based on a causal graph named backdoor criterion[8]. A set of variables Z satisfies the back-door criterion relative to an ordered pair of variables (X_i, X_j) in a DAG G if: (1) no node in Z is a descendant of X_i: and (2) Z blocks every path between X_i and X_j that contains a causal arrow into X_i[8].

3.5 Balancing scores

Balancing many covariates across groups at once has to deal with the curse of dimensionality. Balancing scores help to lower the dimensionality when balancing cohorts. Rosenbaum and Rubin proved that the lowest dimensional i.e. scalar balancing score is the propensity score e, i.e. the probability of assignment: $e_i(X) = p(t = i|X)$[9]. In general, the balancing score $b(X)$ is a function of the covariates X that has conditional independence $Y \perp\!\!\!\perp t|b(X)$. Note that $b(X) = X$ itself is a balancing score[9].

3.6 Counterfactual regression

Given a balanced setting, we can use counterfactual regression to find out the individual treatment effect. In the binary setting, a counterfactual estimator $f(X)$, predicts the response Y to a treatment t, given individual covariates X. $f(X,t) = \binom{y_0}{y_1}$, $f(X,t) = y_T = p(Y|X, do(t = T))$ In general, $p(Y|X, do(t = T))$ is not equal to predicting conditional probabilities $(p(Y|X,t))$. Thus, supervised learning cannot be applied to solve the problem[4].

4 Causal representation models

4.1 Counterfactual representation learning

Treatment agnostic regression networks (TARNet) are models to estimate individual treatment effects a in binary treatment case[5]. It is a neural network model with a shared base representation to counteract sparsification of features. In the binary case, the performance can be assessed by precision in estimating heterogeneous effects(PEHE) with μ_j as the unknown noiseless outcome distribution for outcome t_j: $\epsilon_{\text{PEHE}} = \frac{1}{N}\sum_{n=0}^{N}(E_{y_j(n)\sim\mu_j(n)}[y_1(0) - y_0(n)] - [\hat{y}_1(n) - \hat{y}_0(n)])^2$[10].

Based on the model in binary treatment, Perfect Match trains neural networks for counterfactual inference to compare any number of treatments[11]. It augments samples within minibatch with their propensity-matched nearest neighbours to control the biased assignment of treatments in observational data. The score used for evaluating the model gets extended to $\hat{\epsilon}_m\text{PEHE} = \frac{1}{\binom{k}{2}}\sum_{i=0}^{k-1}\sum_{j=0}^{i-1}\hat{\epsilon}_{\text{PEHE},i,j}$, i.e. pairwise PEHE between all treatments.

The model was recently generalised to the multiple treatment case with continuous dosages in DRNet. The computing time of the model grows very big but decreases the counterfactual error. Higher resolution dosage subdivision increases predictive performance but also increases the computation required. The error measuring equation is extended to the following errors:[12]

Mean Intergated Square Error(MISE): MISE=$\frac{1}{N}\frac{1}{|T|}\sum_{t\in T}\sum_{n=1}^{N}\int_{s=a_t}^{b_t}(y_{n,t}(s) - \hat{y}_{n,t}(s))^2 ds$

Dosage Policy error(DPE): DPE=$\frac{1}{N}\frac{1}{|T|}\sum_{t\in T}\sum_{n=1}^{N}(y_{n,t}(s_t^*) - y_{n,t}(\hat{s}_t^*))^2$, where $s_t^* = arg_{s\in[a_t,b_t]}\max y_{n,t}(s)$ and $\hat{s}_t = arg_{s\in[a_t,b_t]}\max\hat{y}_{n,t}(s)$.

Policy error(PE): PE=$\frac{1}{N}\sum_{n=1}^{N}(y_{n,t^*}(s^*_{t^*}) - y_{n,\hat{t}^*}(\hat{s}^*_{\hat{t}^*}))^2$, where $t^* = arg_{t\in T}\max y_{n,t}(s^*_t)$ and $\hat{t}^* = arg_{t\in T}\max \hat{y}_{n,t}(\hat{s_t}^*)$.

4.2 Causal explanation models

For neural network models, we often input many factors(e.g. age, weight and blood pressure) and obtain the output (e.g. how the factors affect heart failure risk). In some cases, we are not only interested if the risk of the certain patient increases but also the reason of why the risk increased[12]. We may take different actions with different medical factors. Thus, we would like to look into the black box of the model to interpret and estimate the feature importance. It can also be useful to debug models that should make sense and should be justifiable.

4.2.1 Attentive mixtures of experts network

We can output accurate predictions and the estimation of feature importance at the same time in attentive mixtures of experts networks(AME)[13]. There is one independent expert per feature group. The feature group only contributes to the final output passing through an attentive gating network. A local estimation is calculated at the attentive gate for a given expert. Attentive gates control expert contributions and the modulation at the gate is called Granger-causually grounded factor[13]. This model aims to jointly produce accurate predictions as well as estimating feature importance but have no incentive to solely output accurate feature importance estimates[14] and often collapses to use very few or even a single expert during training[15].

Granger-causality declares a relationship $X \rightarrow Y$ to be causal if we can better predict Y using all information compared to if all information apart from X had been used. We can define feature importance as the marginal reduction in prediction error associated with adding that feature. $\Delta\epsilon_{X,i} = \epsilon_{X\setminus\{i\}} - \epsilon_X$[13]. We then have a differentiable link between labels(prediction error) and feature importance which can be the objective function used to train models. Compared to SHAP(SHapley Additive exPlanations)[16] AME has similar or even better estimation accuracy in some cases but significantly lower running time. Lower mean Granger-causal error correlates with better feature importance estimates which justify the objective function used. Furthermore, AME discriminates well between cancer types and important and unimportant genes compared to LIME(perturbation-based approach)[17] and SHAP. Association discovered by AMEs are consistent with those reported by domain experts. The AME architecture has slightly lower prediction accuracy compared to RNN and FNN architectures.

There are many limitations to the AME model as we can only learn the importance of the feature without direction, e.g. we do not know if it is a positive or negative effect. The large number of experts increases the training time of the model. Further research on how to group the features may help to improve the model. The prediction performance decreases when we require joint learning. A special-purpose explainer-model is introduced to extend the Granger-causal loss to any machine learning model to help improving the prediction power in CXPLAIN model[18], where a predictive model and the adjacent explanation model are trained separately for a specified purpose, in which the explanation model only trains under a fixed predictive model only if a certain feature is important.

5 Conclusion

Causal inference is important to machine learning applications in medicine. The case-controlled study can help substantiate hypothetical causal effects, in the case of not being able to prove the randomisation in the experiment. Expressive counterfactual estimators, such as TARNet and DRNet, can be used to estimate counterfactual outcomes, under some essential assumptions. An explanation is an essential component of the machine learning system that is required to contextualise model outputs. Using a causal objective function, we can train explanation models that learn to explain the predictions of another model. Causal explanation(CXPlain) models produce accurate estimates of feature importance with comparatively low compute requirements at runtime. We have seen counterfactual representation learning and causal explanation models in this lecture. They can help choosing in between treatments for a given patient by comparing the outcome effects and propose new treatment options for further scientific research by providing the importance or feature groups.

Acknowledgments

These notes are based on a guest lecture given by Dr. Patrick Schwab on 10th and 24th of Nov 2020 (title: Causal Machine Learning in Healthcare) as part of the Causal Representation Learning seminar course at ETH Zurich during the fall semester of 2020.

References

[1] E. Hariton et al. “Randomised controlled trials—The gold standard for effectiveness research”. In: *BJOG: an international journal of obstetrics and gynaecology*, 125(13) (2018): 1716.

[2] J. Pearl et al. *The book of why: the new science of cause and effect*. Basic Books, 2018.

[3] P. Research et al. “Biopharmaceutical Industry-Sponsored Clinical Trials: Impact on State Economies”. In: (2015).

[4] M. A. Hernán et al. “Causal inference: what if”. In: *Boca Raton: Chapman & Hill/CRC*, 2020 (2020).

[5] U. Shalit et al. “Estimating individual treatment effect: generalization bounds and algorithms”. In: *International Conference on Machine Learning*. PMLR. 2017, pages 3076–3085.

[6] R. Doll et al. “Smoking and carcinoma of the lung”. In: *British medical journal*, 2(4682) (1950): 739.

[7] M. Lechner. “Identification and estimation of causal effects of multiple treatments under the conditional independence assumption”. In: *Econometric evaluation of labour market policies*. Springer, 2001, pages 43–58.

[8] R. Stone. “The assumptions on which causal inferences rest”. In: *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(2) (1993): 455–466.

[9] P. R. Rosenbaum et al. “The central role of the propensity score in observational studies for causal effects”. In: *Biometrika*, 70(1) (1983): 41–55.

[10] J. L. Hill. “Bayesian nonparametric modeling for causal inference”. In: *Journal of Computational and Graphical Statistics*, 20(1) (2011): 217–240.

[11] P. Schwab et al. “Perfect match: A simple method for learning representations for counterfactual inference with neural networks”. In: *arXiv preprint arXiv:1810.00656* (2018).

[12] P. Schwab et al. “Learning Counterfactual Representations for Estimating Individual Dose-Response Curves.” In: *AAAI*. 2020, pages 5612–5619.

[13] P. Schwab et al. “Granger-causal attentive mixtures of experts: Learning important features with neural networks”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Volume 33. 2019, pages 4846–4853.

[14] M. Sundararajan et al. “Axiomatic attribution for deep networks”. In: *arXiv preprint arXiv:1703.01365* (2017).

[15] N. Shazeer et al. “Outrageously large neural networks: The sparsely-gated mixture-of-experts layer”. In: *arXiv preprint arXiv:1701.06538* (2017).

[16] S. M. Lundberg et al. “A unified approach to interpreting model predictions”. In: *Advances in neural information processing systems*. 2017, pages 4765–4774.

[17] M. T. Ribeiro et al. “" 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*. 2016, pages 1135–1144.

[18] P. Schwab et al. “CXPlain: Causal explanations for model interpretation under uncertainty”. In: *Advances in Neural Information Processing Systems*. 2019, pages 10220–10230.