Causal Inference & Deep Learning

Notes on Statistical Learning & Causal Learning

MIT IAP 2018 Compiled by Max W. Shen

Notes are heavily derived from references and textbooks.

Machine learning has been a powerful tool for modeling statistical associations in data. However, causality as a concept cannot be discussed properly with solely the language of statistical learning theory. Here, the relationship between statistical reasoning to causal reasoning is discussed.

1 Statistical Reasoning and Learning

A common aim in statistical learning is to infer the correlative structure between random variables using observational data. Consider the regime of two random variables with a latent underlying joint distribution $p(x, y)$ or $P_{X,Y}$ from which observations $(x_1, y_1), ..., (x_n, y_n)$ are sampled i.i.d. from $P_{X,Y}$.

In a supervised machine learning setting, $x_i \in X$ could be understood as inputs or covariates and $y_i \in Y$ as outputs or labels. If $Y = \mathbb{R}$, regression could be of interest:

$$f(x) = \mathbb{E}[Y|X = x]. \tag{1}$$

And if $Y = \{\pm 1\}$, binary classification could be of interest:

$$f(x) = \arg\max_{y \in Y} P(Y = y|X = x). \tag{2}$$

In unsupervised machine learning, a task of interest could be generating realistic images x and captions y using deep neural nets. Such a task involves inferring a joint distribution $\hat{P}_{X,Y}$ from real images and captions such that the sampled images and captions are realistic:

$$(\hat{x}, \hat{y}) \sim \hat{P}_{X,Y}. \tag{3}$$

All of these tasks involve inferring information that is captured by the joint distribution $P_{X,Y}$. These examples highlight the point that many statistical learning tasks can be solved by first inferring the joint distribution $P_{X,Y}$. The joint distribution is an object that contains all statistical information regarding its random variables.

However, the task of inferring the latent joint distribution $P_{X,Y}$ from the empirical distribution $P^N_{X,Y}$ is an **ill-posed inverse problem**. As an example, consider the case where $X = \mathbb{R}$. Then, with only a finite dataset, there are infinitely many joint distributions that perfectly fit the empirical dataset but behave differently on unobserved x, e.g., with different predictions $p(y|x)$.

In statistical learning, regularization places additional assumptions on the ill-posed inverse problem in order to attempt to learn a statistical model that generalizes well to held-out data.

2 Causal Reasoning and Learning

In standard statistical settings, the joint distribution is king – it is an object that contains all information regarding its random variables. However, in causal settings, it is dethroned. A causal model entails a joint distribution, but a causal model also contains information not contained in its joint distribution.

Informally, this claim can be understood by noting:

- Correlation does not imply causation – that is, correlation between two random variables can be explained by direct causation in either direction as well as indirect causation from an unobserved confounder
- $p(a, b) = p(a|b)p(b)$ and $p(b|a)p(a)$, with different causal interpretations

In causal learning, a joint distribution can be equally compatible with multiple causal models, but each causal model can produce different causal conclusions on, for example, unobserved causal interventions.

Therefore, just as statistical learning aims to solve an ill-posed inverse problem, in general, causal learning also aims to solve an **ill-posed inverse problem**. An important implication is that attempting to infer a causal model from observational data requires two inverse problems to be solved: first, a joint distribution must be inferred from finite data; second, a causal model must be inferred from a joint distribution. *The task of inferring a causal model from observational data in the simple setting of two random variables, known as cause-effect discovery, will be discussed further by MS.*

In many real-world settings, some causal prior knowledge is available and the underlying causal model can be assumed. One important example is clinical trials where an observed treatment is assumed to cause an observed patient outcome. Clinical trials aim to estimate the utility of a treatment. In order to do so, they must consider **counterfactuals**:

> "What would have happened to patient x if they had received treatment 1 instead of treatment 0?"

The task of counterfactual inference will be discussed further by FJ.

2.1 Example: MNIST, handwritten digits

Consider the MNIST handwritten digit dataset, consisting of images X and labels Y of handwritten numerical digits from 0 to 9. Statistical models that can capture the correlation

between X and Y are capable of predicting the digit label from the image at performance above random guessing.

Consider two causal models, both of which are capable of generating the same statistical joint distribution $P_{X,Y}$.

> Causal Model #1:
> A class label is generated, then given to a human writer who uses it to produce a corresponding handwritten digit image with some noise (e.g., in perceiving the label or in executing motor functions). The class label and the handwritten image are recorded into the dataset.

Here, the class label y is generated from a distribution $p(y)$, and the image x is generated given the label, from the distribution $p(x|y)$.

Consider an intervention on the process generating the label y. Under this causal model, the process is $p(y)$. If the intervention modifies $p(y)$ to a distribution such that $p(y = \text{“9”}) = 1$, then the human writer will only see “9” labels, and the causal generative process will produce only (x_i, y_i) pairs where y_i is always “9”. The intervention impacts both the labels and the images.

Consider instead an intervention on the process generating the handwritten image x. Under this causal model, this process is $p(x|y)$. If the intervention modifies $p(x|y)$ to a distribution such that $p(x = \text{“7”}) = 1$, then labels will not be impacted, but the handwritten image will always be a “7”.

> Causal Model #2:
> The human writer decides on a digit, then records the label with some noise (e.g., in data entry) and also handwrites the digit with some noise (e.g., in motor function). The class label and the handwritten image are recorded into the dataset.

Here, the image x and the class label y are both functions of the writer's intention, a latent variable z with some generative distribution $p(z)$. The image is generated by $p(x|z)$ and the label by $p(y|z)$.

Consider an intervention on the process generating the label y. Under this causal model, this process is $p(y|z)$, and intervening on it will impact the labels but will not impact the associated handwritten images x.

Similarly, consider an intervention on the process generating the image x. Under this causal model, this process is $p(x|z)$, and intervening on it will impact the images but will not impact the associated labels y.

In sum, multiple causal models can be equally compatible with the statistical joint distribution, but can assert differing conclusions regarding causal interventions.

3 Structural Causal Models

The previous example highlights the usage of directed acyclic graphs where nodes are random variables and edges describe causal generative processes from causes to effects. This language for describing causal models can be generalized to arbitrarily many random variables and formalized as a "structural causal model" (SCM) as follows:

Definition: A structural causal model $C = (S, P_N)$ consists of a collection S of d structural assignments

$$X_j \leftarrow f_j(PA_j, N_j), \tag{4}$$

for $j = 1, ..., d$, where $PA_j \subseteq \{X_1, ..., X_d\} \backslash \{X_j\}$ are called parents of X_j; and a joint distribution $P_N = P_{N_1,...,N_d}$ over the noise variables, which is required to be jointly independent, that is, P_N is a product distribution.

The graph G of an SCM is obtained by creating one vertex for each X_j and drawing directed edges from each parent in PA_j to X_j. This graph is assumed to be acyclic.

Here, "$\leftarrow$" is defined and read as "is caused by", in contrast to algebraic equality "=", read as "is equal to". The significance of causal assignment (akin to the "=" symbol in programming languages) is in reasoning about causal interventions.

Note that the "$\rightarrow$" and "$\leftarrow$" symbols are also used to describe functions mapping from their input space to their output space within these notes, but context should be sufficient for identifying the intended meaning.

4 Interventional Distributions

Different causal models can generate the same joint distribution but offer different conclusions to causal interventions. Interventions in SCMs can be described using the "do" operator, introduced by Judea Pearl.

Consider an SCM $C = (S, P_N)$ and its entailed joint distribution P_X^C. An intervention replaces one (or several) of the structural assignments to obtain a new SCM $\tilde{C}$. For example, when the causal assignment for X_k is replaced by

$$X_k \leftarrow \tilde{f}(\tilde{PA}_k, \tilde{N}_k), \tag{5}$$

the entailed joint distribution of the new SCM is called an **interventional distribution**. The new entailed joint distribution is denoted by

$$P_X^C = P_X^{C;do(X_k \leftarrow \tilde{f}(\tilde{PA}_k, \tilde{N}_k))}. \tag{6}$$

The set of noise variables in $\tilde{C}$ now contains both some "new" Ns and some "old" Ns, all of which are still required to be jointly independent.

When $\tilde{f}(\tilde{PA}_k, \tilde{N}_k)$ puts a point mass on a real value a, the interventional distribution is written

$$P_X^{C;do(X_k \leftarrow a)}, \tag{7}$$

and called a "hard intervention". Hard interventions induce statistical independence between the original causes and effect. In contrast, "soft interventions" preserve some or all of the original causes on an effect.

The new SCM $\tilde{C}$ is still required to induce an acyclic graph. Thus, the set of allowed interventions depends on the graph induced by C.

4.1 Calculating interventional distributions

Calculating the impact of an intervention in a causal model can be performed from observational data alone when properly controlling for confounding variables. When this is the case, the interventional distribution is identifiable.

Back-Door Criterion: In an SCM, we wish to calculate the impact that an intervention on X has on Y: $p(Y|do(X = x))$. A set of variables S in the SCM satisfies the back-door criterion and is sufficient for identifiability, giving:

$$p(Y|do(X = x)) = \sum_s p(Y|X = x, S = s)p(S = s) \tag{8}$$

if the two following conditions are satisfied:

i S blocks every path from X to Y that has an arrow into X.

ii No node in S is a descendant of X.

Disjunctive Cause Criterion: In an SCM, we wish to calculate the impact that an intervention on X has on Y: $p(Y|do(X = x))$. A set of variables S in the SCM satisfies the disjunctive-cause criterion and is sufficient for identifiability, giving:

$$p(Y|do(X = x)) = \sum_s p(Y|X = x, S = s)p(S = s) \tag{9}$$

if the following condition is satisfied:

i S is the set of all direct causes of X, Y, or both.

These are two sufficient criteria for identifiability given observational data on an assumed causal graph. Others exist, including the front-door criterion which is built on the back-door criterion.

In reality, the true causal graph is not known. The existence of unobserved confounding variables breaks identifiability of interventional distributions. As a result, a common assumption causal inference theory is the strong assumption of "no unobserved confounders". Without adequate control of confounding, the impact of an intervention can still be bounded. Research into this area is sometimes called *partial identification*.

The back-door criterion tends to require smaller sets of variables to control for than the disjunctive cause criterion. However, the disjunctive cause criterion can be useful in practice because it can be easier to have high confidence on direct causes in the true underlying causal graph.

5 References

Notes on Causal Inference, Lecture 31. CMU 36-350 "Data Mining", 2009. http://www.stat.cmu.edu/∼cshalizi/350/lectures/31/lecture-31.pdf

Peters, J. (2017) Elements of Causal Inference: Foundations and Learning Algorithms, Final Textbook Draft. https://mitpress.mit.edu/books/elements-causal-inference (Available for free, search for "This is an open access title" on left side of webpage)

Roy, J. A. (2018) A Crash Course in Causality: Inferring Causal Effects from Observational Data - Disjunctive Cause Criterion. Coursera.