Causal Inference in Machine Learning

From Prediction to Decision Making: A review

Finnian Lattimore
Australian National University

NICTA

Introduction

Would reducing salt intake reduce heart attack risk? Would upping the minimum wage increase unemployment? Causal problems differ from the traditional machine learning setting in that they require us to predict the consequences of an intervention, which may change the properties of the distribution from which our data is sampled. Without experimentally testing the intervention or making assumptions to constrain its effect, such inference is impossible. There are many important questions where direct experimentation is expensive, unethical or impossible. Recent research on causality has clarified what assumptions allow causality to be determined and has demonstrated causal inference and discovery is possible with some very general assumptions.

Causal Frameworks

Causal Directed Acyclic Graphs

- A causal DAG is a Bayesian network where $A \to B$ is defined to mean A causes B.
- Variables are independent of their non-effects given their direct causes (Causal Markov Property)
- An intervention that sets a subset of variables $\boldsymbol{X}$ to $\boldsymbol{x}$, denoted $do(\boldsymbol{X} = \boldsymbol{x})$, has a simple graphical representation in a causal DAG, G. All links entering intervened on variables, $\boldsymbol{X}$, are deleted, resulting in the mutilated network $G_{\overline{\boldsymbol{X}}}$ (figure 1). Thus, a causal DAG represents the set of all possible interventional distributions over its variables.

Figure 1: *Intervention in a causal DAG*

(Causal) Structural Equation Models (SEMs)

- Represent each variable as a deterministic function of its direct causes and a noise term, where the noise terms are mutually independent.
- If the set of equations does not create a cycle then the Causal Markov Property holds and the SEM is a causal DAG, (but not visa-versa - SEMs can encode more information).

Counterfactuals

- Counterfactuals are statements about what would happen under alternate realities where some specified thing differs. For example, consider people taking a medical drug:
  For an individual, i, let: $\begin{cases} y_i^0 = & \text{outcome if } x_i = 0 \text{ (not treated)} \\ y_i^1 = & \text{outcome if } x_i = 1 \text{ (treated)} \end{cases}$
- We can define a random variable Y^1, where $P(Y^1)$ is the distribution of outcome, Y, that would occur if everyone was treated. Similarly $P(Y^0)$ is the distribution of outcome if no-one was treated.
- If $(X \perp\!\!\!\perp Y^0|\boldsymbol{Z})$ & $(X \perp\!\!\!\perp Y^1|\boldsymbol{Z})$: $\longleftarrow$ Ignoreability Assumption
  we can calculate counterfactual distributions from observed ones:

$$P(Y^1|Z) = P(Y|X = 1, Z) \text{ and } P(Y^0|Z) = P(Y|X = 0, Z)$$

- Distributions over counterfactual variables that correspond to interventions can be translated directly to the do notation $P(Y^1) = P(Y|do(X = 1))$. However we can phrase queries with counterfactual variables that are not interventional. For example: *what is the proability that Joe, who was not treated and died, would have recovered had he been treated?*. This query asks about the joint distribution of $P(Y^0, Y^1)$.

group	placebo	treatment	probability of group
1	die	die	$\alpha = P(Y^0 = 0, Y^1 = 0)$
2	die	recover	$\beta = P(Y^0 = 0, Y^1 = 1)$
3	recover	die	$\gamma = P(Y^0 = 1, Y^1 = 0)$
4	recover	recover	$\delta = P(Y^0 = 1, Y^1 = 1)$

- Counterfactuals can be defined in terms of SEMs with a slight relaxation of the independence of errors assumption [8].

Causal Inference

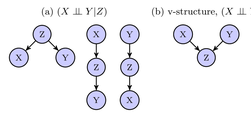
Consider the problem where the causal DAG is known, due to theory or prior knowledge and we wish to infer the outcome of an intervention of the form $P(\boldsymbol{Y}|do(\boldsymbol{X} = \boldsymbol{x}))$ using observational data.

- If there are no latent variables, we can compute outcome of any intervention by simply multiply the factors in the mutliated network (figure 1).

The Do Calculus

The do-calculus consists of three rules, which result from d-separation in a causal DAG [6]. It is complete: a causal effect is non-parametrically identifiable if and only if the interventional query can be reduced to an observational one via these rules [10].

Figure 2: *d-separation allows us to read conditional independences off a DAG. If a set of variables $\boldsymbol{Z}$ d-separates $\boldsymbol{X}$ and $\boldsymbol{Y}$ in G then $(\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z})$ in all distributions P compatible with G.*



The Three Rules

1. A causal DAG remains a causal DAG after an intervention so d-separation still applies.

   if $(\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{W}|\boldsymbol{X})$ in $G_{\overline{\boldsymbol{X}}}$
   $$P(\boldsymbol{Y}|do(\boldsymbol{X} = \boldsymbol{x}), \boldsymbol{W} = \boldsymbol{w}) = P(\boldsymbol{Y}|do(\boldsymbol{X} = \boldsymbol{x}))$$

2. If $\boldsymbol{Y}$ is independent of *how* variables $\boldsymbol{X}$ take their values then the effect on $\boldsymbol{Y}$ of setting $\boldsymbol{X}$ to some value is equivalent to observing it take that value. If this rule is satisfied the corresponding ignoreability assumption in the counterfactual framework is satisfied.

   if $(\boldsymbol{Y} \perp\!\!\!\perp \hat{\boldsymbol{X}}|\boldsymbol{X}, \boldsymbol{Z})$ in $G^\dagger$
   $$P(\boldsymbol{Y}|do(\boldsymbol{X} = \boldsymbol{x}), \boldsymbol{Z}) = P(\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Z})$$

3. If there is no direct causal path from $\boldsymbol{X}$ to $\boldsymbol{Y}$ then intervention on $\boldsymbol{X}$ does not change the distribution of $\boldsymbol{Y}$.

   if $(\boldsymbol{Y} \perp\!\!\!\perp \hat{\boldsymbol{X}}|\boldsymbol{Z})$ in $G^\dagger$
   $$P(\boldsymbol{Y}|do(\boldsymbol{X} = \boldsymbol{x}), \boldsymbol{Z}) = P(\boldsymbol{Y}|\boldsymbol{Z})$$

(For readability, this is a simplified version of the do-calculus that covers interventions on a single variable or cases where it is sufficient for identifiability to consider intervention on all variables together. The fully general version is only slightly more complex see [6])

Causal Discovery

Causal discovery attempts to infer causal structure from data based on more general assumptions.

Indepdendence Based Methods

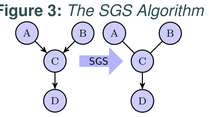
Without Latent Variables

1. We assume our distribution P was generated by some (unknown) causal DAG over our observed variables (causal sufficiency)
2. We assume that all the conditional independences in P are implied by d-separation in the true causal network (***faithfulness***)
3. Finding the causal structure equates to finding perfect maps for P

Algorithm 1: *SGS or IC Algorithm [11, 6].*
Input: A Distribution P over variables $\boldsymbol{V}$
Output: A partially directed network representing the Markov equivalence class for the generating causal model.

1. Create a complete undirected graph over V. For all pairs of variables $(a, b) \in V$ search for a set S_{ab} s.t $a \perp\!\!\!\perp b|S_{ab}$. If such a set exists, delete the link $a - b$
2. For all pairs of unlinked-nodes (α, β) with a common neighbour c, if $c \notin S_{\alpha\beta}$ direct links towards c.
3. Recursively direct any remaining links for which there is only one orientation that does not create a cycle or any additional v-structures ($\bullet \to \bullet \leftarrow \bullet$).

Figure 3: *The SGS Algorithm*

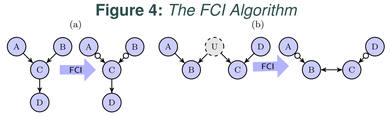


- The SGS algorithm is infeasible in practise due to the exponential number of (high order) conditional independence tests it requires.
- The PC algorithm [11] modifies the SGS algorithm to exploit any sparsity in the true network, leading to much better average case performance.

With Latent Variables

- For every latent structure there is a dependency equivalent structure in which every latent variable is a root node with exactly two children [12]. This key to the FCI algorithm [11], which generalizes the PC algorithm to handle latent and selection variables.
- The FCI algorithm returns an equivalence class of Maximal Ancestral Graphs (a generalization of DAGs) since DAGs are not closed under marginalization (figure 4b).

Figure 4: *The FCI Algorithm*



- The FCI algorithm discovers all aspects of causal structure identifiable from conditional independence relations [13].
- It can be made to require a worst case polynomial (rather than exponential) number of conditional independence tests for sparse graphs [1].
- Implementations of both the PC and FCI algorithm are available in the R package pcalg [4]

Beyond independence

Independence based methods have the advantage that they require only very general, non-parametric, assumptions. However they cannot distinguish between causal graphs with equivalent dependency structure; for example, between $X \to Y$ and $X \leftarrow Y$.

Figure 5: *Figure from [3]. Additive noise models, $Y = f(X) + \epsilon$, are identifiable for most combinations of f and $P(\epsilon)$ but not in linear-gaussian case*

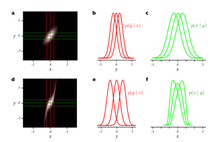


Figure 6: *Figure from [2]. The causal direction can be identifiable even where the relationship between X and Y is deterministic and invertible. Let $Y = f(X) \Leftrightarrow X = g(Y)$. For most input distributions, $p_X(x)$, the distribution $p_Y(y)$ will be higher where f' is small and a larger region of X maps to similar values of Y. If X causes Y (but not if Y causes X) we would expect f and $p_X(x)$ to be independent and $p_Y(y)$ should be correlated with f'.*

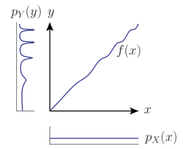
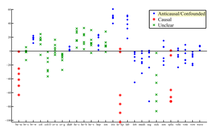


Figure 7: *Figure from [9]. The idea of independence of mechanism and input can be generalized to the non-deterministic setting. If $X \to Y$ then $P(X)$ and $P(Y|X)$ should be independent, but $P(Y)$ and $P(X|Y)$ are not. Therefore, semi-supervised learning should yield no benefit when trying to learn in the causal direction, (estimating $P(Y|X)$) but could help when learning in the anti-causal direction.*



Learning what causality looks like [5]

Suppose we had M different causal pairs data sets.

$$D = \{\{x_j, y_j\}_{j=1}^{N_i}, l_i\}_{i=1}^{M}$$

where l_i is a binary label that indicates if $X \to Y$ or $Y \to X$ in dataset i.

- Kernel mean embedding allows us to take a distribution P and transform it to a point in some Hilbert space.
- We expect there to be differences in the relationships between $P(X)$, $P(Y)$ and $P(Y|X)$ for $X \to Y$ and $Y \to X$

Algorithm 2:

1. Let μ be a kernel mean embedding that maps a distribution P into some Hilbert space.
2. For each data set $i = 1...M$, construct a feature vector that approximates $\mu(P(X)), \mu(P(Y)), \mu(P(X, Y))$
3. Apply a standard classification algorithm to learn if $X \to Y$ or $Y \to X$

With more than two variables

- If you can come up with a condition, on the triple $(P(Y), P(\epsilon), f)$, that guarantees identifiability for the bivariate SEM $Y = f(X) + \epsilon$, you can extend that result to get the conditions under which the multivariate case is identifiable [7].

References

[1] Tom Claassen, J Mooij, and Tom Heskes. Learning sparse causal models is not NP-hard. In *UAI*, 2013.

[2] P Daniusis, Dominik Janzing, and Joris Mooij. Inferring deterministic causal relations. In *UAI*, 2010.

[3] PO Hoyer, Dominik Janzing, and JM Mooij. Nonlinear causal discovery with additive noise models. In *NIPS*, 2009.

[4] Markus Kalisch and M Mächler. Causal inference using graphical models with the R package pcalg. *JSS*, VV(Ii), 2012.

[5] David Lopez-Paz, Krikamol Muandet, and Benjamin Recht. The Randomized Causation Coefficient. *arXiv Prepr. arXiv1409.4366*, September 2014.

[6] Judea Pearl. *Causality: models, reasoning and inference*. MIT Press, Cambridge, 2000.

[7] Jonas Peters, J Mooij, Dominik Janzing, and B Schölkopf. Causal discovery with continuous additive noise models. *JMLR*, 15:2009–2053, 2014.

[8] TS Richardson and JM Robins. Single world intervention graphs (SWIGs): a unification of the counterfactual and graphical approaches to causality. *Cent. Stat. . . .*, (128), 2013.

[9] B Schölkopf, Dominik Janzing, and Jonas Peters. On causal and anticausal learning. In *ICML*, 2012.

[10] Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. *JMLR*, 9:1941–1979, 2008.

[11] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*, volume 81. MIT press, 2000.

[12] TS Verma. Graphical aspects of causal models. Technical report, 1993.

[13] Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, November 2008.