Causal inference cheat sheet

Author: Juvid Aryaman
Last compiled: October 5, 2020

1. Basic probability

- Law of total probability: $P(A) = \sum_i P(A, B_i)$ (a.k.a. marginalizing over B)
- Chain rule of probability: $P(A, B) = P(A|B)P(B)$
- Thus, $P(A) = \sum_i P(A|B_i)P(B_i)$
- Expectation: $E(g(X)) = \sum_x g(x)P(x)$
- Conditional mean: $E(X|Y) = \sum_x xP(x|y)$
- Variance: $\sigma_X^2 = E[(X - E(x))^2]$
- Covariance: $\sigma_{XY} = E[(X - E(X))(Y - E(Y))]$
- Correlation coefficient: $\rho_{XY} = \sigma_{XY}/(\sigma_X \sigma_Y)$
- Regression coefficient of X on Y: $r_{XY} = \rho_{XY}\sigma_X/\sigma_Y = \sigma_{XY}/(\sigma_Y^2)$ (for the equation $X = r_{XY}Y + c + \mathcal{N}(0, \sigma^2)$)
- Conditional independence: $(X \perp\!\!\!\perp Y|Z) \iff P(x|y, z) = P(x|z)$

The recursive decomposition of the joint distribution into parents which characterises Bayesian networks is

$$P(x_1, ..., x_n) = \prod_i P(x_i|pa_i) \tag{1.1}$$

1.1. d-separation (blocking) in Bayesian networks

A path p is d-separated (or blocked) by a set of notes 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 a 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

where an arrow $pa_j \to x_j$ denotes part of a directed acyclic graph (DAG) in which variables are represented by nodes and arrows are drawn from each node of the parent set PA_j towards the child node X_j.

Probabilistic implications of d-separation Consequently, if X and Y are d-separated by Z in a DAG G, then $(X \perp\!\!\!\perp Y|Z)$ in every distribution compatible with G. Conversely, if X, Y, and Z are *not* d-separated by Z in a DAG G then X and Y are dependent conditional on Z in almost all distributions compatible with G (assuming no parameter fine-tuning).

2. Functional causal models

A functional causal model consists of a set of equations of the form

$$x_i = f_i(pa_i, u_i), \quad i = 1, ..., n \tag{2.1}$$

where pa_i are the set of variables (parents) that directly determine the value of X_i and U_i represents errors (or "disturbances") due to omitted factors. When some disturbances U_i are judged to be dependent, it is customary to denote such dependencies in a causal graph with double-headed arrows. If the causal diagram is acyclic, then the corresponding model is called *semi-Markovian* and the values of the variables X are uniquely determined by those of the variables U. If the error terms U are jointly independent, the model is called *Markovian*.

Linear structural equation models obey

$$x_i = \sum_{k \neq i} \alpha_{ik} x_k + u_i, \quad i = 1, ..., n \tag{2.2}$$

In linear models, pa_i corresponds to variables on the r.h.s. of the above equation where $\alpha_{ik} \neq 0$.

2.1. Counterfactuals in functional causal models: An example

Consider a randomized clinical trial, where patients are/are not treated $X \in \{0, 1\}$. We also observe whether the patients die after treatment $Y\{0, 1\}$. We wish to ask the question: did the patient die *because of* the treatment, *despite* the treatment, or *regardless* of the treatment.

Assume $P(y|x) = 0.5$, and therefore $P(y, x) = 0.25$ for all x and y. We can write two models with the same joint distribution

Model 1 (treatment no effect):

$$x = u_1 \tag{2.3}$$

$$y = u_2 \tag{2.4}$$

$$P(u_1 = 1) = P(u_2 = 1) = \frac{1}{2} \tag{2.5}$$

Model 2 (treatment has an effect):

$$x = u_1 \tag{2.6}$$

$$y = xu_2 + (1 - x)(1 - u_2) \tag{2.7}$$

$$P(u_1 = 1) = P(u_2 = 1) = \frac{1}{2} \tag{2.8}$$

Let Q=fraction of deceased subjects from the treatment group who would not have died had they not taken the treatment. In model 1, $Q = 0$ since X has no effect on Y. In model 2, subjects who died ($y = 1$) and were treated ($x = 1$) must correspond to $u_2 = 1$. If $u_2 = 1$ then the only way for $y = 0$ is for $x = 0$. I.e. if you are a patient for whom $u_2 = 1$ then the only way not to die is to not take the treatment, so the treatment caused your death. So $Q = 1$.

Consequence 0: joint probability distributions are insufficient for counterfactual computation

Consequence 1: stochastic causal models are insufficient for counterfactual computation

Consequence 2: functional causal models are sufficient to define and compute counterfactual statements.

2.2. General method to compute counterfactuals

Given evidence $e = \{X_{obs}, Y_{obs}\}$, to compute probability of $Y = y$ under hypothetical condition $X = x$ apply the following steps:

1. Abduction: Update the probability of disturbances $P(u)$ to obtain $P(u|e)$
2. Action: Replace the equations corresponding to variables in the set X by the equations $X = x$
3. Prediction: Use the modified model to compute the probability $Y = y$.

3. Causal Bayesian networks

Given two disjoint sets of variables X and Y, the **causal effect** of X on Y, denoted as $P(y|\hat{x})$ or $P(y|do(x))$, is the probability of $Y = y$ by deleting all equations from Eq.(2.1) where variables X are on the l.h.s., and substituting $X = x$ in the remaining equations.

This corresponds to mutilating the DAG such that all arrows pointing directly to X_i are removed. **Amputation is the difference between seeing and doing**.

For an atomic intervention, we get the *truncated factorization* formula

$$P(x_1, ..., x_n|\hat{x}'_i) = \begin{cases} \prod_{j \neq i} P(x_j|pa_j) & \text{if } x_i = x'_i \\ 0 & \text{if } x_i \neq x'_i \end{cases} \tag{3.1}$$

The $j \neq i$ denotes the removal of the term $P(x_i|pa_i)$ from Eq.(1.1) (i.e. amputation). A $do(x_i)$ is a severely limited sub-space of the full joint distribution, since the distribution only has support where the intervention variable x_i is equal to its particular intervention value x'_i, rather than a continuum of values in Eq.(1.1).

Multiplying and dividing by $P(x'_i|pa_i)$ yields

$$P(x_1, ..., x_n|\hat{x}'_i) = \begin{cases} P(x_1, ..., x_n|x'_i, pa_i)P(pa_i) & \text{if } x_i = x'_i \\ 0 & \text{if } x_i \neq x'_i \end{cases} \tag{3.2}$$

Marginalization of the above leads to the following theorem.

Adjustment for direct causes Let PA_i denote the set of direct causes of variable X_i, and let Y be any set of variables disjoint of $\{X_i \cup PA_i\}$. The causal effect of $do(X_i = x'_i)$ on Y is

$$P(y|\hat{x}'_i) = \sum_{pa_i} P(y|x'_i, pa_i)P(pa_i) \tag{3.3}$$

where $P(y|x'_i, pa_i)$ and $P(pa_i)$ are preintervention probabilities. This is called "adjusting for PA_i".

Identifiability Causal quantities are defined relative to a causal model M, not the joint distribution $P_M(v)$ over the set of observed variables V. Non-experimental data provides information about $P_M(v)$ alone, and several graphs can give rise to the same $P_M(v)$. Thus, not all quantities are unambiguously **identifiable** from observational data, **even with infinite samples**. Added assumptions by specifying a particular M can provide enough details to compute quantities of interest without explicating M in full.

Theorem 3.2.5: Given a causal diagram G of any Markovian model in which a subset of variables V are measured, the causal effect $P(y|\hat{x})$ is identifiable whenever $\{X \cup Y \cup PA_X\} \subseteq V$. I.e. *all parents of the cause are necessary to estimate the causal effect.*

4. Inferring causal structure

- IC algorithm is for inferring causal structure given observational data when there are no latent variables
- IC* algorithm is for inferring causal structure given observational data when there are latent variables. The PC algorithm is apparently more contemporary (see Spirtes et al 2010)
- There are local criteria for potential cause and genuine cause
- Spurious association: X and Y are spuriously associated if they are dependent in some context and there exists a latent common cause, as exemplified in the structure $Z_1 \to X \to Y \leftarrow Z_2$
- NOTEARS (Zheng et al. 2018) casts the structure learning problem as a continuous optimization problem over real matrices to avoid the superexponential combinatorial explosion with number of variables.

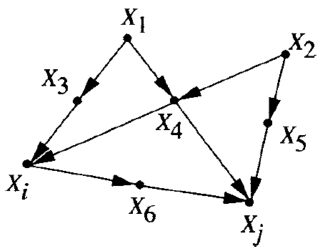

Figure 3.4 A diagram representing the back-door criterion; adjusting for variables $\{X_3, X_4\}$ (or $\{X_4, X_5\}$) yields a consistent estimate of $P(x_j \mid \hat{x}_i)$. Adjusting for $\{X_4\}$ or $\{X_6\}$ would yield a biased estimate.

Figure 1. Example of the back-door criterion

5. Adjusting for confounding bias

When seeking to evaluate the effect of one factor (X) on another (Y), we should ask **whether** we should *adjust* for possible variations in other factors (Z, known as "covariates", "concomitants" or "confounders"). This becomes apparent in **Simpson's paradox**: any statistical relationship between two variables may be reversed by including additional factors in the analysis.

5.1. The back-door criterion

This criterion demonstrates how confounders that *affect* the treatment variable can be used to facilitate causal inference.

Back-door A set of variables Z satisfy the back-door criterion relative to an ordered pair of variables (X_i,X_j) in a DAG G if:

1. no node in Z is a descendant of X_i; and
2. Z blocks every path between X_i and X_j that contains an arrow into X_i

Similarly, if X and Y are two disjoint subsets of nodes in G, then Z satisfies the back-door criterion relative to (X, Y) if it satisfies the criterion relative to any pair (X_i, X_j) such that $X_i \in X$ and $X_j \in Y$.

Back-door adjustment If a set of variables Z satisfies the back-door criterion relative to (X, Y), then the causal effect of X on Y is identifiable and is given by

$$P(y|\hat{x}) = \sum_z P(y|x, z)P(z). \tag{5.1}$$

This corresponds to partitioning the population into groups that are homogeneous relative to Z, assessing the effect of X on Y in each homogeneous group, and then averaging the results. Conditioning in this way means that the observation $X = x$ cannot be distinguished from an intervention $do(x)$.

5.2. The front-door criterion

This criterion demonstrates how confounders that are *affected by* the treatment variable can be used to facilitate causal inference.

Front-door A set of variables Z satisfy the front-door criterion relative to an ordered pair of variables (X, Y) if:

1. Z intercepts all directed paths from X to Y;
2. there is no unblocked back-door path from X to Z; and
3. all back-door paths from Z to Y are blocked by X.

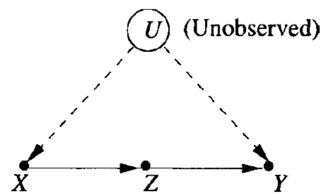

Figure 3.5 A diagram representing the front-door criterion. A two-step adjustment for Z yields a consistent estimate of $P(y \mid \hat{x})$.

Figure 2. Example of the front-door criterion

Front-door adjustment If Z satisfies the front-door criterion relative to (X, Y) and if $P(x, z) > 0$, then the causal effect of X on Y is identifiable and is given by

$$P(y|\hat{x}) = \sum_{z} P(z|x) \sum_{x'} P(y|x', z)P(x') \tag{5.2}$$

Conditions (2) and (3) of the front-door definition are overly restrictive: e.g. nested combinations of back-door and front-door conditions are permissible (see Section 6 for a more general set of conditions).

6. Do-calculus

The back-door and front-door criteria do not provide a complete set of rules for when/how causal effects can be computed. Do-calculus sidesteps the need for algebraic manipulation and provides a complete set of inference rules by which probabilistic sentences involving interventions and observations can be transformed into other such sentences, allowing a method of deriving/verifying claims about interventions. The aim is to compute causal effect expressions for $P(y|\hat{x})$ where Y and X are subsets of variables. When $P(y|\hat{x})$ can be reduced to an expression involving observable probabilistic quantities, we say that the causal effect of X on Y is **identifiable**.

6.1. Notation

- $G_{\overline{X}}$ = graph obtained by deleting from G all arrows pointing into nodes in X
- $G_{\underline{X}}$ = graph obtained by deleting from G all arrows pointing out of nodes in X
- $G_{\overline{X}\underline{Z}}$ = graph obtained by deleting from G all arrows pointing into nodes in X and out of nodes in Z
- $P(y|\hat{x}, z) := P(y, z|\hat{x})/P(z|\hat{x})$, meaning the probability of observing $Y = y$ given an *intervention* $X = x$ and an *observation* $Z = z$

6.2. Rules

Rule 1 (Insertion/deletion of observations)

$$P(y|\hat{x}, z, w) = P(y|\hat{x}, w) \quad \text{if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}}}. \tag{6.1}$$

This rule is a reaffirmation of d-separation (Section 1.1) as a valid test for conditional independence in the distribution resulting from $do(X = x)$. The rule follows from the fact that deleting equations from the system ($G_{\overline{X}}$) does not introduce any dependencies among the remaining disturbance terms.

Rule 2 (Action/observation exchange)

$$P(y|\hat{x}, \hat{z}, w) = P(y|\hat{x}, z, w) \quad \text{if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\underline{Z}}}. \tag{6.2}$$

This rule provides a condition for an external intervention $do(Z = z)$ to have the same effect on Y as the passive observation $Z = z$. The condition amounts to $\{X \cup W\}$ blocking all back-door paths from Z to Y (in $G_{\overline{X}}$), since $G_{\overline{X}\underline{Z}}$ retains all (and only) such paths.

Rule 3 (Insertion/deletion of actions)

$$P(y|\hat{x}, \hat{z}, w) = P(y|\hat{x}, w) \quad \text{if } (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}, \overline{Z(W)}}} \tag{6.3}$$

where $Z(W)$ is the set of Z-nodes that are not ancestors of any W-node in $G_{\overline{X}}$.

This rule provides conditions for introducing (or deleting) an external intervention $do(Z = z)$ without affecting the probability of $Y = y$. The validity of this rule stems from simulating the intervention $do(Z = z)$ by the deletion of all equations corresponding to the variables in Z (hence $G_{\overline{XZ}}$).