Non-probabilistic Markov categories for causal modeling in machine learning

Dhurim Cakiqi[1], Max A. Little[1,2]

[1]*School of Computer Science, University of Birmingham, UK and* [2]*Media Lab, MIT, USA*

Statistical machine learning (ML) and structural causal modeling have in common the use of *Bayesian networks*, usually represented using *directed acyclic graphs* (DAGs). Such networks represent conditional relationships between random variables at the DAG nodes, where directed edges between nodes $A \to B$ denote that variable B is conditioned on A. When given an explicitly causal interpretation, this conditioning coincides with the direction of mechanistic influence between variables in the network [6]. Factorized according to the chain rule of probability and simplified by the implied conditional independence relations between variables encoded in these edges, the joint distribution over all the variables on the nodes is given by the product of the conditional distributions of each node. For example, given a DAG $D \leftarrow C \leftarrow A \to B \to D$, the joint distribution is given by $P(A,B,C,D) = P(A)\,P(B|A)\,P(C|A)\,P(D|B,C)$.

Causal inference in Bayesian networks involves isolating the *direct* causal influence between variables from the DAG of observed relationships, whilst eradicating other, unwanted (spurious, indirect) relationships. This involves simulating experimental *interventions*, that is, changes to the structure of the DAG [6]. In particular, the interventional probability $P(Y|\mathrm{do}\,(X=x))$ is computed after replacing all incoming edges to the variable X with the (unconditional) constant $X = x$. Crucially, this differs from the ordinary conditional probability $P(Y|X=x)$ which is computed from the unchanged DAG. This *calculus of* (*causal*) *interventions* ("do-calculus") [6] can be used to derive *adjustment methods* to isolate direct effects. For instance, assume variable A is the age of a patient, D is whether a drug treatment is taken, and H is the status of the patient's health. The model $H \leftarrow A \to D \to H$ captures the typical situation where age influences health status independent of the drug, while older subjects are less likely to take the drug if made available to them. If we are interested in the direct causal effect of D on H (i.e. whether the treatment works absent the effect of patient age A), we must somehow remove the indirect relationship between D and H which exists due to the spurious causal path through the *confounding* variable A. In this DAG, the so-called *back-door adjustment* formula

$$P(H|\mathrm{do}\,(D=d)) = \sum_{a\in\Omega_A} P(H|A=a,D=d)\,P(A=a) = \sum_{a\in\Omega_A} \frac{P(H,A=a,D=d)\,P(A=a)}{P(A=a,D=d)},$$

allows us to compute the desired direct effect. Where the structure of the DAG permits, adjustment methods for arbitrary DAGs with hidden variables have recently been devised [7].

Such probabilistic causal modeling and adjustment is widely used in disciplines such as epidemiology, macroeconomics and bioinformatics where controlled experimentation is generally impossible, allowing the estimation of direct effects from purely observational data. In a similar way, ML and artificial intelligence (AI) learn sophisticated prediction models from large-scale observational data. However, the highly pernicious problems of spurious correlations inherent to this "causally-blind" approach have only recently been acknowledged, for instance, in ML models for medical prognosis [2]. Therefore, while it would be of substantial advantage to ML and AI to make use of the techniques of causal inference, unfortunately, the best performing ML algorithms are nearly non-probabilistic models. For instance, *kernel support vector machine classifiers* and *deep learning* algorithms make heavy use of compositions of deterministic matrix transformations, nonlinear *activation* functions such as the rectified linear unit (ReLU), and loss functions such as the *cross-entropy* or *hinge loss*. These are all deterministic operations which have no meaningful probabilistic interpretation. It follows that we cannot directly apply the techniques of causal inference to these algorithms because the mathematical foundations of these structures are fundamentally incompatible.

To address this problem, we note that causal modeling and adjustment requires computations over Bayesian networks which must, at the minimum, support;

- *Joint states*: combines random variables,
- *Marginals*: separate variables and normalizes distributions,
- *Disintegration*: conditioning to enable the chain rule of factorization.

At the root of these operations are the usual arithmetic computations of multiplication (joint states), summation (marginalization, normalization) and division (disintegration), carried out in the *probability algebra* in which joint states are the usual products e.g. $P(X,Y) = P(X) \times P(Y)$ (where X and Y are independent), marginals and normalization involve integrals e.g. $\int_{\Omega_X} P(X=x)\,dx = 1$,

and conditioning requires dividing distributions e.g. $P(X|Y) = P(X,Y) \times (P(Y))^{-1}$. While these arithmetic computations suffice for probabilistic Bayesian networks, they are inappropriate for modeling in non-probabilistic ML. However, we note that they have the properties of an abstract semifield:

Definition. A semifield $\mathbf{S}$ consists of a set S with two associative binary operations $\otimes, \oplus$ and corresponding identities $i_\otimes, i_\oplus$ such that $(S, \oplus, i_\oplus)$ is a commutative monoid, $(S \setminus \{i_\oplus\}, \otimes, i_\otimes)$ is a group, $\otimes$ distributes both to the left and right over $\oplus$, and, $i_\oplus$ annihilates $\otimes$, e.g. $i_\oplus \otimes s = s \otimes i_\oplus = i_\oplus$ for all $s \in S$.

To exploit this abstraction for our purposes, a straightforward but rigorous formalization of the above operations are given by *Markov categories*, a special class of symmetric monoidal categories (upon which string diagrams can be based [3]):

Definition. A *Markov category* $\mathbf{C}$ with morphisms f and objects X, Y, is a category with symmetric monoidal bifunctor $(\times, 1)$ in which every object X is augmented with a commutative comonoid structure given by a comultiplication (copy) $\Delta_X : X \to X \times X$ and counit (terminal, delete) $1_X : X \to 1$. Additionally, these must satisfy the commutative comonoid equations as well as compatibility with the monoidal structure.

Such categories have recently been proposed for structural causal models and probabilistic causal adjustment [1]. Morphisms in these categories are *Markov kernels* (i.e. conditional probabilities such as $P(X|Y)$), and morphism composition is formally equivalent to matrix multiplication [3]. We define objects as arbitrary sets X, Y and morphisms are maps $f : X \to D(Y)$ in which $D(Y) = S^Y$, i.e. the set of all maps $f : Y \to S$. Morphism composition is defined by the underlying semifield. Joint states are matrix Kronecker products $P(X|Y) \times P(U|V)$. Normalization for disintegration is computed using the terminal morphism 1_X and division. We also consider affine Markov categories satisfying the natural transformation $1_X \cdot f = 1_Y$.

We show that a functor constructed using a semifield homomorphism can be used to map between Markov categories where morphism compositions, marginalization, products and disintegrations are computed over an arbitrary semifield. In particular, we show a functor mapping between the usual probability and *min-plus* $\mathbf{S}_\infty = (\mathbb{R}^+, \min, +, \infty, 0)$ and *softmin-plus* $\mathbf{S}_1 = (\mathbb{R}^+, -\ln(e^{-x} + e^{-y}), +, \infty, 0)$, semifields. These algebras are particularly important for ML/AI, because they are the usual context in which *negative log-likelihood* and *log-posterior*-based models are formulated. To illustrate, a string diagram representation of a non-probabilistic causal model over the $\mathbf{S}_\infty$ semifield, can be understood as the composition and (parallel) summation of propagating "prediction errors" and "inferential biases" through a machine learning algorithm, and minimization marginalizes out results which are irrelevant to the next stage of algorithm computation. The softmin-plus semifield behaves much like min-plus, except that, additionally, the entire string diagram model is fully differentiable, so that gradient-based parameter estimation in the algorithm can be performed easily using modern techniques such as automatic differentiation [5].

Furthermore, the functor translates, into these machine learning semifields, all causal adjustment operations in the usual probability semifield. This enables new capabilities in machine learning such as a novel form of *causally-adjusted ML model parameter inference*. We illustrate this idea in the case of general string diagrams for back-door and front-door adjustment, formulated using a novel point state variation of the causal intervention *cut functor* given in Jacobs et al. [4].

References

[1] Kenta Cho and Bart Jacobs. "Disintegration and Bayesian inversion via string diagrams". In: *Mathematical Structures in Computer Science* 29.7 (2019), 938â971. DOI: 10.1017/S0960129518000488.

[2] Alex J. DeGrave et al. "AI for radiographic COVID-19 detection selects shortcuts over signal". In: *Nat Mach Intell* 3 (May 2021), 610â619. ISSN: 2522-5839. DOI: 10.1038/s42256-021-00338-7.

[3] Tobias Fritz. "A synthetic approach to Markov kernels, conditional independence and theorems on sufficient statistics". In: *Advances in Mathematics* 370 (2020), p. 107239. ISSN: 0001-8708. DOI: https://doi.org/10.1016/j.aim.2020.107239. URL: https://www.sciencedirect.com/science/article/pii/S0001870820302656.

[4] Bart Jacobs et al. "Causal inference via string diagram surgery: A diagrammatic approach to interventions and counterfactuals". In: *Mathematical Structures in Computer Science* 31.5 (2021), 553â574. DOI: 10.1017/S096012952100027X.

[5] Arthur Mensch and Mathieu Blondel. "Differentiable Dynamic Programming for Structured Prediction and Attention". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, Oct. 2018, pp. 3462–3471.

[6] Judea Pearl. *Causality*. 2nd ed. Cambridge University Press, 2009. DOI: 10.1017/CBO9780511803161.

[7] Thomas S. Richardson et al. "Nested Markov Properties for Acyclic Directed Mixed Graphs". In: *arXiv e-prints*, arXiv:1701.06686 (Jan. 2017).