Summary for Introduction to Probabilities, Graphs, and Causal Models[Causality Chap.1 (Judea Pearl)]

Shirley Wu
2021.2.1

1 Basic Concepts

- Three basic axioms of probability calculus:

$$0 \leq P(A) \leq 1$$

$$P(\text{ sure proposition }) = 1$$

$$P(A \text{ or } B) = P(A) + P(B) \text{ if } A \text{ and } B \text{ are mutually execlusive.}$$

- Law of total probability

$$P(A) = \sum_i P\left(A, B_i\right)$$

- Conditional probability

$$P(A \mid B) = \frac{P(A, B)}{P(B)}$$

  Combine the above two equalities, we have

$$P(A) = \sum_i P\left(A \mid B_i\right) P\left(B_i\right)$$

$$P(A \mid K) = \sum_i P\left(A \mid B_i, K\right) P\left(B_i \mid K\right)$$

- Bayesian Equation:

$$P(H \mid e) = \frac{P(e \mid H) P(H)}{P(e)}$$

  So far we can define probabilistic model as an encoding of information that permits us to compute the probability of every well-formed sentence S in accordance with the axioms of probability. Actually, any valid joint probability function can represent a complete probabilistic model by uniquely determining the probability of every elementary event in the domain as well as satisfying the axioms.

  Let's further portray the essence of Bayes's rule,

  Defining the **prior odds** on H as

$$O(H) = \frac{P(H)}{P(\neg H)} = \frac{P(H)}{1 - P(H)}$$

and the **likelihood ratio** as

$$L(e \mid H) = \frac{P(e \mid H)}{P(e \mid \neg H)}$$

the **posterior odds**

$$O(H \mid e) = \frac{P(H \mid e)}{P(\neg H \mid e)} = \frac{P(e \mid H)P(H)}{P(e)} \frac{P(e)}{P(e \mid \neg H)P(\neg H)} = L(e \mid H)O(H)$$

where the prior odds O(H) measures the predictive or prospective support accorded to H by the background knowledge alone, while the likelihood ratio L(e | H) represents the diagnostic or retrospective support given to H by the evidence actually observed.

- **Random variables** are brought to present a certain issue in real world. And by a variable we will mean an attribute, measurement or inquiry that may take on one of several possible outcomes, or values, from a specified domain.
- **Statistical magnitude** like the mean or expected value and variance of X; we can also define similar statistical magnitudes (expectation and co-varience) for multi-variable.

$$E[g(X,Y)] \triangleq \sum_{x,y} g(x,y)P(x,y)$$

$$\sigma_{XY} \triangleq E[(X - E(X))(Y - E(Y))]$$

and which is often normalized to yield the correlation coefficient

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

and the regression coefficient (of X on Y)

$$r_{XY} \triangleq \rho_{XY} \frac{\sigma_X}{\sigma_Y} = \frac{\sigma_{XY}}{\sigma_Y^2}$$

- **Conditional independence** is a central notion in causal modelling, and it is defined as:

$$(XY \mid Z)_P \quad \text{iff} \quad P(x \mid y, z) = P(x \mid z)$$

along with its properties

Symmetry: $(XY \mid Z) \Longrightarrow (YX \mid Z)$
Decomposition: $(XYW \mid Z) \Longrightarrow (XY \mid Z)$
Weak union: $(XYW \mid Z) \Longrightarrow (XY \mid ZW)$
Contraction: $(XY \mid Z)\&(XW \mid ZY) \Longrightarrow (XYW \mid Z)$
Intersection: $(XW \mid ZY)\&(XY \mid ZW) \Longrightarrow (XYW \mid Z)$

Weak union and Contraction reveal that regardless of the engagement of irrelevant information, what was relevant remains relevant, and same does the irrelevant one.

- **Marginal independence** is defined as

$$(XY \mid \varnothing) \text{ iff } P(x \mid y) = P(x) \quad \text{whenever } P(y) > 0$$

2 Graphs for Representing Causality

Bayesian networks:
a directed acyclic graph representing causal or temporal relationships, and the core is that, knowing the values of other preceding variables is redundant once we know the values pa_j of the parent set PA_j, which is also called Markovian Parents.

Markovian Parents:
Let $V = \{X_1, \ldots, X_n\}$ be an ordered set of variables, and let $P(v)$ be the joint probability distribution on these variables. A set of variables PA_j is said to be Markovian parents of X_j if PA_j is a minimal set of predecessors of X_j that renders X_j independent of all its other predecessors. In other words, PA_j is any subset of $\{X_1, \ldots, X_{j-1}\}$ satisfying

$$P\left(x_j \mid pa_j\right) = P\left(x_j \mid x_1, \ldots, x_{j-1}\right) \tag{1}$$

The way we construct a BN is by recursion with the assistance of some screening conditions. We therefore conclude that a necessary condition for a DAG G to be a Bayesian network of probability distribution P is for P to admit the product decomposition dictated by G, as

$$P\left(x_1, \ldots, x_n\right) = \prod_i P\left(x_i \mid pa_i\right) \tag{2}$$

Markov Compatibility:
If a probability function P admits the factorization of Eq.2 relative to DGA G, we say that G represents P, that G and P are compatible, or that P is Markov relative to G.
It is a necessary and sufficient condition for a DAG G to explain a body of empirical data represented by P, that is, to describe a stochastic process capable of generating P.

d-Separation:
A path p is said to be d -separated (or blocked) by a set of nodes Z if and only if

1. p contains a chain $i \to m \to j$ or a fork $i \leftarrow m \to j$ such that the middle node m is in Z, or

2. p contains an inverted fork (or collider) $i \to m \leftarrow j$ such that the middle node m is not in Z and such that no descendant of m is in Z.

A set Z is said to d-separate X from Y if and only if Z blocks every path from a node in X to a node in Y.
Figuratively, for the 1st condition, conditioning on m appears to "block" the flow of information along the path, since learning about i has no effect on the probability of j, given m. While for the 2nd condition, if the two extreme variables are (marginally) independent, they will become dependent (i.e. connected through unblocked path) once we condition on the middle variable (i.e., the common effect) or any of its descendants.

Probabilistic Implications of d -Separation:
If sets X and Y are d -separated by Z in a DAG G, then X is independent of Y conditional on Z in every distribution compatible with G. Conversely, if X and Y are not d -separated by Z in

a DAG G, then X and Y are dependent conditional on Z in at least one distribution compatible with G.
The probabilistic notion of conditional independence and the graphical notion of d-separation are not equivalent, unless they satisfy some compatibility laws.

Equivalence of d -Separation between DAG and probilistic distribution:
For any three disjoint subsets of nodes (X, Y, Z) in a DAG G and for all probability functions P, we have:

(i) $(XY \mid Z)_G \Longrightarrow (XY \mid Z)_P$ whenever G and P are compatible; and

(ii) if $(XY \mid Z)_P$ holds in all distributions compatible with G, it follows that $(XY \mid Z)_G$

Ordered Markov Condition:
A necessary and sufficient condition for a probability distribution P to be Markov relative a DAG G is that, conditional on its parents in G, each variable be independent of all its predecessors in some ordering of the variables that agrees with the arrows of G . A consequence of this theorem is an order-independent criterion for determining whether a given probability P is Markov relative to a given DAG G.

Parental Markov Condition:
A necessary and sufficient condition for a probability distribution P to be Markov relative a DAG G is that every variable be independent of all its nondescendants (in G), conditional on its parents. (We exclude X_i when speaking of its "nondescendants.")
Ordered Markov Condition and Parental Markov Condition have specified the condition for distribution P to be Markov relative a DAG G.

Observational Equivalence:
Two DAGs are observationally equivalent if and only if they have the same skeletons and the same sets of v-structures, that is, two converging arrows whose tails are not connected by an arrow (Verma and Pearl 1990).
Observational Equivalence places a limit on our ability to infer directionality from probabilities alone. Two networks that are observationally equivalent cannot be distinguished without resorting to manipulative experimentation or temporal information. It nevertheless gives us a technical direction to distinguish two networks.
Given a task of prediction, it is necessary to find a coherent interpretation of incoming observations that is consistent with both the observations and the prior information at hand. Some resort to Baysian Networks, which boils down this task to the computation of $P(y|x)$.
The challenge, however, lies in performing these computations efficiently and within the representation level provided by the network topology. The earlier methods adpoted a message-passing architecture but were limited to trees, while join-tree propagation and cut-set conditioning methods were try to extend this tree propagation to general networks.

3 Causal Baysian Networks

The interpretation of direct acyclic graphs as carriers of independence assumptions does not necessarily imply causation; in fact, it will be valid for any set of recursive independencies along any

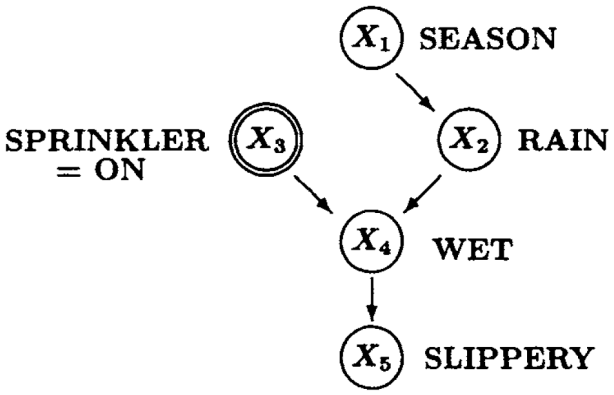

Figure 1: Exmaple 1.1

ordering of the variables, not necessarily causal or chronological. The advantages of building DAG models around causal rather than associational information are (i) the judgments required in the construction of the model are more meaningful, more accessible and hence more reliable. (ii) the ability to represent and respond to external or spontaneous changes.

Intervention do is an action rather then observation. Note the difference between the action $do(X_3 = On)$ and the observation $X_3 = On$. The effect of the latter is obtained by ordinary Bayesian conditioning, that is, $P(x_1, x_2, x_4, x_5|X_3 = On)$, while that of the former by conditioning a mutilated graph with the link removed.

Causal Bayesian Network:
Let $P(v)$ be a probability distribution on a set V of variables, and let $P_x(v)$ denote the distribution resulting from the intervention $do(X = x)$ that sets a subset X of variables to constants x. Denote by $\boldsymbol{P}_*$ the set of all interventional distributions $P_x(v), X \subseteq V$ including $P(v)$, which represents no intervention (i.e. $, X = \phi$). A $DAGG$ is said to be a causal Bayesian network compatible with $\boldsymbol{P}_*$ if and only if the following three conditions hold for every $P_x \in P_*$:

(i) $P_x(v)$ is Markov relative to G

(ii) $P_x\left(v_i\right) = 1$ for all $V_i \in X$ whenever v_i is consistent with $X = x$

(iii) $P_x\left(v_i \mid pa_i\right) = P\left(v_i \mid pa_i\right)$ for all $V_i \notin X$ whenever pa_i is consistent with $X = x$
i.e., each $P\left(v_i \mid pa_i\right)$ remains invariant to interventions not involving V_i

This understanding of causal influence permits us to see precisely why, and in what way, causal relationships are more "stable" than probabilistic relationships. We expect such difference in stability because causal relationships are ontological, describing objective physical constraints in our world, whereas probabilistic relationships are epistemic, reflecting what we know or believe about the world.
There are generally two points:

(i) in contrast to probabilistic relationships, causal relationships remain invariant to changes in the mechanism that governs the causal variables.

(ii) causal models need not encode behavior under intervention but instead aim primarily to provide an "explanation" or "understanding" of how data are generated. Regardless of what use

is eventually made of our "understanding" of things, we surely would prefer an understanding in terms of durable relationships, transportable across situations, over those based on transitory relationships.

4 Functional Causal Models

Causal relationships in some models are expressed in the form of deterministic, functional equations, and probabilities are introduced through the assumption that certain variables in the equations are unobserved, which reflects Laplace's (1814) conception of natural phenomena, indicating that nature's laws are deterministic and randomness surfaces owing merely to our ignorance of the underlying boundary conditions.
In contrast, all relationships in the definition of causal Bayesian networks were assumed to be inherently stochastic and thus appeal to the modern (i.e., quantum mechanical) conception of physics, according to which all nature's laws are inherently probabilistic and determinism is but a convenient approximation.
The author adopts the first concepts due to its generality and coincidence towards human intuition as well as applicability for some problems (e.g. counterfactual).
In its general form, a functional causal model consists of a set of equations of the following form,

$$\begin{aligned} x_i &= f_i\left(pa_i, u_i\right), \quad i = 1, \ldots, n \\ x_i &= \sum_{k \neq 1} \alpha_{ik} x_k + u_i, \quad i = 1, \ldots, n \end{aligned} \tag{3}$$

And we can compare it with the formulation of the linear structural equation models (SEMs). The 1st equation specifies what value nature would assign to X_i in response to every possible value combination that (PA_i, U_i) might take on. Mathematically, the distinction between structural and algebraic equations is that any subset of structural equations is, in itself, a valid structural model – one that represents conditions under some set of interventions. Also, the difference between controlled and observed changes is essential for the correct interpretation of structural equation models.

4.1 Probabilistic Predictions in Causal Models

Given a causal model , if we draw an arrow from each member of PA_i toward X_i, then the resulting graph G will be called a causal diagram. If the causal diagram is acyclic, then the corresponding model is called semi-Markovian and the values of the X variables will be uniquely determined by those of the U variables. Under such conditions, the joint distribution $P(x_1, , x_n)$ is determined uniquely by the distribution $P(u)$ of the error variables. If, in addition to acyclicity, the error terms are jointly independent, the model is called Markovian.

Causal Markov Condition:
Every Markovian causal model M induces a distribution $P\left(x_1, \ldots, x_n\right)$ that satisfies the parental Markov condition relative the causal diagram G associated with M; that is, each variable X_i is independent of all its nondescendants, given its parents PA_i in G (Pearl and Verma 1991).

The causal Markov condition implies that characterizing each child–parent relationship as a deterministic function, instead of the usual conditional probability $P(x_i|pa_i)$, imposes equivalent independence constraints on the resulting distribution and leads to the same recursive decomposition that characterizes Bayesian networks. Also, Druzdzel and Simon (1993) showed that, for every Bayesian network G characterized by a distribution P, there exists a functional model that generates a distribution identical to P, which indicates that we can regard functional models as just another way of encoding joint distribution functions.
The advantages for us to use functional model are

(i) it is invariant to parametric changes in the mechanisms represented by the functions fi and the distributions $P(u_i)$.

(ii) it yields a small number of parameters.

(iii) judgmental assumptions of conditional independence among observable quantities are simplified, because such assumptions are cast directly as judgments about the presence or absence of unobserved common causes.

(iv) instead of reestimating the entire model, it can just reassess a few local parameters involving the condition change.

4.2 Interventions and Causal Effects in Functional Models

When an intervention forces a subset X of variables to attain fixed values x, then a subset of equations is to be **pruned** from the functional model, one for each member of X, thus defining a new distribution over the remaining variables that characterizes the effect of the intervention and coincides with the truncated factorization obtained by pruning families from a causal Bayesian network.
However, while context itself varies with the intervention and so the conditional probabilities $P(x_i|pa_i)$ are altered in the process, the functional relationships $x_i = f_i(pa_i, u_i)$ remain invariant.

4.3 Counterfactuals in Functional Models

Assume that a given subject, Joe, has taken the treatment and died; we ask whether Joe's death occurred because of the treatment, despite the treatment, or regardless of the treatment. In other words, we ask for the probability Q that Joe would have died had he not been treated.
Given evidence e, to compute the probability of $Y = y$ under the hypothetical condition $X = x$ (where X is a subset of variables). It is easy to get the answer through inference on any causal model M using the following steps:

Step 1 (abduction): Update the probability $P(u)$ to obtain $P(u \mid e)$.

Step 2 (action): Replace the equations corresponding to variables in set X by the equations $X = x$

Step 3 (prediction): Use the modified model to compute the probability of $Y = y$.

Step 1 explains the past (U) in light of the current evidence e; step 2 bends the course of history (minimally) to comply with the hypothetical condition $X = x$; finally, step 3 predicts the

future (Y) based on our new understanding of the past and our newly established condition.
The author also constructed two tricky networks to illustrate why the stocastic causal networks are insufficient to give such answer — the knowledge of the actual process behind $P(y|x)$ is needed for the computation.
The inference process on functional causal models for counterfactual are clearly stated in the books, and we can now see why the stocastic causal networks will fail, since the $P(y|x)$ is the same for both models while the situations are different. Or, the U variables do not appear explicitly in stochastic models, we cannot apply step 1 so as to update $P(u)$ with the evidence e at hand.
So far, prediction, intervention, and counterfactuals have form a natural hierarchy of causal reasoning tasks. Prediction is the simplest of the three, requiring only a specification of a joint distribution function. The analysis of interventions requires a causal structure in addition to a joint distribution. Finally, processing counterfactuals is the hardest task because it requires some information about the functional relationships and/or the distribution of the omitted factors.

5 Causal Versus Statistical Terminology

Examples of statistical concepts are: correlation, regression, conditional independence, association, likelihood, collapsibility (?), risk ratio, odds ratio, propensity score, Granger's causality, and so on.
Examples of causal concepts are: randomization, influence, effect, confounding, exogeneity, ignorability, disturbance, spurious correlation, path coefficients, instrumental variables, intervention, explanation, and so on.
This author also summaries two mental barriers to causal analysis

(i) to commence causal analysis with untested, judgmental assumptions, and

(ii) to extend the syntax of probability calculus, constitute the two main obstacles to the acceptance of causal analysis among professionals with traditional training in statistics.

6 Exercise(in P34 of Causality[Pearl])

Question: Based on and barring sampling variations, the percentage Q of deceased subjects from the treatment group who would have recovered had they not taken the treatment precisely equals the percentage Q' of surviving subjects in the nontreatment group who will die if given treatment. Whereas Q is hypothetical, Q' is unquestionably testable.

My Solution:

Note tht there are two conditions:

$P(y|x) = 0.5$ for all x, y (a)

random uniform sampling (b)

$$Q = P(y = 1|do(A = 0))$$

Treatment Group		Nontreatment Group	
50%	50%	50%	50%
die	recover	die	recover
A	B	C	D

$$Q' = P(y = 0|do(D = 1))$$

Thus,

$$\begin{aligned} Q - Q' &=_{(b)} P(y = 1|do(C \cup D = 0)) - P(y = 0|do(A \cup B = 1)) \\ &=_{\text{consistency}} P(y = 1|C \cup D = 0) - P(y = 0|A \cup B = 1) \\ &=_{(a)} 0.5 - 0.5 = 0 \\ \Rightarrow Q &= Q' \end{aligned}$$