Case Study : Causal Inference

Mohammed FELLAJI - Ahmed BEN AISSA

Supervisor : Frédéric PENNERATH †

†Associate professor at CentraleSupélec

CentraleSupélec

Introduction

One of the most challenging questions in every problem is the one related to understanding the reason(s) why an action happened and whether or not we can explain it with the information we have at our disposal. Determining the relationships between the inputs can be done on different levels : when recording the data or even once the data collection is done. Unlike correlation, causation cannot be extracted only from the distribution of the data. In this case study, we introduce the main concepts in the study of causality, we formulate the underlying mathematical problem, and then we offer a road map to solve this problem.

Definitions

Here is some of the most important definitions [2, 3] :

- **Associational concept** : any relationship that can be defined in terms of a joint distribution of observed variables. For example: regression, correlation.
- **Causal concept** : any relationship that cannot be defined from the distribution alone. One should also rely on causal assumption that explains the different relationships. For example: randomisation, confounding.

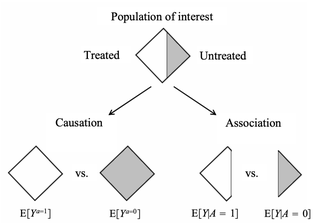

Fig. 1: a visualisation of the difference between causation and association

- **Counterfactuals** : the unobserved outcome. It is the answer to the question : what if?
- **Confounders** : an unmeasured variable that influences both the supposed cause and the supposed effect.

Notation

Here is a list of some useful notations :

- Y_i : random variable of the outcome for the "unit" i (a unit is a physical object, for example, a person, at a particular point in time [4]),
- N : the number of units,
- Y : $(Y_1, Y_2, ..., Y_n)$,
- A : treatment (a treatment is an action that can be applied or withheld from the unit [4]),
- Y_i^a : controlled treatment

Metrics and Interpretations

In the case of a treatment (A) that has only 2 possible values {0,1}, we can define these metrics:

- **Individual level Causal Effect (ICE)** : the difference between an individual's two potential outcomes.

$$ICE = \delta = Y_i^{a=1} - Y_i^{a=0}$$

- **Average Causal Effect (ACE)** : it is a mesure of causation; it determines if the treatment has a causal effect on the outcome or not.

$$ACE = E[\delta] = E[Y^{a=1}] - E[Y^{a=0}]$$

- **Standard estimator (S*)** : it is a mesure of association; it is computed from the treatment and control groups.

$$S^* = E[Y^{a=1}|A = 1] - E[Y^{a=0}|A = 0]$$

Fundamental Problem of Causal Inference

Because one can never observe both potentiel outcomes (the outcome under different treatments), it is impossible to mesure directly the causal effects (ACE and ICE). This is what Holland called as the "fundamental problem of causal inference" [1].

This can be seen as missing data problem, and cannot be resolved. Thus, causal inference aims to use several methods to obtain the best approximate measure of ACE , through randomized experimentation and the study of confounders and counterfactuals.

Randomised experimentation

A randomised experiment is when the investigator carried out the action of interest and it was randomised because the decision to act on any study subject was made by a random device.

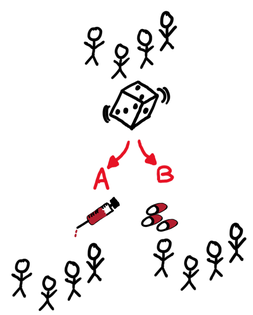

Fig. 2: an illustration of a randomised experiment

In the case of randomised experiment, we have the following :

- $E[Y^{a=1}|A = 1] = E[Y^{a=1}|A = 0] = E[Y^{a=1}]$ (1)
- $E[Y^{a=0}|A = 1] = E[Y^{a=0}|A = 0] = E[Y^{a=0}]$ (2)
- from the equations (1) and (2) we can prove that : $ACE = S^*$

Causal discovery

A traditional way to discover causal relations is to use interventions or randomized experiments, which is, however, in many cases of interest too expensive, too time-consuming, unethical, or even impossible. Therefore, inferring the underlying causal structure from purely observational data, or from combinations of observational and experimental data, has drawn much attention in various disciplines. With the rapid accumulation of huge volumes of data, it is necessary to develop automatic causal search algorithms that scale well.

Methods: From classical to machine learning

In order to put to use the knowledge gathered through the randomized experimentation and test the hypothesis formulated through causal discovery, we need methods that allow for counterfactual estimation and confounding factors elimination.
Many classical methods exist in the litterature :

- **Re-weighting** : To assign a different weight to each sample in the observational dataset, in order to better reflect on the population and eliminate any confounding effect of a selection bias.
- **Matching** : After defining distance between samples to reflect on their similarity, we study the outcomes of neighboring samples, which can reveal confounders. We can also estimate counterfactuals by inferring from similar sample points.
- **Tree-based methods** : Tree-based methods refer to the use of decision-trees on an elementary level. A decision- tree allows for a partition of the sample space. The criterias used to partition the data can allow the detection of confounders and counterfactuals, and thus offer a better estimate of the causal effect.

These classical methods offer a good starting point, but aren't suited to the databases of today, containing millions of samples, with data that is so deeply intertwines that these methods simply aren't realistic. It's where machine learning steps in, by putting to use the massive calculatory powers of modern computers in order to better tackle the causal inference problem.
Doubly Robust Regression, **Subspace learning**, and **Deep Representation learning** all offer a promising set of models to tackle selection bias, underlying confounders and missing counterfactuals, which in turn allows for a better estimation of the average causal effect.

References

[1] Paul W Holland. "Statistics and causal inference". In: *Journal of the American statistical Association* 81.396 (1986), pp. 945–960.

[2] Judea Pearl. "The mathematics of causal relations". In: *Causality and Psychopathology: Finding the Determinants of Disorders and their Cures (P. Shrout, K. Keyes and K. Ornstein, eds.). Oxford University Press, Corvallis, OR* (2010), pp. 47–65.

[3] Judea Pearl et al. "Causal inference in statistics: An overview". In: *Statistics surveys* 3 (2009), pp. 96–146.

[4] Donald B Rubin. "Causal inference using potential outcomes: Design, modeling, decisions". In: *Journal of the American Statistical Association* 100.469 (2005), pp. 322–331.