The Symphony of the Eighteenth Century and the Modern Times: Chain Rule, Bayesian Networks, and Hume - Probability, Correlation, and Causation


Jingming 2025-03-20 14:50:30 Edited United States

Let's start with a motivation: What is the relationship between the two events **rain (Rain, R)** and **traffic jam (Traffic, T)** ?

If, like me, you've observed that traffic jams always occur when it rains, and are about to respond without a second thought, "Cause and effect!", then you must be an observant driver, but lack some forethought. Is the relationship really so obvious?

First, let's define some notation:

- $P(T)$: Probability of traffic jam
- $P(R)$: Probability of Rain
- $P(T|R)$: Conditional probability of traffic jam when it rains
- $P(R|T)$: Conditional probability of rain when there is a traffic jam
- $P(R,T)$: The probability of traffic jam and rain occurring at the same time

If we focus on $P(T,R)$, we can get two decompositions according to **the chain rule** :

$P(T,R) = P(T) \cdot P(R|T)$

$P(T,R) = P(R) \cdot P(T|R)$

However, according to **the Bayesian network** proposed by **Judea Pearl (1936-**), they model two opposite directions:

$P(T,R) = P(T) \cdot P(R|T)$ ==> T->R

$P(T,R) = P(R) \cdot P(T|R)$ ==> R->T

Both of these modeling approaches can mathematically represent the same joint probability distribution $P(T,R)$.

This result reveals a disturbing fact: Only **Observational data** , we cannot determine the direction of causality. Whether it is T->R or R->T, the Bayesian network is compatible with the data.

However, in reality, these two causal relationships are obviously not equivalent. It is common sense that rain causes traffic jams, but it is absurd that traffic jams cause rain. This contradiction reveals that **observational data** It is adept at describing correlation, but struggles with causality.

David Hume (1711-1776) expressed profound skepticism about causality in his Treatise of Human Nature, writing:

> "We have no other notion of cause and effect, but that of certain objects, which have been always conjoin'd together." Our understanding of causality is limited to the fact that certain objects are always connected together.

Hume believed that causality is not a logical necessity, but a habitual association based on experience. We cannot perceive some mysterious "causal force", but can only observe the sequence of events and their frequent co-occurrence. In the example of rain and traffic jams, Hume would probably say this: the reason why we think that rain causes traffic jams may be just because the two often appear at the same time, rather than truly understanding the necessary connection between them.

In response to Hume's skepticism, modern science introduced **experimental data** or **mathematically simulated experimental intervention data** , **with randomized controlled trials (RCT)** and **do-calculus** being the two corresponding methods.

Randomized controlled trials (RCTs) experiment Imagine an ideal experiment that uses artificially forced rain and artificially forced rain avoidance to isolate causal effects.

- Experimental group: forced rain (R = 1)
- Control group: forced no rain (R = 0)

By comparing the traffic jam probabilities of the two groups, we can calculate the average causal effect (ATE):

ATE = P(T = 1 | R = 1) - P(T = 1 | R = 0)

If ATE > 0, it means that rain does increase the probability of traffic jams, thus supporting the causal direction of R->T.

When experiments are not feasible, **the do-operator** proposed by **Judea Pearl** provides a Experimental intervention with mathematical simulations method. Assume that there is a confounding variable Z (such as holidays), which affects both rain and traffic jams. We can use the backdoor adjustment formula to eliminate its interference:

P(Y=y|do(X=x)) = ∑_z P(Y=y|X=x,Z=z) · P(Z=z)

By controlling Z, we are able to estimate the direct effect of rain on traffic jams rather than relying solely on correlations.

RCT and do-operator have their own advantages. The former requires real experiments, while the latter can intervene through mathematical simulation experiments when experiments are not possible. Compared with the limitations of observational data, these two methods allow us to go beyond correlation and approach causality.

However, even with experimental data or simulation interventions, can we fully respond to Hume's skepticism?

Hume would likely partially agree that experiments can strengthen causal beliefs because they actively create a “traffic jam after rain” pattern and reduce confusion.

But Hume might have asked further: Even if the experiment shows that rain causes traffic jams, why do we believe that this causal relationship will still hold true in the future? The experiment only increases the frequency and consistency of experience, and still cannot prove the "necessity" of causality - it is still a belief.

Nevertheless, experimental data or experimental intervention data from mathematical simulations is indeed a practical response to Hume - we admit that causality cannot be absolutely proven, but through experiments and models, we can predict and intervene more reliably. Although these methods cannot prove the ultimate inevitability of causality, they have more or less soothed our wandering hearts, and we no longer always feel like reeds in the wind or duckweed in the water.