Causal Inference for Machine Learning
An Exploratory Study

Anol Kurian Vadakkeparampil *

1. Introduction

Causal inference in machine learning is a statistical approach used to understand cause-and-effect relationships between attributes. With the use of this technology, our AI and machine learning algorithms can reason similarly to how people do.

The goal of causal inference is to explain what factors lead (are influential) to the outcome. In the marketing domain, for example, decision-makers would be interested in knowing which campaign has the best conversion rate. The emphasis is on investigating and explaining the role of individual factors in the outcome.

Conversely, the majority of machine learning projects place more emphasis on the result and try to determine if the result will happen again in the future. However, understanding the cause-and-effect relationships can help improve future outcomes by changing our behavior based on understanding why something happened.

A Stanford Graduate School of Business course, for instance, covers fundamental ideas in machine learning-based causal inference, along with techniques for calculating causal effects in observational research. It also presents how machine learning can be used to measure the effects of interventions, understand the heterogeneous impact of interventions, and design targeted treatment assignment policies.

Recent developments in causal inference have shown exciting areas of development at the intersection of causal inference with machine learning. The incorporation of machine learning in causal inference enables researchers to better address potential biases in estimating causal effects and uncover heterogeneous causal effects.

For this project, I plan to study causal inference techniques in machine learning and their applications. Given below is a list of papers that I have studied:

- Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, *3*, 96-146. (Pearl, 2009)
- Shalit, U., Johansson, F. D., & Sontag, D. (2017). Estimating individual treatment effect: Generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning* (Vol. 70, pp. 3076-3085). (Shalit et al., 2017)
- Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., & Welling, M. (2017). Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems* (pp. 6446-6456). (Louizos et al., 2017)
- Schölkopf, B. (2019). Causality for machine learning. *arXiv preprint arXiv:1911.10500*. (Schölkopf, 2019)
- Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, *20*(1), 217-240. (Hill, 2011)
- Glymour, M., Zhang, K., & Spirtes, P. (2019). Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, *10*, 524. (Glymour et al., 2019)

2. Papers

2.1. Causal inference in statistics: An overview

(Pearl, 2009)

Recent innovations in causal inference are highlighting major shifts required to transition from conventional statistical analysis to causal evaluation of multivariate information. This paper spotlights the premises underlying all causal conclusions, the terminology used to articulate those premises, the conditional essence of causal and counterfactual arguments, and the techniques developed to appraise such arguments.

Utilizing a broad theory of causality centered on the Structural Causal Model (SCM) framework, which encapsulates and synthesizes other causal approaches and erects a unified mathematical basis for analyzing causes and counterfactuals, the paper elucidates these advancements. The SCM framework facilitates coherent causal analysis while unifying distinct causal philosophies. Through this model, the paper aims to demonstrate the pivotal principles and techniques emerging within causal inference.

The paper surveys the development of mathematical tools for inferring (from a combination of data and assumptions)

answers to three types of causal queries:

- Queries about the effects of potential interventions (also called "causal effects" or "policy evaluation").
- Queries about probabilities of counterfactuals (including assessment of "regret," "attribution" or "causes of effects").
- Queries about direct and indirect effects (also known as "mediation").

Some important concepts gleaned from the paper:

Structural Causal Model (SCM): This paper emphasizes the SCM as a unified framework for causal analysis, providing mathematical tools to evaluate interventions, counterfactual probabilities, causal effects, and more.

Three Types of Causal Queries: Pearl discusses mathematical tools for inferring answers to queries about effects of potential interventions, probabilities of counterfactuals (including assessment of regret, attribution, or causes of effects), and direct and indirect effects (mediation).

Relationships Between Structural and Potential-Outcome Frameworks: The paper outlines the formal and conceptual relationships between these frameworks and presents tools for symbiotic analysis utilizing the strengths of both.

Counterfactual Analysis: Pearl elaborates on the use of counterfactual analysis in structural models, providing a comprehensive understanding of its application in causal inference.

Identifiability and Estimation of Causal Effects: The paper discusses the concepts of identifiability and the estimation of causal effects, crucial for understanding the impact of interventions in various settings.

Graphical Models and Causal Inference: Pearl highlights the importance of graphical models in simplifying and elucidating causal inference problems.

2.2. Estimating individual treatment effect: Generalization bounds and algorithms

(Shalit et al., 2017)

This paper addresses the challenge of predicting individual treatment effects (ITE) from observational data under the assumption of strong ignorability. The authors introduce a new theoretical analysis and a family of algorithms for this purpose. The algorithms learn a "balanced" representation such that the treated and control distributions appear similar. The paper also provides a novel generalization-error bound for ITE estimation, incorporating standard generalization-error and the distance between treated and control distributions.

Key Contributions:

- Introduction of Integral Probability Metrics (IPMs): The paper uses IPMs, specifically the Wasserstein and Maximum Mean Discrepancy (MMD) distances, to measure the distances between distributions, deriving explicit bounds for these distances.
- Novel Generalization-Error Bound: The paper presents a new generalization-error bound that shows the expected ITE estimation error of a representation is bounded by the sum of the standard generalization-error of that representation and the distance between the treated and control distributions induced by the representation.
- Algorithm Development: Based on their theoretical findings, the authors develop a family of representation-learning based algorithms for ITE estimation.

This paper makes significant contributions to the field of causal inference in machine learning by providing a robust theoretical framework and practical algorithms for estimating individual treatment effects from observational data, with potential applications in healthcare, economics, and education.

2.3. Causal Effect Inference with Deep Latent-Variable Models

(Louizos et al., 2017)

The paper focuses on the challenge of inferring individual-level causal effects from observational data, particularly when dealing with confounding factors. The authors introduce a deep latent-variable model, specifically a Variational Autoencoder (VAE), to address this challenge. The VAE is used to infer complex non-linear relationships between observed variables and hidden confounders.

Key technical aspects include:

- Latent Variable Modeling: The approach involves estimating a latent space that summarizes the confounders and infers their impact on treatment and outcomes.
- Use of Proxies for Confounders: The model addresses the issue of unmeasured confounders by using proxy variables.
- Variational Autoencoders (VAEs): The paper builds upon VAEs for approximate inference, allowing for weaker assumptions about the data generating process and the structure of the hidden confounders.

The paper presents an innovative approach to causal inference using deep learning and latent-variable modeling, offering significant advantages in handling hidden confounders

and proxy variables. This method opens new avenues for robust causal effect estimation in complex observational data settings.

2.4. Causality for machine learning

(Schölkopf, 2019)

In Bernhard Schölkopf's paper "Causality for Machine Learning," he examines the integration of causal reasoning into machine learning (ML). The paper critically analyzes how conventional ML relies heavily on statistical correlations, often failing to consider the underlying causal structures. This limitation can lead to models that don't generalize well across different settings or when the data distribution changes.

Key Technical Details:

- Causal Representation Learning: Schölkopf emphasizes the importance of learning representations that capture the causal structure underlying the data.
- Independent Causal Mechanisms (ICM): He introduces the principle of ICM, which suggests that the mechanisms causing the data are independent of each other.
- Causal Inference in AI: The paper explores the use of causal inference in AI, particularly in areas where traditional ML struggles, like in changing environments.
- Counterfactual Reasoning: Schölkopf discusses the role of counterfactuals in understanding causality, which is crucial for making predictions under interventions or changes.
- Applications and Future Directions: The paper highlights potential applications of causal learning in various fields and suggests directions for future research.

In conclusion, Schölkopf's paper is a call to action for the ML community to embrace causal reasoning, as it offers the potential to build more robust, understandable, and generalizable models.

2.5. Bayesian Nonparametric Modeling for Causal Inference

(Hill, 2011)

The paper introduces Bayesian Additive Regression Trees (BART) as a strategy for estimating causal effects in non-experimental settings. This method focuses on flexibly modeling the response surface, handling a large number of predictors, and accommodating continuous treatment variables and missing outcome data. BART is shown to be more accurate in estimating average treatment effects compared to traditional methods like propensity score matching and regression adjustment, especially in nonlinear simulation settings.

Key aspects of the paper include:

- BART and Estimating Causal Effects: Discussion on using BART for causal inference, including generalizability, and handling missing outcome data.
- Simulations Based on Real Data: Simulation studies to demonstrate the effectiveness of BART in various settings.
- Robustness of Simulation Findings: Exploration of the robustness of BART's performance under different scenarios.
- Estimating Dosage Effects: Application of BART in real-world scenarios, comparing its performance with other methods.

Hill's paper provides a significant contribution to causal inference methodology, demonstrating the effectiveness of Bayesian nonparametric modeling, particularly BART, in accurately estimating causal effects in complex scenarios.

2.6. Bayesian Nonparametric Modeling for Causal Inference

(Glymour et al., 2019)

The paper "Review of Causal Discovery Methods Based on Graphical Models" by Glymour, Zhang, and Spirtes (2019) presents a comprehensive overview of methods for causal discovery using graphical models. It highlights the evolution of computational methods for causal discovery over three decades, focusing on constraint-based and score-based methods, as well as those based on functional causal models. The paper discusses the challenges and practical issues in causal discovery, particularly in biological applications, and offers guidance for the choice and use of various methods.

Key technical aspects of the paper include:

- Directed Graphical Causal Models (DGCMs): It explains DGCMs and their components, including variables, directed edges, and probability distributions.
- Constraint-Based and Score-Based Methods: The paper discusses traditional methods like the PC and FCI algorithms for causal discovery, highlighting their assumptions and limitations.
- Functional Causal Models (FCMs): It reviews causal discovery methods based on linear non-Gaussian models and non-linear models, stressing their importance in identifying causal relationships.

- Practical Issues in Causal Discovery: The paper addresses challenges like causality in time series, measurement errors, selection bias, and missing data in causal discovery.

This paper provides an extensive review of the methodologies and challenges in causal discovery using graphical models, emphasizing the interdisciplinary nature and application of these methods in various scientific domains.

3. Algorithm Implemented : CEVAE

(Louizos et al., 2017)

3.1. Dataset Used

The Infant Health and Development Program (IHDP) dataset, when used in the context of the Counterfactual Variational Autoencoder (CEVAE) algorithm, serves as a benchmark for evaluating the algorithm's ability to estimate causal effects from observational data. Here are some specific details regarding the IHDP dataset in relation to CEVAE:

Data Characteristics: The IHDP dataset includes numerous covariates such as demographic information, health status of the infants, and family background. These covariates are essential for CEVAE as they are used to predict outcomes and understand the impact of treatment.

Treatment and Outcome: In the IHDP data, the 'treatment' typically refers to a special intervention program aimed at enhancing children's cognitive development. The 'outcome' is usually a measure of cognitive test scores at a certain age. CEVAE aims to estimate the causal effect of this treatment on the outcome.

Selection Bias Simulation: For the use in CEVAE and other causal inference models, the IHDP data is often manipulated to simulate observational data with selection bias. This is done by selectively removing a subset of treated individuals, which breaks the random assignment of the original experiment. This setup allows CEVAE to demonstrate its ability to handle confounding bias and estimate causal effects accurately.

Evaluation of CEVAE Performance: The primary goal in using the IHDP dataset with CEVAE is to evaluate the model's performance in estimating the Average Treatment Effect (ATE). The ATE is calculated as the difference in outcomes between the treated and control groups, averaged over the population.

Benchmarking and Comparisons: Since the IHDP dataset includes ground truth for counterfactual outcomes (due to its origin from a randomized controlled trial), it is an ideal benchmark for comparing CEVAE's performance against other causal inference methods.

Data Preprocessing for CEVAE: Before being fed into the CEVAE model, the IHDP data typically undergoes preprocessing steps like normalization and encoding, which are crucial for the performance of deep learning models.

Research Implications: The use of IHDP in conjunction with CEVAE provides insights into the effectiveness of deep generative models in causal inference, particularly in scenarios where traditional methods might struggle due to confounding variables and selection biases.

Limitations in Generalization: While IHDP provides a controlled environment to test CEVAE, the specific nature of the dataset (focused on infant health) means that the generalization of the results to other domains should be done with caution.

3.2. Metrics for Causal Inference

1. **ITE (Individual Treatment Effect)**: ITE refers to the effect of a treatment or intervention on an individual level. It is the difference in outcome for the same individual if they were treated versus if they were not. Mathematically, for an individual i, ITE is defined as $ITE_i = Y_i(1) - Y_i(0)$, where $Y_i(1)$ is the potential outcome if the individual receives the treatment and $Y_i(0)$ is the potential outcome if the individual does not receive the treatment. Estimating ITE is challenging because we can never observe both potential outcomes for the same individual in reality.

2. **ATE (Average Treatment Effect)**: ATE is the average effect of a treatment across a population. It is the expected difference in outcomes between treated and untreated groups. In mathematical terms, ATE is defined as $ATE = E[Y(1) - Y(0)]$, where $Y(1)$ and $Y(0)$ are the potential outcomes with and without the treatment, respectively, and E denotes the expected value. ATE provides a summary measure of the effectiveness of a treatment at the population level.

3. **PEHE (Precision in Estimation of Heterogeneous Effect)**: PEHE measures the precision in estimating heterogeneous treatment effects, which are variations in treatment effects across individuals or groups. It is often used to evaluate the performance of models that estimate ITE. The metric is defined as the mean squared error between the true ITE and the estimated ITE for each individual: $PEHE = E[(\hat{ITE}_i - ITE_i)^2]$, where $\hat{ITE}_i$ is the estimated treatment effect for individual i. Lower PEHE values indicate higher precision in estimating individual-level treatment effects.

3.3. Algorithm

Objective: CEVAE aims to estimate individual treatment effects (ITE) from observational data. It is particularly useful in scenarios where experimental data (like randomized control trials) are not available or ethical to obtain.

Framework: The model is based on Variational Autoencoders (VAEs), a class of deep latent-variable models. VAEs are generally used for learning complex data distributions, but CEVAE adapts this approach for causal inference.

Structure: CEVAE consists of several neural network components:

Encoder Networks: These networks infer the latent variables from observed variables (including treatment and outcome variables).

Decoder Networks: They are used to reconstruct the observed variables from the latent ones. This reconstruction is crucial for learning the joint distribution of observed and latent variables.

Latent Variables: CEVAE introduces latent variables that capture confounders - variables that affect both treatment and outcome. This is a crucial aspect, as confounders are often unobserved in observational studies.

Inference: The model uses variational inference to estimate the posterior distribution of latent variables. This involves optimizing a lower bound on the likelihood of the observed data (Evidence Lower BOund, or ELBO).

Treatment Effect Estimation: After training, CEVAE can predict potential outcomes under different treatment scenarios for the same individual. The difference in these potential outcomes is used to estimate the ITE.

Advantages: CEVAE addresses some limitations of traditional causal inference methods, like the need for strong assumptions about confounders and limitations in handling high-dimensional, non-linear relationships.

Applications: The model has potential applications in various fields like healthcare, economics, and social sciences, where understanding the causal effect of interventions is crucial.

The CEVAE model represents a significant advancement in the field of causal inference, leveraging deep learning techniques to tackle complex, real-world problems where causal relationships are essential yet challenging to discern.

3.4. PseudoCode

1. **Input:** Observational data (X, T, Y), where X are covariates, T is the treatment indicator, and Y is the outcome.
2. **Initialize** parameters of encoder networks ϕ, decoder networks θ, and the distribution of latent variable Z.
3. **Repeat for** a number of training iterations:
   (a) **For each** minibatch $\{X_i, T_i, Y_i\}$:
      i. **Encode** X_i, T_i, Y_i into latent space: $Z_i \sim q_\phi(Z|X_i, T_i, Y_i)$.
      ii. **Reconstruct** X_i, T_i, Y_i from latent space: $\hat{X}_i, \hat{T}_i, \hat{Y}_i \sim p_\theta(X, T, Y|Z_i)$.
      iii. **Calculate** loss function, typically a combination of reconstruction loss and KL divergence: $L = \text{ReconstructionLoss}(X_i, \hat{X}_i, T_i, \hat{T}_i, Y_i, \hat{Y}_i) + \text{KLDivergence}(q_\phi(Z|X_i, T_i, Y_i)||p(Z))$.
      iv. **Update** parameters ϕ and θ using gradient descent to minimize loss L.
   (b) **End For**
4. **End Repeat**
5. **Output:** Trained CEVAE model with parameters ϕ and θ.
6. **For Counterfactual Inference:** Use the trained model to estimate counterfactual outcomes given new data.

3.5. CEVAE performance on IHDP

Some observations on the performance of a model on training and testing datasets using the Individual Treatment Effect (ITE), Average Treatment Effect (ATE), and Precision in Estimation of Heterogeneous Effect (PEHE) metrics:

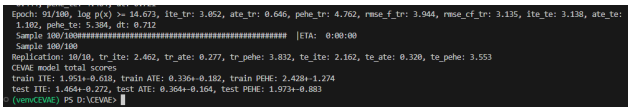

Figure 1. Execution of CEVAE

Training Scores:

For ITE, the score is approximately 0.5 with a relatively small confidence interval, suggesting that the model has moderate accuracy in predicting individual treatment effects on the training data.

The ATE score is close to zero with a very tight confidence interval, which indicates that the model is very consistent in estimating the average treatment effect on the training data.

The PEHE score is high (around 3.5) with a wide confidence interval. This suggests that the model has considerable error in estimating the heterogeneous treatment effects across individuals in the training set.

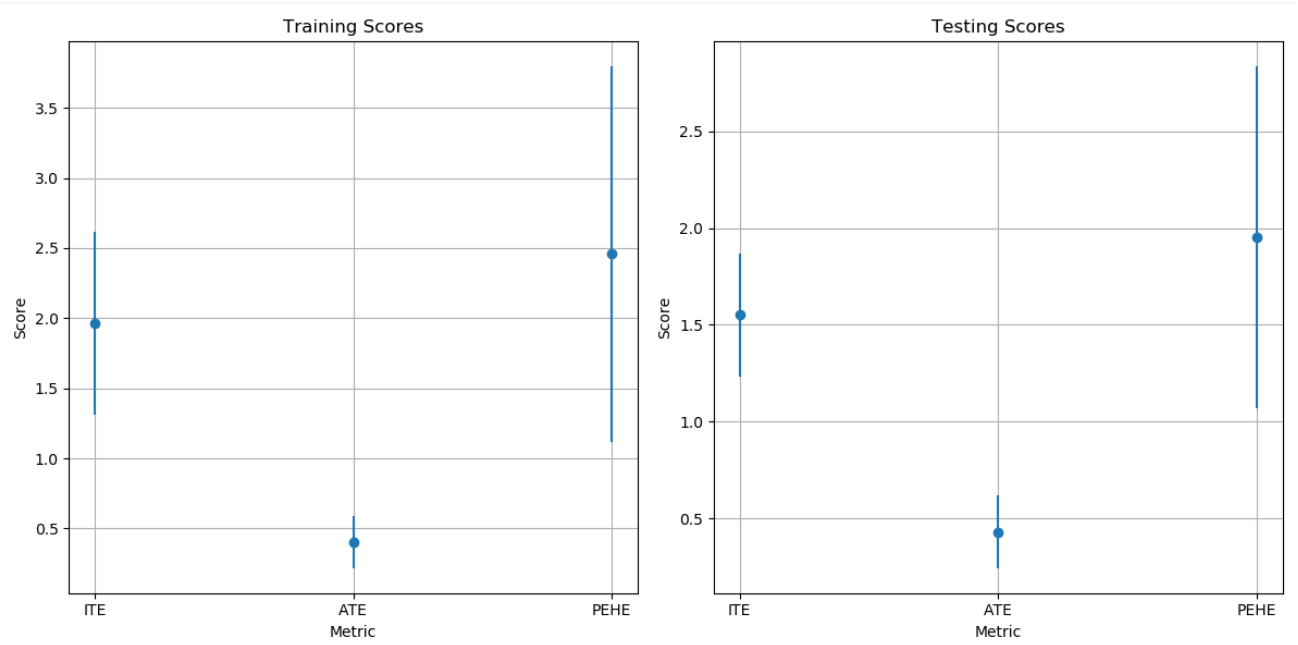

Figure 2. Metrics for CEVAE on IHDP

Testing Scores:

The ITE score on the testing data is similar to the training data, around 0.5, but with a slightly wider confidence interval. This suggests a consistent performance of the model in estimating individual treatment effects but with a bit more uncertainty on unseen data.

The ATE score on the testing set is also close to zero, similar to the training set, and the confidence interval remains tight. This indicates that the model's estimation of the average treatment effect is consistent and generalizes well.

For PEHE, the score is much higher on the testing set (just below 2.5) compared to the training set, and the confidence interval is also wider. This indicates a significant drop in performance when estimating heterogeneous treatment effects on unseen data. Overall Comments:

The model seems to estimate the average treatment effect (ATE) consistently well both on training and testing data, given the low scores and tight confidence intervals.

There is a clear drop in performance from training to testing for the PEHE metric, indicating that the model's ability to capture the heterogeneity in treatment effects is less robust on unseen data. The moderate ITE scores with relatively small confidence intervals suggest that the model has a reasonable performance in estimating individual treatment effects, although the confidence interval widens in the testing data.

The wide confidence intervals for PEHE, especially in the testing data, suggest that the model's precision in estimating heterogeneous effects is not very reliable, and there could be overfitting to the training data.

The model's performance might benefit from additional tuning, feature engineering, or exploration of model complexity to improve the estimation of heterogeneous effects (as indicated by PEHE) while maintaining the strong performance on ATE.

In conclusion, the model appears to be reliable in estimating the ATE but shows limitations in capturing the variability of treatment effects across individuals, especially when applied to new data. Further investigation into the model's architecture and training process could potentially improve its generalizability and precisio

4. Conclusion

In conclusion, my exploratory study, informed by the foundational works of Pearl, Shalit, Louizos, Schölkopf, Hill, and Glymour, involved the implementation of the Counterfactual Variational Autoencoder (CEVAE) algorithm to enhance my understanding of causal inference in complex systems. I have explored the effectiveness of CEVAE in revealing hidden causal relationships within datasets. The study not only buttressed the theoretical concepts proposed by these authors but also provide practical perspectives on the application of deep learning in causal analysis.

This study explores the significance of merging advanced machine learning models with classical causal inference theory, offering a more intricate and comprehensive perspective on causal understanding.

Future works would include, implementation of more algorithms and further exploration into this topic to capture its depth.

References

Glymour, M., Zhang, K., and Spirtes, P. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10:524, 2019.

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

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pp. 6446–6456, 2017.

Pearl, J. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009.

Schölkopf, B. Causality for machine learning. arXiv preprint arXiv:1911.10500, 2019.

Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: Generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 3076–3085, 2017.